VARIANCE REDUCTION VIA AN APPROXIMATING MARKOV PROCESS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF OPERATIONS RESEARCH
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Shane G. Henderson
April 2002

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Peter W. Glynn
(Principal Adviser)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Donald L. Iglehart

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Nicholas Bambos

Approved for the University Committee on Graduate Studies:

Abstract

Many stochastic processes may be approximated by another mathematically more tractable Markov process. For example, the waiting time sequence in a single-server queue may be approximated by a reflected Brownian motion. It seems reasonable that knowledge of the approximation might enable a simulator to increase the efficiency of a simulation of the original process. The method of external control variates is one possible approach in which the approximating process is simulated in parallel with the original process. In this thesis, we present a new method of exploiting the approximation that *does not require the simulation of the approximating process.*

To be specific, suppose that $X = (X(t) : t \geq 0)$ is a Markov process living on state space S. Suppose further that X has a unique stationary probability distribution π. The goal is to estimate the steady-state cost πf, where f is a real-valued cost function on S, and $\pi f = \int_S f(x)\pi(dx)$. Define

$$\alpha_1(t) = t^{-1} \int_0^t f(X(s))\, ds.$$

It is well known that under mild conditions, $\alpha_1(t)$ is a strongly consistent estimator of πf.

Suppose that h is a second real-valued function, and $\pi h = 0$. Then another strongly consistent estimator of πf is given by

$$\alpha_2(t) = t^{-1} \int_0^t (f + h)(X(s))\, ds.$$

If h is chosen appropriately, then it is conceivable that $\alpha_2(t)$ might be a "better" estimator of πf than $\alpha_1(t)$. As an extreme example, if $h(x) = -(f(x) - \pi f)$, then $\alpha_2(t)$ has zero variance!

The approximating Markov process (AMP) method uses information from the approximation to suggest such an h, and then estimates πf by $\alpha_2(t)$.

In this thesis we present the general methodology behind the AMP method, and then discuss applications to the waiting time sequence and queue size process in the single-server queue. In particular, we demonstrate that if heavy-traffic approximations are used for these processes, then order of magnitude variance reductions are possible when estimating steady-state moments and tail probabilities.

Acknowledgments

Were it not for a (disturbingly) large number of people who have contributed more than they will ever know, this little red book would not exist. In fact, were it not for these folks, I would probably be sandblasting right now (which you might conclude would not be such a bad thing, should you venture into the body of this thesis; then again, you've probably never seen me sandblast).

It would take more beer than has ever been brewed to thank my advisor Peter Glynn for all that he has done for me. Quite aside from his encyclopaedic knowledge of the field and his incredible insight, his patience, generosity with time, patience, sense of humour, and patience are unsurpassed. Believe me, I know. I pushed him pretty hard on some of those things occasionally. I also benefitted enormously from the quiet yet sure guidance of Professor Iglehart, and it was very much appreciated. Many thanks also to my third reader Professor Bambos, and to Professors Veinott and Harrison for rounding out the orals committee. I would be remiss if I did not also mention the folks from the University of Auckland and Massey University who helped put me on the path to this degree. I really am grateful to you for all you have done for me.

I particularly want to thank Karl and Matt, specialists in remote thesis development, who were instrumental in getting this thesis to the right office at the right time.

Next, I'd like to thank the friends I made here at Stanford for all they did for (to?) me, and the fun we shared. Thank you for getting me drunk every night before I had to present in the uncertainty seminar, for introducing me to the culinary delights of Denny's at 3am, for explaining the real meaning of "diffusion approximation", for teaching me the art of indoor golf, for nocturnal sunbathing trips to the beach, for the "head-size is correlated with psychic ability" experiment, for taking me down black diamonds the second time I went skiing, for joining in hysterical laughter during qual preparations, for lifting the foul limit to six million so I didn't foul out, for explaining who "Dennis" is, for teaching me satisfying German swear words, for the games of frisbee golf that ended up as Cat Burglary 101, for the card games till dawn (NZers are so much better card players than Germans!), for nearly killing me the day before a certain flight to Toronto, and for just being great mates. If it weren't for you I'd be in a rubber-room right now. And to the Princes of Darkness: I don't know how to thank you. But I'm going to try really, really hard.

I would also like to thank my family, who right from the start encouraged me to aim high, celebrated my successes with me, taught me to learn from and laugh at my failures, and perhaps most importantly, to laugh at myself. You kept my feet firmly on the ground (no shoes), while still giving me room to explore. Although you're on the other side of the Pacific, you're never far away, and I am so grateful. To you I can only say Arohanui, and please get the time differences sorted out before ringing us at 3am again. And I am also grateful to the newest "wing" of my family, the Rooks clan, for their love, their encouragement, and their maintenance of a sense of humour during the annual rendition of fine country music ala "Coward of the County".

Finally I would like to thank my wife Ally, who has taught me so much (but not how to cook). Throughout this project she has been there, always with words of encouragement, always with faith, always with patience, and always with a smile. Over the last few years she has shouldered much so that I could concentrate on my studies, and it has made this task so much easier. Ally, through it all I have seen and felt your warmth, your quiet strength and your smiling eyes. I hope I always will.

To Boris

Contents

List of Tables

List of Figures

Chapter 1

The Approximating Markov Process Method

1.1 Introduction

In many application areas, it is of interest to compute steady-state performance measures. Specifically, suppose that the system under consideration can be modelled by a Markov chain $X = (X(t) : t \geq 0)$ on a state space S. Let $P_x(\cdot) \stackrel{\triangle}{=} P(\cdot | X(0) = x)$ and let $f : S \to \mathbb{R}$ be a "cost function" or "reward function" on S. If the process X satisfies some form of "positive recurrence" condition, then in great generality, the strong law of large numbers (SLLN)

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t f(X(s))\, ds = \alpha \qquad P_x \text{ almost surely (a.s.)} \tag{1.1}$$

holds for all $x \in S$, where α is a deterministic constant that is independent of x. We call α the steady-state mean of f. It is typically impossible to obtain a closed form expression for α and so we must turn to computational methods. We refer to the task of determining α by simulation as the *steady-state estimation problem.*

To determine α, we may define an estimator

$$\alpha_1(t) = \frac{1}{t}\int_0^t f(X(s))\,ds, \tag{1.2}$$

and in view of (1.1), we see that $\alpha_1(t)$ is (strongly) consistent for α. But $\alpha_1(t)$ is just one possible estimator for α. If $h : S \to \mathbb{R}$ is such that $t^{-1}\int_0^t h(X(s))\,ds \to 0$ P_x a.s. for all $x \in S$, then we could also estimate α with

$$\alpha_2(t) \stackrel{\triangle}{=} \frac{1}{t}\int_0^t (f+h)(X(s))\,ds. \tag{1.3}$$

Which estimator should we use?

Under further mild conditions on X, f and h (see Glynn and Meyn (1996)), it can be shown that for $i = 1, 2$,

$$\sqrt{t}(\alpha_i(t) - \alpha) \Rightarrow \sigma_i N(0,1) \tag{1.4}$$

as $t \to \infty$, where $\Rightarrow$ denotes weak convergence, $N(0,1)$ is a standard normal random variable, and σ_i^2 is termed the *time average variance constant (TAVC)* for $\alpha_i(t)$.

The central limit theorem (CLT) (1.4) provides a basis for developing confidence intervals for α, and the widths of the intervals are proportional to σ_i. Therefore, if the computational cost of evaluating the estimators $\alpha_i(t)$ is approximately the same, then it is natural to use the estimator with the lower TAVC. If the cost of computing the two estimators is not the same, then the simulator must take this factor into account in choosing between the estimators. It is no longer the case that the simulator should simply use the estimator with the lowest TAVC, because of the differences in computational cost in computing the two estimators. Glynn and Whitt (1992) examine this issue and develop a framework for quantifying the trade-off between lower variance and higher computational time. Their approach is also applied to several traditional efficiency improvement techniques in Glynn (1994a), including importance sampling and control variables.

We simply assume for now that the costs associated with computing the estimators $\alpha_i(t)$ $(i = 1, 2)$ are comparable, and consider the problem of how to choose the

function h in (1.3) so that $\sigma_2^2 < \sigma_1^2$. Clearly, an optimal choice for h is the function $h(x) = -f(x)+\alpha$. In that case $\sigma_2^2 = 0$! Of course, we do not know the centering term α, so that this function is unimplementable, but it does suggest what the function h should look like. The function h should have steady-state mean 0, and satisfy $h(x) \approx -f(x)+\alpha$. Graphically, the process $h(X(\cdot))$ behaves antithetically to $f(X(\cdot))$, in that $-h(X(\cdot))$ "follows" $f(X(\cdot))$, with the vertical separation between the two processes being approximately α. The effect of using such a function is to obtain a moderated process $((f+h)(X(\cdot)) : s \geq 0)$ with the same steady-state mean as $f(X(\cdot))$. This motivates the following definition.

Definition 1.1 *We call* $h : S \to \mathbb{R}$ *a* shadow function *if* $\pi h = 0$.

Of course, this definition encompasses shadow functions that may result in $\sigma_2^2 > \sigma_1^2$, but to confine the definition to functions that reduce the TAVC is, although theoretically desirable, practically infeasible. We defer the details of how to choose a particular shadow function to later sections, except to say that the approximating Markov process (AMP) method obtains the shadow function through the use of analytical information from a second Markov process. In fact, it is often the case that while the original Markov process X is mathematically intractable (and therefore must be simulated), a more tractable process that approximates the dynamics of X is known. For example, reflected Brownian motion may be used to approximate the sequence of waiting times of customers in a single server queue (see for example, Glynn (1990)).

Another efficiency improvement technique that takes advantage of knowledge of an "approximating process" is the method of (external) control variates (see for example, Gaver and Thompson (1973) p. 586, Bratley, Fox and Schrage (1987), p. 59, or Law and Kelton (1991) p. 641). In this method, both the original process X, and the approximating process Y are simulated. Observations from both simulations are collected and then combined into a single estimator of α. Suppose (for the sake of discussion) that we observe $\alpha_1(t)$ from the simulation of X, and ξ from that of Y,

where $E\xi$ is known. Typically, a linear rule is used to combine the two observations into an estimator of α of the form

$$\hat{\alpha} = \alpha_1(t) + \beta(\xi - E\xi),$$

for some parameter β that is at our disposal. The parameter β is then chosen in an attempt to minimize the variance of $\hat{\alpha}$. When the rv's $\alpha_1(t)$ and ξ are strongly correlated, significant variance reductions are possible. Two main issues arise in implementing external control variates. First, in addition to simulating X, the simulator must also simulate Y, thus incurring a possibly substantial overhead. Second, the simulator is usually required to "synchronize" the simulations of X and Y in order that the correlation referred to above be significant. Law and Kelton (1991) p. 617 discuss the issue of synchronization in the context of another efficiency improvement technique (common random numbers), but the same ideas apply equally well in this context. For example, if we are interested in computing the mean steady-state waiting time in a single-server queue with uniform interarrival and service time distributions, then we might use as the approximating process an M/M/1 queue with the same mean interarrival and service times. One approach to synchronization is to use the same random numbers to generate arrival times and service times in both queues (via the inverse transform method, for example). The hope is that the behaviour of the two waiting time sequences is then very similar, resulting in a high correlation and substantial efficiency improvement. In this example it is easy to obtain a reasonable synchronization, but what if, instead of using an M/M/1 approximation, we instead chose to use a heavy-traffic reflected Brownian motion approximation? It is not at all clear how to synchronize such processes.

In contrast, the AMP method does not require the simulation of the approximating process, so that the additional work in doing so is avoided. Furthermore, synchronization is not an issue. However, we shall see that these advantages come at a price. To implement the AMP method more analytical work is required prior to the simulation.

In Section 1.2 we discuss the application of the AMP method in the case where

both X and its approximation are continuous time Markov chains (CTMC's), and demonstrate the application of the AMP method through a simple example. Then, in Section 1.3, we extend the ideas of Section 1.2 to more general Markov processes, in preparation for later applications.

1.2 Continuous Time Markov Chains

In this section we describe the AMP method in the case where both the original process X and the approximating process are CTMC's. Our intention is to explain the main ideas in a (relatively) simple setting, and motivate the general case which follows in Section 1.3.

Suppose that X is an irreducible positive recurrent CTMC on a countable state space S. Since X is positive recurrent, it possesses a unique stationary probability measure π (Wolff (1989) p. 217). For convenience, we write π_x for $\pi(\{x\})$. Let $f : S \to \mathbb{R}$ be a given cost function and suppose that f is π-integrable, i.e.,

$$\pi|f| \stackrel{\triangle}{=} \sum_{x \in S} |f(x)|\pi_x < \infty.$$

Define $\alpha = \pi f$, and $\alpha_1(t)$ as in (1.2). Let $P_x(\cdot)$ be defined as $P(\cdot|X(0) = x)$. If g is any π-integrable function, then the SLLN holds, i.e.,

$$\frac{1}{t}\int_0^t g(X(s))\,ds \to \pi g \qquad P_x \text{ a.s.} \tag{1.5}$$

as $t \to \infty$ (see Chung (1967), and note that he only proves this result for discrete time Markov chains, but his proof is easily extended to cover the CTMC case considered here). Hence we immediately see that $\alpha_1(t)$ is a consistent estimator of α. Furthermore, if h is a shadow function, then the estimator $\alpha_2(t)$ given by (1.3) is also consistent.

But how do we obtain a shadow function when we do not know π? Recall that if A is the generator (rate matrix) of X, then $\pi A = 0$. Hence, if $h = Ag$, and

$\pi(Ag) = (\pi A)g = 0$, then $\pi h = 0$. The interchange of the order of summation in $\pi(Ag) = (\pi A)g$ will hold if

$$\sum_{x \in S} \sum_{y \in S} \pi_x |A_{xy}| \, |g_y| < \infty.$$

Clearly this is true if S is finite. It is easy to show that this condition also holds in the countably infinite state space case if A is uniformizable (i.e., $\sup_x |A_{xx}| < \infty$) and g is π-integrable, since

$$\sum_{x \in S} \sum_{y \in S} \pi_x |A_{xy}| \, |g_y| = 2 \sum_{y \in S} \pi_y |A_{yy}| \, |g_y|$$

in the sense that if one side is finite then both sides are finite and they are equal, and if one side is infinite (the sum diverges) then both sides are infinite. We may summarise this discussion with the following proposition.

Proposition 1.1 *Suppose that X is a positive recurrent CTMC on a countable state space S, with stationary distribution π and generator A. If $g : S \to \mathbb{R}$ is such that*

$$\sum_{y \in S} \pi_y |A_{yy}| \, |g_y| < \infty, \tag{1.6}$$

then Ag is π-integrable, and $\pi(Ag) = (\pi A)g = 0$, so that Ag is a shadow function.

Proposition 1.1 defines a large class of shadow functions $h = Ag$ for which the estimator $\alpha_2(t) = t^{-1} \int_0^t (f + h)(X(s)) \, ds$ is consistent. We now explore the choice of a function within this class.

Recall that the "optimal" function is $h(x) = -(f(x) - \alpha)$, because then $\alpha_2(t) = \alpha$ for all t, which is a zero variance estimator. Hence, we are naturally led to the set of equations

$$Ag(x) = -(f(x) - \alpha) \qquad \text{for all } x \in S. \tag{1.7}$$

Equation (1.7) is known in the literature as "Poisson's equation", and it occupies a central role in the asymptotic analysis of the estimator $\alpha_1(t)$ (Glynn and Meyn

(1996)). Notice that (1.7) does not uniquely determine g. If g is a solution to (1.7), then so is $g+c$, where c is a constant function. We shall see that this does not matter for our analysis, and *any* π-integrable solution to (1.7) will do.

Of course, we cannot solve (1.7), because of the presence of the centering constant α. However, it does provide guidance in selecting a suitable function g. In particular, it suggests that if g is an approximation to the solution to Poisson's equation, then $h = Ag$ might be a useful choice for the estimator (1.3). For convenience, we will refer to g as a *surrogate function* if g is an approximation to the solution to Poisson's equation.

But where can we obtain a reasonable surrogate function?

We propose to use the solution to Poisson's equation for a stochastic process that is similar to X. Suppose that X can be approximated by another positive recurrent CTMC $\tilde{X}$ on a countable state space $\tilde{S}$. Let $\tilde{X}$ have stationary distribution $\tilde{\pi}$ and generator $\tilde{A}$.

We assume that there is a natural correspondence between the elements of S, and the elements of $\tilde{S}$, which is represented by a mapping $r : S \to \tilde{S}$ that is not necessarily invertible (see Example 1.1). Further, we let $\tilde{f} : \tilde{S} \to \mathbb{R}$ be a function that is (in some sense) intimately related to the original cost function $f : S \to \mathbb{R}$.

Let us further suppose that $\tilde{\pi}|\tilde{f}| < \infty$, and we can compute $\tilde{\pi}\tilde{f}$ and solve Poisson's equation

$$\tilde{A}\tilde{g}(\tilde{x}) = -(\tilde{f}(\tilde{x}) - \tilde{\pi}\tilde{f}) \quad \forall \tilde{x} \in \tilde{S}.$$

A solution $\tilde{g}$ to this equation then furnishes the required surrogate function via

$$g(x) = \tilde{g}(r(x)) \quad \forall x \in S. \tag{1.8}$$

If g satisfies the conditions of Proposition 1.1, then a consistent estimator of α is given by $\alpha_2(t) = t^{-1} \int_0^t (f + Ag)(X(s))\, ds$.

Example 1.1 Suppose that customers arrive at a service facility according to a Poisson process with rate λ and are served in a FIFO fashion. Upon entry to the service facility, service times are determined to be exponential with mean μ_1^{-1} or μ_2^{-1}, with respective probabilities $0 < p_1 < 1$ and $p_2 = 1 - p_1$. Let $X_1(t)$ be the number of customers in the system at time t, $X_2(t)$ be the class of the customer currently in service and set $X(t) = (X_1(t), X_2(t))$. Then $X = (X(t) : t \geq 0)$ is a CTMC with state space $S = \{0, 1, 2, \ldots\} \times \{1, 2\}$. (The states $(0, 1)$ and $(0, 2)$ could be coalesced into a single state, but having two states simplifies the notation.) The non-zero transition rates λ_{xy} of X are given by

$$\begin{aligned} \lambda_{(0,j),(1,k)} &= \lambda p_k, \\ \lambda_{(i,j),(i+1,j)} &= \lambda \text{ and} \\ \lambda_{(i,j),(i-1,k)} &= \mu_j p_k \end{aligned}$$

for $j, k \in \{1, 2\}$ and $i \geq 1$. The process $X_1(\cdot)$ is the queue size process in an M/G/1 queue, and so X is positive recurrent if

$$\lambda < \mu \stackrel{\triangle}{=} (p_1 \mu_1^{-1} + p_2 \mu_2^{-1})^{-1},$$

so let us assume that this is the case in what follows. Suppose we are interested in computing the probability that there are q or more customers in the system in steady-state. Then the cost function f may be defined by $f(i, j) = I(i \geq q)$, where I is the indicator function that is 1 if its argument is true, and 0 otherwise.

This queueing system may be approximated by an M/M/1 queue with arrival rate λ, service rate μ, and traffic intensity $\rho \stackrel{\triangle}{=} \lambda/\mu$. Define $\tilde{X}(t)$ to be the number of customers in the system at time t, so that $\tilde{S} = \{0, 1, 2, \ldots\}$ and set $\tilde{f}(i) = I(i \geq q)$. Finally, define the mapping $r : S \to \tilde{S}$ by $r(i, j) = i$.

Let $\tilde{g}$ be a solution to Poisson's equation for the M/M/1 queue and cost function $\tilde{f}$, i.e., $-\lambda \tilde{g}(0) + \lambda \tilde{g}(1) = \rho^q$, and for $i \geq 1$,

$$\mu \tilde{g}(i-1) - (\lambda + \mu)\tilde{g}(i) + \lambda \tilde{g}(i+1) = -(I(i \geq q) - \rho^q).$$

Set $g(i,j) = \tilde{g}(r(i,j)) = \tilde{g}(i)$. Then, for $j = 1, 2$,

$$\begin{aligned} h(0,j) = Ag(0,j) &= \lambda p_1(g(1,1) - g(0,j)) + \lambda p_2(g(1,2) - g(0,j)) \\ &= \lambda(\tilde{g}(1) - \tilde{g}(0)) \end{aligned}$$

and for $i \geq 1$, $j = 1, 2$,

$$\begin{aligned} h(i,j) = Ag(i,j) &= \lambda(g(i+1,j) - g(i,j)) + \mu_j p_1(g(i-1,1) - g(i,j)) \\ &\quad + \mu_j p_2(g(i-1,2) - g(i,j)) \\ &= \lambda(\tilde{g}(i+1) - \tilde{g}(i)) - \mu_j(\tilde{g}(i) - \tilde{g}(i-1)). \end{aligned}$$

The estimator of the probability that the system size is q or more is then given by

$$\frac{1}{t}\int_0^t I(X_1(s) \geq q) + h(X(s))\, ds.$$

We may determine that this estimator is consistent as follows. Glynn and Torres (1996) compute $\tilde{g}$ explicitly. They state that for $i < q$,

$$\tilde{g}(i) = \frac{\rho^q(\rho^{-i} - 1)}{\mu(1-\rho)^2} - \frac{\rho^q i}{\mu(1-\rho)},$$

and for $i \geq q$,

$$\tilde{g}(i) = \frac{1-\rho^q}{\mu(1-\rho)^2} - \frac{q\rho^q}{\mu(1-\rho)} + \frac{1-\rho^q}{\mu(1-\rho)}(i-q).$$

We note that $\tilde{g}(i)$ is bounded by a linear function in i, and therefore $g(i,j) \leq a_1 + a_2 i$ for some non-negative constants a_1 and a_2. Since A is uniformizable, the estimator that we derived will be consistent if $\pi|g| < \infty$. But this follows since $X_1(t)$ is the queue size in an M/G/1 queue, and therefore (by the Pollaczek-Khintchine formula) has a finite steady-state mean.

To compare the standard and AMP estimators for this example, we performed a simulation experiment. We chose $p_1 = p_2 = 1/2$, $\lambda = 1$ and $q = 4$. In order to obtain results which reflect the "closeness" of the approximating M/M/1 queue to the original process, we chose several values of the two service rates μ_1 and μ_2, while maintaining $\rho = 1/2$. We ran the simulations for 20 repetitions of 2000 regenerative

cycles, where the regeneration times were returns to the states $(0,1)$ or $(0,2)$. The results are presented in Table 1.1. The first column gives the rate μ_1, and the second column the value of μ_2 required to give $\rho = 1/2$. Column 3 contains the true value of the tail probability (πf where π is the stationary distribution of X). Columns 4, 5 and 6 contain the point estimate, TAVC estimate, and a 95% confidence interval for the TAVC estimate for the standard estimator. Columns 7, 8 and 9 contain the corresponding results for the AMP estimator.

We should mention how we computed the true value of the tail probability. In general, expressions for tail probabilities in the M/G/1 queue are not known, but for this example, we are able to compute their value as follows. First, using a standard approach (see e.g., Gross and Harris (1985) p. 257), we find the generating function $\Pi(z) = \sum_{n=0}^{\infty} \pi_n z^n$ for the steady-state number of customers in the system. After substituting $p_0 = p_1 = 0.5$, $\lambda = 1$, and using the constraint that $\rho = 0.5$, we obtain

$$\Pi(z) = \frac{\mu_1\mu_2(3-z)/4}{z^2 - (1+\mu_1\mu_2)z + 3\mu_1\mu_2/2}.$$

Now for each of the values for μ_1 and μ_2 we find the partial fraction expansion of $\Pi(z)$, expand those expressions as power series, and thus obtain the steady-state probabilities. In particular, if $\Pi(z) = a(r_1 - z)^{-1} + b(r_2 - z)^{-1}$, then

$$\pi f = \frac{a r_1^{-q}}{r_1 - 1} + \frac{b r_2^{-q}}{r_2 - 1}.$$

Looking at columns 5 and 8, we see that as μ_1 and μ_2 approach one another, the variance reductions become quite considerable. This is perhaps to be expected, because in some sense, the M/M/1 approximation becomes better as μ_1 approaches μ_2.

μ_1	μ_2	πf	Standard Estimator			AMP Estimator		
			Est	TAVC	CI	Est	TAVC	CI
5	1.25	0.0963	0.095	0.52	± 0.06	0.096	0.12	± 0.01
4	1.33	0.0864	0.086	0.45	± 0.06	0.086	0.087	± 0.007
3	1.5	0.0734	0.077	0.43	± 0.08	0.074	0.038	± 0.004
2.5	1.67	0.0665	0.067	0.32	± 0.06	0.067	0.013	± 0.002
2.1	1.91	0.0627	0.064	0.30	± 0.04	0.063	7.3E-4	± 6E-5
2.01	1.99	0.0625	0.064	0.29	± 0.04	0.062	8.0E-6	± 6E-7
2	2	0.0625	0.0625	0.29	± 0.03	0.0625	0	± 0

Table 1.1: Simulation results for the CTMC example.

1.3 The General Case

Our intention in this section is to explain how the ideas in the previous section extend to more general Markov processes. Recall the major points in the previous section.

1. If $\pi|f| < \infty$ then $t^{-1}\int_0^t f(X(s))\,ds \to \alpha \quad P_x$ a.s., i.e., $\alpha_1(t)$ as defined in (1.2) is consistent.

2. If h is a shadow function (i.e., $\pi h = 0$), then the estimator $\alpha_2(t)$ given by (1.3) is also consistent.

3. If g satisfies a certain integrability requirement (Proposition 1.1), then Ag is a shadow function.

4. If we use $\alpha_2(t)$ to estimate α, then an optimal choice for g is the solution to Poisson's equation. We cannot solve this equation explicitly, but we can find approximate solutions to it (surrogate functions) by solving a similar equation from a second Markov process.

To extend these results, we first need to place our discussion within a sound theoretical framework. The following definition is given in Sigman (1990), and Azema, Duflo, and Revuz (1969). Let X be a Markov process in discrete or continuous time

on some state space S. For certain technical reasons we require that S be a complete, separable metric space, that X has paths in $D_S[0,\infty)$ (the space of right-continuous paths with left limits everywhere), and that X is a strong Markov process. These assumptions are certainly satisfied for almost all practical applications, and so, to all intents and purposes, can be ignored.

To conclude that X possesses a unique stationary probability distribution π, we must assume some form of positive recurrence condition. We shall therefore assume that X is *positive Harris recurrent*.

Definition 1.2 *A Markov process satisfying the above regularity conditions is said to be* Harris recurrent *if there exists a non-trivial σ-finite non-negative measure η on S such that whenever $\eta(B) > 0$,*

$$P_x(\int_0^\infty I(X(s) \in B)\, ds = +\infty) = 1$$

for all $x \in S$. (We have given this definition in the case where X is a continuous-time process. If X is a discrete-time process, then we require that

$$P_x(\sum_{k=0}^{\infty} I(X_k \in B) = +\infty) = 1$$

for all $x \in S$.)

Such processes automatically possess a unique non-trivial σ-finite invariant measure π. If π is finite, then it can be normalized to a probability, and we then say that X is positive Harris recurrent.

The notion of positive Harris recurrence generalizes that of positive recurrence for discrete and continuous time Markov chains to processes with general state spaces. We are nearly in a position to state our first results, but first we need the SLLN for positive Harris recurrent processes.

Theorem 1.1 *Suppose that X is a positive Harris recurrent process on state space S with stationary probability distribution π. If $g : S \to \mathbb{R}$ is π-integrable (i.e., $\pi|g| \stackrel{\triangle}{=} \int_S |g(x)|\pi(dx) < \infty$), then*

$$t^{-1} \int_0^t g(X(s))\, ds \to \pi g$$

P_x a.s. as $t \to \infty$ for all $x \in S$.

For a proof of this result see Sigman (1990) or Azema, Duflo and Revuz (1969).

So suppose that $f : S \to \mathbb{R}$ is a π-integrable cost function, and h is a shadow function. The above SLLN proves that the estimators (1.2) and (1.3) of α are consistent.

Our next problem is to determine a class of shadow functions. In the CTMC case we used functions of the form Ag, where A was the rate matrix of the CTMC. We shall use the same method here, but A will no longer necessarily be a matrix. We need to introduce the notion of the *generator* of a Markov process.

1.3.1 The Generator of a Markov Process

There are several definitions of the generator of a Markov process. We shall use the following one, taken from Kurtz (1969). This definition of the generator is usually referred to as the *bounded-pointwise* generator.

Let L be the Banach space of bounded, measurable, real-valued functions from S to $\mathbb{R}$. (A Banach space is a complete, normed, linear space.) Let $P(t)f(x) \stackrel{\triangle}{=} E_x f(X(t))$. Then $(P(t) : t \geq 0)$ is known as the transition semigroup of X on L. We say that a sequence $\{f_k\} \subseteq L$ converges boundedly and pointwise to f if

$$\sup_k \|f_k\| < \infty \text{ and } \lim_{k \to \infty} f_k(x) = f(x),$$

and we shall write

$$bp - \lim_{k\to\infty} f_k = f.$$

Definition 1.3 *The (bounded-pointwise) generator of a continuous time Markov process $X = (X(t) : t \geq 0)$ is the linear operator $\bar{A}$ defined by*

$$\bar{A}f = bp - \lim_{t\downarrow 0} \frac{P(t)f - f}{t}. \tag{1.9}$$

The domain $D(\bar{A})$ of $\bar{A}$ is the set of all $f \in L$ for which this limit exists. Notice that if the limit exists, then it is automatically in L.

A similar definition applies for discrete time processes.

Definition 1.4 *The (bounded-pointwise) generator of a discrete time process $X = (X_n : n \geq 0)$ is the linear operator $\bar{A}$ defined by*

$$\bar{A}f = Pf - f \tag{1.10}$$

where $Pf(x) \stackrel{\triangle}{=} E_x f(X_1)$. The domain $D(\bar{A})$ of $\bar{A}$ is the set of all $f \in L$ for which $Pf - f \in L$, which is simply L.

A natural question to ask is whether the generator provides enough information to uniquely determine the process X, once the initial distribution of X is known. The answer is yes, subject to a regularity condition on the transition probabilities (see e.g., Breiman (1968) p. 343). Therefore, the term "generator" is justified.

To understand what the generator of a Markov process actually is, we may reason as follows. The expression (1.10) states that $\bar{A}f(x)$ is the expected change in the value of the stochastic process $f(X_n)$ over the next transition, if X_n is currently equal to x. Similarly, (1.9) states that $\bar{A}f(x)$ is the forward derivative of the function $E_x f(X(t))$ at $t = 0$, when $X(0) = x$. So we see that in both discrete and continuous time,

the generator of X provides information on the expected local (in the sense of time) change in functions of $X(t)$.

There are two cases where the generator will be familiar. First, if X is a CTMC on a countable state space, then $\bar{A}$ is the associated rate matrix (see Proposition 1.2 below). Second, if X is a countable state space DTMC with transition matrix P, then $\bar{A} = P - I$, where I is the identity matrix. It is also possible to describe the generator of far more complicated Markov processes, such as diffusions, in which case (as we shall see), the generator takes the form of a differential operator.

Proposition 1.2 *Let X be a non-explosive (only a finite number of transitions can occur in finite time) CTMC with no instantaneous states (states for which the holding time is 0 with probability 1), and let Q be the associated rate matrix. Define $f(x) = I(x = m)$ for some fixed state m. Then*

$$\lim_{t \downarrow 0} \frac{P(t)f(x) - f(x)}{t} = Qf(x) = Q_{xm}.$$

(Notice that this limit is a pointwise *limit.) Furthermore, if X is uniformizable, then the limit is bounded-pointwise, so that $\bar{A}f = Qf$, where $\bar{A}$ is the bounded-pointwise generator of X.*

Remark 1.1 *If X is uniformizable, so that there is a $\lambda > 0$ such that $|Q_{xx}| < \lambda$ for all $x \in S$, then we may write*

$$P(t) = \sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} B^n$$

where B is stochastic (see e.g., Ross (1983) p. 176). Using this representation for $P(t)$ it is straight-forward to show that $D(\bar{A})$ contains all bounded functions, and $\bar{A}f = Qf$. Our approach to the proof was adopted in order to illustrate the distinction between pointwise convergence, and bounded-pointwise convergence. Typically, additional conditions are required to obtain bounded-pointwise convergence from pointwise convergence.

Proof. Suppose first that X is a non-explosive CTMC on state space S, and has no instantaneous states. Recall that $P_x(X(t) = m)$ is continuously differentiable in t and $P'_x(X(t) = m) = \sum_z Q_{xz} P_z(X(t) = m)$ (Asmussen (1987) p. 36). Define $\delta_{xy} = I(x = y)$, and note that for $x \in S$,

$$\begin{aligned} t^{-1}(P(t)f(x) - f(x)) &= t^{-1}(P_x(X(t) = m) - \delta_{xm}) \\ &= t^{-1} \int_0^t P'_x(X(s) = m)\, ds \\ &= P'_x(X(\xi_t) = m) \\ &= \sum_z Q_{xz} P_z(X(\xi_t) = m) \end{aligned}$$

for some (deterministic constant) $\xi_t \in [0, t]$. But $P_z(X(\xi_t) = m) \to P_z(X(0) = m) = \delta_{zm}$ as $t \downarrow 0$, and since

$$|\sum_z Q_{xz} P_z(X(\xi_t) = m)| \leq \sum_z |Q_{xz}| < \infty, \tag{1.11}$$

it follows that

$$t^{-1}(P(t)f(x) - f(x)) \to \sum_z Q_{xz} \delta_{zm} = Q_{xm} = Qf(x).$$

Thus, we have established pointwise convergence to $Qf(x)$. To show bounded-pointwise convergence, we further require that the functions $t^{-1}(P(t)f(x) - f(x))$ are bounded in x and t for all sufficiently small t. A simple sufficient condition for this (from equation (1.11)) is that $\sum_z |Q_{xz}| \leq \lambda < \infty$ for some constant λ, i.e., X is uniformizable. ■

Notice that we did not specify the domain of the generator for a CTMC. This is perhaps the most difficult aspect of defining the generator of a process.

In the application of the AMP method, we will encounter functions which are not in the domain of the generator, simply because they are not bounded. For that reason, we will shortly define the extended generator of a Markov process. First, we provide some motivation for that definition.

Proposition 1.3 *Let $\bar{A}$ be the bounded-pointwise generator of a positive Harris recurrent continuous time process $X = (X(t) : t \geq 0)$ with transition semigroup $(P(t) : t \geq 0)$. Let $g \in D(\bar{A})$. Then for $t \geq 0$,*

1. $P(t)g - g = \int_0^t P(s)\bar{A}g\, ds$
2. $M(t) \stackrel{\triangle}{=} g(X(t)) - g(X(0)) - \int_0^t \bar{A}g(X(s))\, ds$ *is a P_μ martingale, for all initial distributions μ.*
3. *If π is the stationary distribution of X, then $\pi(\bar{A}g) = 0$.*
4. *The above results also hold if X is a discrete time process. In that case, $P^n g - g = \sum_{k=0}^{n-1} P^k \bar{A}g$, and $M_n = g(X_n) - g(X_0) - \sum_{k=0}^{n-1} \bar{A}g(X_k)$ is a P_μ martingale for any initial distribution μ. Also, if π is the stationary distribution for X then $\pi(\bar{A}g) = 0$.*

Proof. The first statement is proved in Lemma 3.4 of Kurtz (1969), and the second result is proved in Ethier and Kurtz (1986), p. 162. To prove the third statement, we use the martingale $M(t)$. Since $M(0) = 0$, it follows that $E_\pi M(t) = 0$ for all $t \geq 0$. (The notation $E_\pi(\cdot)$ is defined to mean the conditional expectation of $\cdot$, given that $X(0)$ is distributed according to π.) Since π is stationary for X, $E_\pi g(X(t)) = \pi g = E_\pi g(X(0))$, so that $E_\pi M(t) = E_\pi \int_0^t \bar{A}g(X(s))\, ds = 0$ for all $t \geq 0$. Since $\bar{A}g$ is bounded, we may apply Fubini's theorem to obtain

$$\begin{aligned} 0 &= E_\pi \int_0^t \bar{A}g(X(s))\, ds \\ &= \int_0^t E_\pi \bar{A}g(X(s))\, ds \\ &= t\,\pi(\bar{A}g) \end{aligned}$$

Hence $\pi(\bar{A}g) = 0$. The discrete time case is similar. ■

We remark that the martingale $M(t)$ is sometimes referred to as the Dynkin martingale (p. 298 of Karlin and Taylor (1981)), and it plays a significant role in the

characterization of Markov processes (see Ethier and Kurtz (1986) for an extensive discussion of the *martingale problem*).

The observations in Proposition 1.3 motivate the following definition of the extended generator.

Definition 1.5 *Let X be a continuous time Markov process with transition semigroup $(P(t) : t \geq 0)$. Denote by $D(A)$, the set of all functions $g : S \to \mathbb{R}$ (not necessarily bounded) for which there exists a function $h : S \to \mathbb{R}$ such that for each $x \in S$,*

$$M(t) \stackrel{\triangle}{=} g(X(t)) - g(X(0)) - \int_0^t h(X(s))\, ds \tag{1.12}$$

is a P_x-martingale. (A process is said to be a P_x-martingale if it is a martingale under the probability law P_x.) We then write $Ag = h$, and call A the extended generator *of the process X.*

Proposition 1.3 shows that $D(\bar{A}) \subseteq D(A)$, and that $Ag = \bar{A}g$ whenever $g \in D(\bar{A})$, thereby justifying the term *extended* generator. We remark that the extended generator A may be multi-valued, i.e., there may exist more than one function h such that $Ag = h$. However, for our purposes, any π-integrable h will suffice.

Suppose that $h = Ag$, and both g and h are π-integrable. Then $(M(t) : t \geq 0)$ is also a P_π martingale, and it follows (by the same argument used in Proposition 1.3) that $\pi(Ag) = 0$, so that Ag is a shadow function. Thus, sufficient conditions for Ag to be a shadow function are that $g \in D(A)$, and both g and Ag are π-integrable. The problem of determining whether $g \in D(A)$ is quite difficult in general, and so we will answer this question application by application.

If X is a discrete time process, then we may define the extended generator in a similar fashion to the continuous time case. However, the following (nonstandard) definition will serve our needs better. We define $D(A)$ to be the set of functions $f : S \to \mathbb{R}$ such that $Pf(x) \stackrel{\triangle}{=} E_x f(X_1)$ exists and is finite for every $x \in S$, and

set $Af(x) = Pf(x) - f(x)$. Now, if f is a π-integrable function, then Pf is also π-integrable, and $\pi(Pf) = (\pi P)f = \pi f$. Therefore, if $f \in D(A)$ and f is π-integrable, then

$$\pi(Af) = \pi(Pf) - \pi f = 0,$$

i.e., Af is a shadow function.

Thus, when we refer to discrete time Markov processes we shall adopt this definition of the extended generator, and when we are dealing with continuous time processes, we shall use the martingale definition provided above. In any case, we have proved the following proposition.

Proposition 1.4 *Suppose that X is a Markov process (in discrete or continuous time), with stationary distribution π, and extended generator A. If $g \in D(A)$, and both g and Ag are π-integrable, then Ag is a shadow function. If X is a discrete time process, the condition that Ag be π-integrable is redundant.*

We now discuss the generators of one dimensional diffusions and generalized semi-Markov processes (GSMP's), because of the importance of these processes to our intended applications. Our discussion of diffusions is based on Chapter 16 of Breiman (1968) and Chapter 15 of Karlin and Taylor (1981). Our principal reference in deriving the generator of a GSMP is Burman (1981). For a comprehensive treatment of the theory of generators, see Ethier and Kurtz (1986).

Generators for Diffusions

Definition 1.6 *We say that $X = (X(t) : t \geq 0)$ is a one-dimensional diffusion if X satisfies the following conditions, where $\Delta_t \stackrel{\triangle}{=} X(t) - X(0)$.*

1. *X is a Markov process with stationary transition probabilities, whose state space I is an interval with endpoints l, r, where $-\infty \leq l < r \leq \infty$.*

2. *Starting from any point $x \in I$, all sample paths are continuous.*

3. *The transition probabilities of X are* stable*; i.e., if ν_n is any sequence of initial distributions such that $\nu_n \Rightarrow \nu_\infty$, and μ_n is the conditional distribution of $X(t)$ given that $X(0)$ is distributed according to ν_n $(1 \leq n \leq \infty)$, then*

$$\mu_n \Rightarrow \mu_\infty$$

as $n \to \infty$, for all $t > 0$.

4. *For every $\epsilon > 0$*

$$t^{-1} P_x(|\Delta_t| > \epsilon) \overset{\text{bp}}{\to} 0,$$

as $t \downarrow 0$.

5. *For every $\epsilon > 0$,*

$$t^{-1} E_x(\Delta_t; |\Delta_t| < \epsilon) \overset{\text{bp}}{\to} \mu(x),$$

as $t \downarrow 0$.

6. *For every $\epsilon > 0$,*

$$t^{-1} E_x(\Delta_t^2; |\Delta_t| < \epsilon) \overset{\text{bp}}{\to} \sigma^2(x),$$

as $t \downarrow 0$.

7. *The functions $\mu(x)$ and $\sigma^2(x)$ are continuous functions on (l, r), and $\sigma^2(x) > 0$ for every $x \in (l, r)$.*

The notation $\overset{\text{bp}}{\to}$ represents bounded and pointwise convergence on every finite interval J whose closure is contained in (l, r).

We remark that processes satisfying conditions 1 to 3 are known as *Feller* processes. Such processes enjoy a kind of "continuity" property which, intuitively speaking, states that the future behaviour of the process does not change dramatically when small changes are made to the initial state.

To completely specify the process X, we must define how X behaves if it hits the boundary $\{l, r\}$. For example, X may be absorbed at l, or reflected at l, and so on. This matter is examined in more detail in Ethier and Kurtz (1986) p. 366.

Denote by $\bar{C}$ the set of all continuous functions f on (l, r) for which $\lim_{x \downarrow l} f(x)$ and $\lim_{x \uparrow r} f(x)$ exist and are finite. We are now ready to define the bounded-pointwise generator $\bar{A}$ of X. For twice continuously differentiable functions f on (l, r), and for every $x \in (l, r)$, set

$$\bar{A}f(x) = \frac{\sigma^2(x)}{2} f''(x) + \mu(x) f'(x). \tag{1.13}$$

The domain $D(\bar{A})$ of $\bar{A}$ includes the set of all twice continuously differentiable functions $f \in \bar{C}$, satisfying certain boundary conditions, and for which $\bar{A}f \in \bar{C}$. For a description of the boundary conditions, see Ethier and Kurtz p. 366, and Mandl (1968) p. 39.

A heuristic argument showing how (1.13) arises as a pointwise limit is given on p. 193 of Karlin and Taylor (1981).

Example 1.2 To illustrate the above definition, and to demonstrate how the extended generator may be derived from knowledge of the bounded-pointwise generator, we present the following example. Let $X = (X(t) : t \geq 0)$ be a reflected (or regulated) Brownian motion with drift $-\mu < 0$, infinitesimal variance $\sigma^2 > 0$, and initial state $X(0) = 0$. Then X is a one-dimensional diffusion with $\mu(x) = -\mu$ and $\sigma^2(x) = \sigma^2$.

For functions f in the domain $D(\bar{A})$ of the bounded-pointwise generator $\bar{A}$,

$$\bar{A}f(x) = \frac{\sigma^2}{2} f''(x) - \mu f'(x). \tag{1.14}$$

To conclude that $f \in D(\bar{A})$ we require, first of all, that both f and $\bar{A}f$ are contained in $\bar{C}$, and that f is twice continuously differentiable. To fully specify the domain of the generator, we need to specify the boundary conditions. From Mandl (1968) p. 39, the appropriate condition is that $f'(0) = 0$.

It will be instructive to discuss the extended generator A of X as well. For this purpose, we introduce the operator Q, where Qg is defined by the right-hand side of (1.14), not only for $g \in D(\bar{A})$, but for all functions g such that the right-hand side of (1.14) is defined. We will show that for certain functions $g \in D(A)$, $Ag = Qg$ using a truncation argument. To begin with, suppose that $g(x)$ is twice continuously differentiable, and $g'(0) = 0$.

For $x \notin (n, n+1)$, define

$$g_n(x) = \begin{cases} g(x) & \text{if } x \leq n \text{ and} \\ g(n+1) & \text{if } x \geq n+1, \end{cases}$$

and for $x \in (n, n+1)$, define $g_n(x)$ so that g_n is twice continuously differentiable over the entire real line. Then $g_n \in D(\bar{A})$ and so

$$\bar{M}_n(t) = g_n(X(t)) - g_n(X(0)) - \int_0^t \bar{A}g_n(X(s))\,ds$$

is a P_x martingale for each $x \in S$. Furthermore $Ag_n(x) = Qg(x)$ for $x \leq n$.

Define $T_n = \inf\{t \geq 0 : X(t) \geq n\}$, and note that

$$\bar{M}_n(t \wedge T_n) = g(X(t \wedge T_n)) - g(X(0)) - \int_0^{t \wedge T_n} Qg(X(s))\,ds \stackrel{\triangle}{=} M_n(t),$$

so that the right hand side of this expression is a P_x martingale for all x and for all n. In particular, $E_x M_n(t) = 0$ for all x and all n. We would like to show that this also holds in the limit as $n \to \infty$, because we would then obtain that $M(t)$ is a P_x martingale for all x (Karlin and Taylor (1981) p. 309), where

$$M(t) \stackrel{\triangle}{=} g(X(t)) - g(X(0)) - \int_0^t Qg(X(s))\,ds,$$

and thus $g \in D(A)$ and $Ag = Qg$.

So suppose that $X(0) = x$, and $x \leq n$. Then

$$\begin{aligned} g(X(t \wedge T_n)) &= |g(X(T_n))| I(T_n \leq t) + |g(X(t))| I(T_n > t) \\ &\leq |g(n)| I(T_n \leq t) + |g(X(t))|. \end{aligned}$$

Under P_x, $X(t) \leq_{\text{stoch}} x + \tilde{X}(t)$, where $\tilde{X}(\cdot)$ is a driftless reflected Brownian motion with the same infinitesimal variance constant as $X(t)$, initial state $\tilde{X}(0) = 0$ and the symbol $\leq_{\text{stoch}}$ denotes stochastic domination (Ross (1983) Chapter 8). But $\tilde{X}(t) \stackrel{\mathcal{D}}{=} |N(0, \sigma^2)|$, and thus has an exponentially decaying tail. Hence, if g is a polynomial, then $E_x|g(X(t))| < \infty$.

Furthermore, $|g(n)|I(T_n \leq t) \to 0$ P_x a.s. as $n \to \infty$, so that if we also show that $|g(n)|P_x(T_n \leq t) \to 0$, then we may conclude that $(g(X(t \wedge T_n)) : n \geq 1)$ is a uniformly integrable sequence of rv's, and hence $E_x g(X(t \wedge T_n)) \to E_x g(X(t))$ as $n \to \infty$. But

$$\begin{aligned} P_x(T_n \leq t) &= P_x(\sup_{0 \leq s \leq t} X(s) \geq n) \\ &\leq P(\sup_{0 \leq s \leq t} \tilde{X}(s) \geq n - x) \\ &\leq 2P(\sup_{0 \leq s \leq t} \bar{X}(s) \geq n - x) \end{aligned}$$

where $\bar{X}$ is a driftless Brownian motion with the same infinitesimal variance constant as $\tilde{X}$. This final expression is bounded above by ae^{-bn} for some constants a and b. (This follows from the distribution of the maximum of a driftless Brownian motion, given on Karlin and Taylor (1975) p. 346). Thus, if g is a polynomial, then we are done.

It remains to show that

$$E_x \int_0^{t \wedge T_n} Qg(X(s))\, ds \to E_x \int_0^t Qg(X(s))\, ds$$

as $n \to \infty$. Suppose that g is a polynomial. Then Qg is too, and we may find another polynomial $\bar{g}$ say, such that $|Qg(x)| \leq \bar{g}(x)$, where $\bar{g}(x)$ is increasing. Then

$$\begin{aligned} \left| \int_0^{t \wedge T_n} Qg(X(s))\, ds \right| &\leq \int_0^t \bar{g}(X(s))\, ds \\ &\leq t \sup_{0 \leq s \leq t} \bar{g}(X(s)) \\ &= t\bar{g}(\sup_{0 \leq s \leq t} X(s)) \end{aligned}$$

which has a finite expectation, so that the required convergence follows from the dominated convergence theorem.

We have thus proved that if g is a polynomial with the property that $g'(0) = 0$, then $g \in D(A)$, and $Ag = Qg$. In fact, the above proof yields a stronger conclusion, but our purpose here is to merely demonstrate the methodology, rather than obtain the tightest possible result.

Generators for Generalized Semi-Markov Processes

Generalized Semi-Markov Processes (GSMP's) are a class of stochastic processes that may be used to model discrete-event dynamical systems (Glynn (1988), Shedler (1993)). We shall define a special class of GSMP's, following the presentation in Burman (1981) closely, and then describe a class of functions that belong to the domain of the bounded-pointwise generator for such processes.

Let $N = (N(t) : t \geq 0)$ be a continuous time parameter stochastic process living on a countable state space J. To each $j \in J$, we associate a set of "active events" $a(j)$, where $a(j) \subseteq E$, a finite space of events. The event $e \in E$ requires T_e units of processing time to be completed, and we let $F_e(t) = P(T_e \leq t)$ be the distribution function of the rv T_e. While in state j, processing occurs on events $e \in a(j)$ at deterministic rate r_{ej}, and for inactive events ($e \notin a(j)$), we define $r_{ej} = 0$.

Suppose that $N(t) = j$, and let $C_e(t)$ be the amount of processing already completed on the event e (where we take $C_e(t) = 0$ for inactive events). The remaining time s until the first event in $a(j)$ has completed processing is given by

$$s = \min_{e \in a(j)} \frac{T_e - C_e(t)}{r_{ej}}.$$

Let $\tilde{e}$ be an event which achieves this minimum. We assume that F_e has a bounded and continuous density f_e (with respect to Lebesgue measure) for each $e \in E$, so that

$\tilde{e}$ may be defined uniquely with probability 1. At time $t+s$, the process $N(t)$ enters state j' with probability $p(j';j,\tilde{e})$, that may depend on both the previous state j and the triggering event $\tilde{e}$. We assume that whenever $p(j';j,\tilde{e})>0$, $a(j)-\{\tilde{e}\}\subseteq a(j')$, i.e., active events that have not completed processing by the time of the state change remain active after the state change. At time $t+s$, $C_e(t+s)=c_e(t)+r_{ej}s$ for $e\in a(j)-\{\tilde{e}\}$, and for $e\in a(j')-(a(j)-\{\tilde{e}\})$, T_e is selected according to F_e, independently of all else. For all other events, $C_e(t)=0$.

The process $N=(N(t):t\geq 0)$ is known as a GSMP, and the augmented process $X=(X(t):t\geq 0)$, where $X(t)=(N(t),C_e(t):e\in E)$ is an SGSMP (supplemented GSMP). The process X is Markov, but this is almost never true of N. (The process N will be Markov if and only if either all the "clock r.v.'s" are exponential r.v.'s or all the clock r.v.'s are deterministic). So we turn now to the derivation of the bounded-pointwise generator $\bar{A}$ of X.

The state space S of X is a subset of $J\times(\mathbb{R}_+)^{|E|}$, where $|E|$ denotes the cardinality of the set E, and $\mathbb{R}_+=[0,\infty)$. Let $g:S\to\mathbb{R}$ be bounded. We wish to obtain conditions on g and X which ensure that $g\in D(\bar{A})$, and determine an expression for $\bar{A}g$.

Suppose that $X(0)=x=(j,t_e:e\in E)$, and let O_t be the number of state transitions N experiences over the time interval $[0,t]$. Then

$$\begin{aligned} t^{-1}(E_xg(X(t))-g(x)) &= t^{-1}(E_x(g(X(t));O_t=0)-g(x)) \\ &+ t^{-1}E_x(g(X(t));O_t=1)+t^{-1}R(t,x) \end{aligned} \tag{1.15}$$

where $|R(t,x)|\leq\|g\|P_x(O_t\geq 2)$.

We would like to show that $R(t,\cdot)$ converges boundedly and pointwise to 0 as $t\downarrow 0$. Suppose that the *hazard rate* functions $r_e(t)\stackrel{\triangle}{=}f_e(t)/\bar{F}_e(t)$ are bounded in t and e (by M_1 say), where $\bar{F}_e(t)\stackrel{\triangle}{=}1-F_e(t)$. (The hazard rate function has the property that $r(t)\Delta t\approx P(T_e\in(t,t+\Delta t]|T_e>t)$; see Ross (1983) for details.) Suppose further that the rates $r_{ej}\leq M_2<\infty$, for every $e\in E$ and $j\in J$.

Because both the clock rates r_{ej} and the hazard rate functions for the rv's T_e are bounded, events in the GSMP may be generated by thinning independent Poisson processes with rate M_1M_2 (Lindvall, (1986); related ideas for nonhomogeneous Poisson processes may be found in Ross (1983) p. 47). Because the number of active events is bounded by $|E|$, the number of events O_t in $[0,t]$ is dominated by the number of arrivals in a Poisson process with rate $M_1M_2|E|$. Therefore $t^{-1}P_x(O_t \geq 2)$ converges to 0 and the convergence is uniform in x (Wolff (1989) p. 70), and so $t^{-1}R(t,\cdot)$ converges boundedly and pointwise to 0.

Turning now to the first term on the right-hand side of 1.15, let $x(t) = (j, t_e + r_{ej}t : e \in E)$, so that $X(t) = x(t)$ conditional on the event that $O_t = 0$. Then

$$\begin{aligned} t^{-1}(E_x(g(X(t)); O_t = 0) - g(x)) &= t^{-1}(g(x(t))P_x(O_t = 0) - g(x)) \\ &= t^{-1}g(x(t))(P_x(O_t = 0) - 1) \\ &\quad + t^{-1}(g(x(t)) - g(x)) \end{aligned} \tag{1.16}$$

Define

$$z(t,x) = P_x(O_t = 0) = \prod_{e \in a(j)} \frac{\bar{F}_e(t_e + r_{ej}t)}{\bar{F}_e(t_e)},$$

and note that our assumptions imply that z possesses a bounded continuous partial derivative with respect to t. Thus $t^{-1}(P_x(O_t = 0) - 1) = z'(\theta_t)$ for some $\theta_t \in (0,t)$. If we assume that g is continuous, it immediately follows that the first term on the right-hand side of (1.16) converges boundedly and pointwise to

$$g(x)z'(0) = -g(x) \sum_{e \in a(j)} r_{ej} r_e(t_e)$$

as $t \downarrow 0$. Furthermore, if g possesses bounded, continuous partial derivatives with respect to the clock variables t_e, then the second term in (1.16) converges boundedly and pointwise to

$$\sum_{e \in a(j)} r_{ej} \frac{\partial g}{\partial t_e}(x)$$

as $t \downarrow 0$.

A similar approach may be used to examine the second term on the right-hand side of (1.15). Under our current assumptions, this term converges boundedly and pointwise to

$$\sum_{\tilde{e}\in a(j)} r_{\tilde{e}j} r_{\tilde{e}}(t_{\tilde{e}}) \sum_{k\in J} p(k;j,\tilde{e}) g(x^{(k,\tilde{e})})$$

where $x^{(k,\tilde{e})} = (k, t'_e : e \in E)$, and

$$t'_e = \begin{cases} t_e & \text{if } e \in a(j) - \{\tilde{e}\}, \\ 0 & \text{otherwise.} \end{cases}$$

We have therefore arrived at the following conclusion. Let $\bar{A}$ be the bounded-pointwise generator of the Markov process X. The domain $D(\bar{A})$ of $\bar{A}$ includes all functions g that are continuously differentiable in the clock variables $(t_e : e \in E)$, and for which g, and $\partial g/\partial t_e$ (for every $e \in E$) are bounded. Furthermore, if $g \in D(\bar{A})$ and $x = (j, t_e : e \in E) \in S$, then

$$Ag(x) = \sum_{e\in a(j)} \left\{ r_{ej} \frac{\partial g(x)}{\partial t_e} + r_{ej} r_e(t_e) \sum_{k\in J} p(k;j,e)(g(x^{(k,e)}) - g(x)) \right\}.$$

Example 1.3 To illustrate the above ideas, we will examine an example related to the single-server queue. For a detailed discussion of this model, see Chapter 2. We specialize to the case where the interarrival and service time distributions possess bounded continuous densities, and bounded hazard rate functions. We will first describe the bounded-pointwise generator, and then move on to the extended generator.

Let $N(t)$ be the number of customers in the system at time t, so that $J = \{0, 1, 2, \ldots\}$. The event space $E = \{a, s\}$, the two events being arrivals and service completions. When $j = 0, a(j) = \{a\}$, and for $j \geq 1$, $a(j) = E$. The rates $r_{ej} = 1$, except when $e = s$ and $j = 0$, in which case $r_{s0} = 0$. The probabilities $p(k;j,e)$ are given by $p(k;j,a) = I(k = j+1)$ and $p(k;j,s) = I(k = j-1)$. Define $X(t) = (N(t), C_a(t), C_s(t))$, so that $X = (X(t) : t \geq 0)$ is a Markov process living on a state space $S \subseteq J \times \mathbb{R}^2_+$, with bounded-pointwise generator $\bar{A}$ defined as follows.

For functions $g \in D(\bar{A})$, and $(q, t_a, t_s) \in S$

$$\begin{aligned} \bar{A}g(q,t_a,t_s) &= r_a(t_a)(g(q+1,0,t_s) - g(q,t_a,t_s)) + \frac{\partial g(q,t_a,t_s)}{\partial t_a} \\ &+ r_s(t_s)(g(q-1,t_a,0) - g(q,t_a,t_s)) + \frac{\partial g(q,t_a,t_s)}{\partial t_s} \end{aligned} \tag{1.17}$$

when $q > 0$, and

$$\bar{A}g(0,t_a,0) = r_a(t_a)(g(1,0,0) - g(0,t_a,0)) + \frac{\partial g(0,t_a,0)}{\partial t_a}.$$

The domain $D(\bar{A})$ contains all functions $g : S \to \mathbb{R}$ that are continuously differentiable in both t_a and t_s, and for which g, $\partial g/\partial t_a$ and $\partial g/\partial t_s$ are bounded.

It will be instructive (and also useful for applications in Chapter 2) to discuss the extended generator A of X as well. Our approach in doing so is very similar to that used in the previous example (reflected Brownian motion). We introduce the operator Q, where Qg is defined by the right-hand side of (1.17), not only for $g \in D(\bar{A})$, but for all functions g such that the right-hand side of (1.17) is defined. We will show that for certain functions $g \in D(A)$, $Ag = Qg$ using a truncation argument.

Suppose that $g(q, t_a, t_s) = f(q)\varphi(t_a, t_s)$ for some real-valued functions f and φ. Assume that φ is continuously differentiable, and that φ, $\partial\varphi/\partial t_a$, and $\partial\varphi/\partial t_s$ are bounded. We will show that if f satisfies certain conditions, then $g \in D(A)$ and $Ag = Qg$. Define $f_n(q) = f(q)I(q < n) + f(n)I(q \geq n)$, and set $g_n(q, t_a, t_s) = f_n(q)\varphi(t_a, t_s)$. Then $g_n \in D(\bar{A})$ and so

$$\bar{M}_n(t) = g_n(X(t)) - g_n(X(0)) - \int_0^t \bar{A}g_n(X(s))\, ds$$

is a P_x martingale for each $x \in S$.

Define $T_n = \inf\{t \geq 0 : N(t) \geq n\}$, where $N(t)$ is the first component of $X(t)$, i.e., the number of customers in the system at time t. By definition, for $q < n$, $Qg(q, t_a, t_s) = \bar{A}g_n(q, t_a, t_s)$, and so

$$\bar{M}_n(t \wedge T_n) = g(X(t \wedge T_n)) - g(X(0)) - \int_0^{t\wedge T_n} Qg(X(s))ds \stackrel{\triangle}{=} M_n(t).$$

Therefore $M_n(\cdot)$ is a P_x martingale and $E_x M_n(t) = 0$ for all x and n.

The final step is to show that $E_x M_n(t) = 0$ also holds in the limit as $n \to \infty$. Recall that the number of events in $[0,t]$ is stochastically dominated by the number of events in the same period in a Poisson process with parameter θ say. Then

$$X(s \wedge T_n) \leq_{\text{stoch}} x_0 + L(s) \qquad \forall s \geq 0$$

where $x_0 = N(0)$, and $L(s)$ is the number of events by time s in the dominating Poisson process. Suppose that $|f(y)| \leq \bar{f}(y)$ where $\bar{f}$ is increasing in y, and $E\bar{f}(x_0 + L(t)) < \infty$. Then

$$|g(X(t \wedge T_n))| \leq_{\text{stoch}} \bar{f}(x_0 + L(t)),$$

and the right-hand side of this relation is P_x integrable. Thus, by dominated convergence, $E_x g(X(t \wedge T_n)) \to E_x g(X(t)) < \infty$ as $n \to \infty$.

For the integral, suppose that $|Qg(y, t_a, t_s)|$ is also bounded by $\bar{f}(y)$. Then

$$\begin{aligned} \left| \int_0^{t \wedge T_n} Qg(X(s)) ds \right| &\leq_{\text{stoch}} \int_0^{t \wedge T_n} \bar{f}(x_0 + L(s))\, ds \\ &\leq \int_0^{t \wedge T_n} \bar{f}(x_0 + L(t))\, ds \\ &\leq t\ \bar{f}(x_0 + L(t)). \end{aligned}$$

Thus, by dominated convergence,

$$E_x \int_0^{t \wedge T_n} Qg(X(s))\, ds \to E_x \int_0^t Qg(X(s))\, ds < \infty$$

as $n \to \infty$. We have therefore shown that $E_x M(t) = 0$ for all $t \geq 0$, where

$$M(t) = g(X(t)) - g(X(0)) - \int_0^t Qg(X(s))\, ds,$$

and may conclude (Karlin and Taylor (1981) p. 309) that $M(t)$ is a P_x martingale, and so $g \in D(A)$ and $Ag = Qg$.

Recall that we required the condition that both $|g|$ and $|Qg|$ be bounded by an increasing function $\bar{f}$, and that $E\bar{f}(x_0 + L(t)) < \infty$. Since $L(t) \sim \text{Poisson}(\theta t)$,

the expectation will be finite (for all t), if $\bar{f}$ increases at most exponentially; i.e., $\bar{f}(x) \leq ab^x$ for positive constants a and b. It is easy to see that $|g|$ and $|Qg|$ will be dominated by such a function if f is. We conclude that if $g(q, t_a, t_s) = f(q)\varphi(t_a, t_s)$, where f is bounded by a function that increases at most exponentially, $\varphi(t_a, t_s)$ is continuously differentiable, and φ, $\partial\varphi/\partial t_a$, and $\partial\varphi/\partial t_s$ are bounded, then $g \in D(A)$, and $Ag = Qg$.

1.3.2 The New Estimator

In the previous section we determined a class of shadow functions $h = Ag$, which may be used to obtain a consistent estimator of α of the form

$$\frac{1}{t}\int_0^t (f + Ag)(X(s))\, ds. \tag{1.18}$$

Our final consideration for this section is to determine what functions g would be effective (in terms of reducing the TAVC) of the estimator (1.18). In fact, this question can be answered in exactly the same manner as in Section 1.2.

Recall that the *optimal* estimator of the form (1.18) occurs when g solves Poisson's equation, $Ag(x) = -(f(x) - \alpha)$. In that case (1.18) is exactly α, and we have an estimator with zero variance. Just as in the CTMC case, we cannot solve this equation explicitly, but it does suggest that if g is an approximation to the solution to Poisson's equation, then the estimator (1.18) should have a TAVC that is small compared with the standard estimator (1.2). Hence, to form the AMP estimator when the processes involved are general state space Markov chains we should simply follow the procedure outlined in Section 1.2. For convenience, we refer to a function g as a surrogate function if g is an approximation to the solution to Poisson's equation $Ag(x) = -(f(x) - \alpha)$, where A is the extended generator of X.

Before concluding this chapter, we mention the situation where the process X is not Markov, but the approximating process is. To apply the AMP method, additional

variables must be adjoined to the state space of X to make it Markov. This occurs, for example, for almost any generalized semi-Markov process where the clock time distributions are not exponential. We shall see in the next chapter that in the case of estimating moments of the steady-state queue size in a single server queue, a mechanical application of the procedure for deriving the AMP estimator does not necessarily result in a useful estimator.

The problem that arises is that the proposed surrogate function does not account for the supplementary variables introduced to ensure that X is Markov. Therefore, the surrogate function suggested by a direct application of the AMP method is a poor approximation to the solution to Poisson's equation (see Section 2.5). Although there is no apparent automatic adjustment for this, our examples suggest that it is possible to either modify the surrogate function so that satisfactory variance reductions may be achieved, or approach the problem using the traditional control variates methodology (see Example 2.6), resulting in moderate variance reductions.

Chapter 2

Applications to the Single-Server Queue

2.1 Introduction

In this chapter we will demonstrate the AMP methodology by deriving new estimators for several quantities associated with the waiting time sequence and the queue length process in the standard single-server queue.

The single-server queue (GI/G/1 queue) may be described as follows. Customer 0 arrives at time $T_0 = 0$ and undergoes a service time V_0. Customer n arrives at time T_n and experiences a service time V_n. Let $U_n = T_n - T_{n-1}$ $(n \geq 1)$ be the associated interarrival times. We assume that $(U_n : n \geq 1)$ and $(V_n : n \geq 0)$ are independent sequences of independent and identically distributed (iid) random variables (rv's), and the server operates under a FIFO service discipline. Let $\lambda^{-1} > 0$ and $\mu^{-1} > 0$ denote the means of the interarrival, and service time distributions respectively, and set $\rho = \lambda/\mu$. We assume that $\rho < 1$.

If W_n denotes the waiting time in the queue that customer n experiences, then it is well known that W_n converges in distribution to a steady-state random variable W say (Asmussen (1987) p. 181). Under very mild conditions (Wolff (1989) p. 92) it is known that if $f : \mathbb{R} \to \mathbb{R}$ is a cost function, then

$$n^{-1} \sum_{k=0}^{n-1} f(W_k) \to Ef(W) < +\infty.$$

It is therefore reasonable to gauge the waiting time performance of the queueing system through measures of W. In particular, EW provides a natural measure of the average time customers spend in the queue. Similarly the variance $E(W - EW)^2$ may be of interest in determining the variability of the waiting times customers experience, and if we are interested in providing guarantees of the form "only $p\%$ of customers wait longer than x minutes" say, then the tail probabilities $P(W > w)$ will be of interest. Similar statements also apply for the number of customers in the system.

In the case where neither the arrival time distribution nor the service time distribution is exponential, there are no closed form analytical results for the moments and tail probabilities of W and the steady-state queue size Q. Approximations of these quantities may be obtained through heavy traffic theory, which asserts that under certain regularity conditions, as $\rho \uparrow 1$, W and Q may be approximated by exponential rv's (Asmussen (1992), Lemoine (1978)). Approximations to the tail probabilities may also be obtained through the theory of large deviations (Glynn and Whitt (1995)), or other forms of asymptotic analysis (Abate et al. (1995), Abate and Whitt (1994)). Large deviations theory also suggests an alternative method for estimating tail probabilities for waiting times. In contrast to the methods studied here, this alternative method performs a terminating simulation to determine the expected value of a certain "overshoot" random variable, and this leads to estimates of the waiting time tail probabilities (Asmussen (1987) p. 269).

The heavy traffic theory alluded to above actually asserts the stronger result that as $\rho \uparrow 1$, scaled versions of both the waiting time sequence and the queue size process converge weakly to reflected Brownian motions (RBM's). In this chapter we exploit

these heavy traffic approximations to develop AMP estimators.

In Section 2.2, we introduce the waiting time sequence, and the queue size process.

Next, in Section 2.3, we develop AMP estimators of the steady-state waiting time moments, and we show that they are statistically more efficient than the "standard" estimators in heavy traffic. Our results also demonstrate that in heavy traffic the AMP estimators are not as efficient as another class of estimators developed by Minh and Sorli (1983).

An AMP estimator for the waiting time tail probability is described in Section 2.4. In estimating tail probabilities for a particular instance of the GI/G/1 queue, there are certain quantities that must be computed. The analytical calculation of these quantities is somewhat clumsy, and may even be impossible (at least in closed form), so we include a brief discussion of a possible numerical approach. Then, for the M/M/1 queue, we show that in heavy traffic the AMP estimator is statistically more efficient than the standard estimator.

We address the problem of estimating steady-state queue size moments in Section 2.5. As mentioned in Chapter 1, a naive application of the AMP method results in estimators which do not perform as well as we might hope. We show how to "fix" the problem, obtaining estimators for the first two moments of the steady-state queue size, and we prove that the estimators are effective in heavy traffic.

In Section 2.6 we develop an AMP estimator of the tail probability $P(Q \geq q)$. We encounter the same problem as in 2.5, but, once again, are able to remedy the situation.

The proofs of several results are collected in Section 2.7.

It is worth noting that although the AMP estimators we present in this chapter are obtained through an approximation derived from a rigorous limit theorem, this

need not be the case in general. For example, one may choose to approximate the queue size process in the GI/G/1 queue not by a reflected Brownian motion (as obtained through heavy traffic theory), but by the M/M/1 queue size process with appropriately chosen parameters. It is not clear whether using an AMP estimator based on such an approximation would result in a gain in efficiency, but nevertheless, such an approach would be easily implemented.

As a final note, we have not developed estimators of the moments and tails of the workload process (or virtual waiting time) (the time the server would have to work to empty the system of customers). This problem can be handled in a similar manner to the queue size process.

2.2 The Waiting Time Sequence and Queue Size Process

In this section we introduce the two stochastic processes related to the GI/G/1 model that we will study in this chapter: the waiting time sequence and the queue size process.

2.2.1 The Waiting Time Sequence

Let W_n denote the waiting time in the queue (ie. not counting service time) of the nth customer. If we define the (iid) sequence $(X_n : n \geq 1)$ by $X_n = V_{n-1} - U_n$, then W_n satisfies the *Lindley Recursion* (Asmussen (1987) p. 79)

$$W_{n+1} = [W_n + X_{n+1}]_+, \tag{2.1}$$

where $[x]_+$ is the non-negative part of x. The stochastic process $(W_n : n \geq 0)$ is therefore a Markov chain on $\mathbb{R}_+ \stackrel{\triangle}{=} [0, \infty)$. Since $\rho < 1$, $W_n \Rightarrow W$ as $n \to \infty$, where

W is the steady-state waiting time (Asmussen (1987) p. 181).

Define the extended generator A for the Markov chain $(W_n : n \geq 1)$ as follows. For functions $f : \mathbb{R}_+ \to \mathbb{R}$ and $x \geq 0$, set

$$\begin{aligned} Af(x) &= E_x f(W_1) - f(x) \\ &= Ef([x + X_1]_+) - f(x). \end{aligned} \tag{2.2}$$

Of course, we restrict this definition to the domain $D(A)$ consisting of all functions f for which $E_x f(W_1)$ is finite for all $x \geq 0$. If we extend the domain of f to $\mathbb{R}$, we may write

$$\begin{aligned} Af(x) &= Ef(x + X_1) - f(x) + f(0)P(x + X_1 < 0) \\ &- E(f(x + X_1); x + X_1 < 0), \end{aligned} \tag{2.3}$$

being careful to extend the definition of f in such a way that the expectations in (2.3) exist. This second way of writing the generator will prove useful in the development of the AMP estimator of the moments of the steady-state waiting time.

2.2.2 The Queue Size Process

Let $Q(t)$ denote the number of customers in the system at time t (including the customer in service, if any). Unless the interarrival and service time distributions are exponential, $(Q(t) : t \geq 0)$ is *not* a Markov process. We therefore consider the process $X = (X(t) : t \geq 0)$, where $X(t) = (Q(t), A(t), S(t))$, $Q(t)$ is defined as before, and $A(\cdot)$ and $S(\cdot)$ are processes associated with the interarrival and service times which ensure that X is Markov. Specifically, we let $A(\cdot)$ be the age process associated with the renewal arrival process, and let $S(\cdot)$ be the length of time that the current customer has been in service if $Q(t) > 0$, and (arbitrarily) 0 otherwise.

Let F_a and F_s be the distribution functions of the interarrival time and service time distributions respectively. In analyzing the queue size process, we shall assume

that both F_a and F_s possess bounded and continuous densities f_a and f_s respectively, so that we may define hazard rate functions for these distributions. Let

$$r_a(t) \triangleq \frac{f_a(t)}{\bar{F}_a(t)} \quad \text{and} \quad r_s(t) \triangleq \frac{f_s(t)}{\bar{F}_s(t)},$$

where $\bar{F}_a(t) \triangleq 1 - F_a(t)$ and $\bar{F}_s(t) \triangleq 1 - F_s(t)$.

If U_1 has a non-lattice distribution, then $Q(t) \Rightarrow Q$ as $t \to \infty$, where Q is the steady-state queue size (Asmussen (1987) p. 193). Because we are assuming that both F_a and F_s have densities, this is clearly the case in our setup.

Let $g : S \to \mathbb{R}$ where $S \subseteq J \times \mathbb{R}_+^2$ is the state space of X. Recall from Example 1.3 that if r_a and r_s are bounded, and $g \in D(\bar{A})$, where $\bar{A}$ is the bounded pointwise generator of X, then for $(q, t_a, t_s) \in S$,

$$\begin{aligned} \bar{A}g(q,t_a,t_s) &= r_a(t_a)(g(q+1,0,t_s) - g(q,t_a,t_s)) + \frac{\partial g(q,t_a,t_s)}{\partial t_a} \\ &+ r_s(t_s)(g(q-1,t_a,0) - g(q,t_a,t_s)) + \frac{\partial g(q,t_a,t_s)}{\partial t_s} \end{aligned} \tag{2.4}$$

when $q > 0$, and

$$\bar{A}g(0,t_a,0) = r_a(t_a)(g(1,0,0) - g(0,t_a,0)) + \frac{\partial g(0,t_a,0)}{\partial t_a}.$$

The domain $D(\bar{A})$ contains all functions $g : S \to \mathbb{R}$ that are continuously differentiable in both t_a and t_s, and for which g, $\partial g/\partial t_a$ and $\partial g/\partial t_s$ are bounded.

Furthermore, we showed that if $g(q, t_a, t_s) = f(q)\varphi(t_a, t_s)$, where f is bounded by a function that increases at most exponentially and $\varphi \in D(\bar{A})$, then $g \in D(A)$ where A is the extended generator of X, and Ag is given by 2.4.

2.3 Waiting Time Moments

In this section we will develop estimators of the steady-state waiting time moments, using the AMP method applied with a heavy traffic approximation to the waiting

time sequence. First we review the relevant heavy traffic theory, and then derive the AMP estimators based on the heavy traffic approximation. Next, we discuss competing estimators that have been proposed. Finally, we analyze the performance of the various estimators through both theoretical results, and numerical examples.

Of course, before we consider the problem of estimating the kth moment of the steady-state waiting time, we must be sure that the moment exists. Kiefer and Wolfowitz (1956) showed that if $\rho < 1$ and $EV_0^{k+1} < \infty$ then $EW^k < \infty$, where W is the steady-state waiting time rv.

We define the *standard estimator* of the kth moment of the steady-state waiting time as

$$\frac{1}{n}\sum_{j=0}^{n-1} W_j^k. \tag{2.5}$$

To show that this estimator is consistent, and satisfies a central limit theorem, we appeal to the regenerative structure of the GI/G/1 queue. The regeneration times coincide with the epochs when a customer arrives to an empty queue. Define $C_0 = 0$, and for $j \geq 1$, $C_j = \inf\{n > C_{j-1} : W_n = 0\}$. Then the sequence $(C_j : j \geq 0)$ defines regeneration times for the waiting time sequence $(W_j : j \geq 0)$. Since $\rho < 1$, $EC_1 < \infty$ (Asmussen (1987) p. 169). It is a standard result from regenerative process theory that the estimator for EW^k in (2.5) is consistent (Wolff (1989) p. 92). To show that the estimator satisfies a CLT, let $f : S \to \mathbb{R}$ be a given function, and define the sequence of random variables

$$Y_n(f) = \sum_{j=C_{n-1}}^{C_n-1} f(W_j), \qquad n \geq 1.$$

If $0 < E[(Y_1(f_k) - C_1 EW^k)^2] < \infty$ (where $f_k(x) = x^k$), then (Glynn and Iglehart (1993)) there exists a finite constant σ_k^2 such that

$$\sqrt{n}\left(\frac{1}{n}\sum_{j=0}^{n-1} W_j^k - EW^k\right) \Rightarrow \sigma_k N(0,1).$$

As discussed in Chapter 1, the time-average variance constant (TAVC) σ_k^2 appearing in this CLT provides a natural means of evaluating the absolute error in the estimator (2.5). (The coefficient of variation would be more appropriate if one is interested in the relative error.) Typically, σ_k^2 is of the order $(1-\rho)^{-2k-2}$ (see Theorem 2.2). Clearly then, in heavy traffic this estimator will require very large run lengths to return reliable estimates, and an improvement (in the form of an estimator with a lower TAVC) is highly desirable.

2.3.1 Heavy Traffic Theory

We will now summarise the relevant heavy traffic theory for the waiting time sequence in the GI/G/1 queue. Heavy traffic results are usually presented in the form of limit theorems, and so we embed our particular GI/G/1 queue in a parameterized family of such queues.

Let $(\bar{V}_n : n \geq 0)$ and $(\bar{U}_n : n \geq 1)$ be independent sequences of iid rv's, with $E\bar{V}_0 = E\bar{U}_1 = \mu^{-1}$. Consider now a family of queueing systems, defined in terms of these basic building blocks, parameterized by ρ ($\rho \leq 1$).

The ρth system consists of the sequences $(V_n(\rho) : n \geq 0)$ and $(U_n(\rho) : n \geq 1)$, where $V_n(\rho) \stackrel{\triangle}{=} \bar{V}_n$ and $U_n(\rho) \stackrel{\triangle}{=} \bar{U}_n/\rho$. The ρth system has a customer arrival rate of $\mu\rho \stackrel{\triangle}{=} \lambda$ and a traffic intensity ρ, justifying the notation. Let $X_n(\rho) \stackrel{\triangle}{=} V_{n-1}(\rho) - U_n(\rho)$ and let $(W_n(\rho) : n \geq 0)$ denote the waiting time sequence for the ρth system, assuming that the 0th customer arrives at time $T_0 = 0$ to an empty queue and experiences a service time $V_0(\rho)$. We will usually suppress the dependence of variables on ρ to improve readability, (as we have already done with λ), so that the rv's in the ρth system are written U_n, V_n, X_n and W_n. We shall write U, V, and X for generic rv's that are equal in distribution to U_1, V_0 and X_1, and W for the steady-state waiting time rv in the ρth system.

Let $\sigma^2 = \text{var}\,\bar{U}_1 + \text{var}\,\bar{V}_1$, $\kappa^2 = \sigma^2\lambda^2$ and $\omega^2 = \sigma^2\lambda$. Note that σ^2 does not depend on ρ, while κ^2 and ω^2 do. Define

$$R^{(\rho)}(t) \triangleq \frac{(1-\rho)}{\omega^2} W_{\left\lfloor \frac{\kappa^2 t}{(1-\rho)^2} \right\rfloor}(\rho),$$

and let $R = (R(t) : t \geq 0)$ be a reflected (or regulated) Brownian motion (RBM) with drift -1, unit infinitesimal variance, and starting state $R(0) = 0$. Adapting a result from Asmussen (1992) we obtain the following theorem.

Theorem 2.1 *Suppose that* $E\bar{U}_1^3 + E\bar{V}_0^3 < \infty$. *Then* $R^{(\rho)} \Rightarrow R$ *in* $D[0,\infty)$ *as* $\rho \to 1$, *where* $D[0,\infty)$ *is the space of real-valued right-continuous functions on* $[0,\infty)$ *that have left limits everywhere.*

Theorem 2.1 suggests the following approximation for the process $(W_n(\rho) : n \geq 0)$:

$$\begin{aligned} W_n(\rho) &\overset{\mathcal{D}}{\approx} \frac{\omega^2 R(n(1-\rho)^2/\kappa^2)}{1-\rho} \\ &\overset{\mathcal{D}}{=} \tilde{W}(n) \end{aligned} \tag{2.6}$$

where $\tilde{W}(\cdot)$ is a reflected Brownian motion with drift $-(1-\rho)/\lambda$ and diffusion coefficient (or infinitesimal variance) σ^2.

2.3.2 Solving Poisson's Equation for the RBM

Equation (2.6) furnishes the required approximating Markov process $\tilde{W}(\cdot)$. The first step in applying the AMP method is to solve Poisson's equation for $\tilde{W}(\cdot)$.

It is a standard result (see for example, Harrison (1990) p. 94) that the stationary distribution ν of an RBM with drift $-a < 0$ and infinitesimal variance b is exponential with parameter $2a/b$. We are interested in estimating the moments of the steady-state waiting time, so that the cost functions we will consider are of the form $f(x) = x^k$,

where k is a non-negative integer. Since $\tilde{W}$ is an RBM, the centering constant for $\tilde{W}$ is the kth moment of an exponential rv. Thus

$$E_\nu f(\tilde{W}(0)) = \frac{k!(\lambda\sigma^2)^k}{(2(1-\rho))^k}.$$

From Example 1.2, the extended generator $\tilde{A}$ of $\tilde{W}(t)$ (which is an RBM) is given by

$$\tilde{A}g = \frac{\sigma^2}{2}g''(x) - \frac{1-\rho}{\lambda}g'(x).$$

The domain $D(\tilde{A})$ contains the set of twice continuously differentiable functions g with $g'(0) = 0$, where g is either bounded, or a polynomial.

We are now in a position to write down Poisson's equation for $\tilde{W}$, and functions of the form $f(x) = x^k$. The equation is

$$\frac{\sigma^2}{2}g''(x) - \frac{1-\rho}{\lambda}g'(x) = -(x^k - \frac{k!(\lambda\sigma^2)^k}{(2(1-\rho))^k}) \qquad g(0) = g'(0) = 0. \tag{2.7}$$

To understand the first boundary condition, note that if g is a solution to Poisson's equation, then so is $g(x) + c$ for any constant c, so we force uniqueness by imposing the condition that $g(0) = 0$.

Equation (2.7) is a second order ordinary differential equation with constant coefficients, and thus is straight-forward to solve. The solution is

$$g(x) = \sum_{j=2}^{k+1} \left(\frac{\lambda\sigma^2}{2(1-\rho)}\right)^{k+1-j} \frac{\lambda k!}{(1-\rho)j!}x^j. \tag{2.8}$$

Thus we have determined the required surrogate function g. The next step in identifying the AMP estimator is to calculate $Ag(x)$, where A is the generator of the waiting time sequence. So that we may investigate the heavy traffic behaviour of our estimators, we will compute $A_\rho g$, where A_ρ is the generator of the ρth system. Since

g is a polynomial, we must first find an expression for $Af_k(x)$, where $f_k(x) = x^k$. From (2.3),

$$A_\rho f_k(x) = E(x + X_1(\rho))^k - x^k - h_k(x;\rho), \tag{2.9}$$

where $h_k(x;\rho) = E((x + X_1(\rho))^k; x + X_1(\rho) < 0)$. We will usually write $h_k(x)$ for $h_k(x;\rho)$.

2.3.3 The First Moment

Let us first consider the estimator for the first moment (i.e., mean) of the steady-state waiting time. For this case, the cost function is $f(x) = x$, and from (2.8), the solution to Poisson's equation (for the approximating RBM $\tilde{W}$) is

$$g(x) = \frac{\lambda x^2}{2(1-\rho)}. \tag{2.10}$$

Then, from (2.9), we compute

$$Ag(x) = \frac{EX^2}{-2EX} - x + \frac{h_2(x)}{2EX}$$

so that the AMP estimator (based on $f(x) + Ag(x)$) is given by

$$\alpha_1(n) = \frac{EX^2}{-2EX} + \frac{1}{2nEX}\sum_{j=0}^{n-1} h_2(W_j). \tag{2.11}$$

The first term in (2.11) is the typical heavy traffic approximation for the expected waiting time (see for example, Asmussen (1987) p. 200). The second term represents a correction to the heavy traffic approximation.

To see why (2.11) should perform better than the standard estimator, we reason as follows. Suppose the interarrival times are bounded by some (positive) constant K. Then the increment rv X is bounded below by $-K$. Thus, $h_2(x) = 0$ for $x > K$, which means that the AMP estimator is only affected when waiting times are "near

the boundary" (i.e., small). In heavy traffic, the waiting times spend very little time near the boundary, and so we would expect that the AMP estimator performs well in this realm. This intuitive argument is supported by Theorem 2.3 below, which proves that under certain conditions on the interarrival and service time distributions, the AMP estimator performs (far) better than the standard estimator in heavy traffic. Before stating this theorem, we examine the form of the AMP estimator for moments higher than the mean.

2.3.4 Higher Moments

We turn now to estimating the kth moment of the steady-state waiting time, where $k > 1$. Equation (2.8) furnishes the required surrogate function. We now need to compute the shadow function Ag. Define

$$c_k(j) \stackrel{\triangle}{=} \left(\frac{\lambda\sigma^2}{2(1-\rho)}\right)^{k+1-j} \frac{\lambda k!}{(1-\rho)j!}, \tag{2.12}$$

for $2 \leq j \leq k+1$ and $c_k(j) = 0$ otherwise, so that

$$g(x) = \sum_{j=2}^{k+1} c_k(j) f_j(x).$$

(Recall that $f_j(x) = x^j$.) Utilizing (2.9), we find that

$$\begin{aligned} A_\rho g(x) &= \sum_{j=2}^{k+1} c_k(j)(E(x+X)^j - x^j - h_j(x)) \\ &= \sum_{j=2}^{k+1} c_k(j) \sum_{i=1}^{j} \binom{j}{i} EX^i x^{j-i} - \sum_{j=2}^{k+1} c_k(j) h_j(x) \\ &= \sum_{m=0}^{k} \sum_{n=1}^{k+1-m} c_k(m+n) \binom{m+n}{n} EX^n x^m - \sum_{j=2}^{k+1} c_k(j) h_j(x) \end{aligned}$$

The coefficient of x^k simplifies to -1, and so

$$\psi_k(x) = f_k(x) + Ag(x)$$

$$= -\sum_{j=2}^{k+1} c_k(j)h_j(x) + \sum_{m=0}^{k-1}\sum_{n=1}^{k+1-m} c_k(m+n)\binom{m+n}{n}EX^n x^m. \quad (2.13)$$

An estimator based on (2.13) of the form $n^{-1}\sum_{i=0}^{n-1}\psi_k(W_i)$ will implicitly estimate the mth moments ($m \leq k$) of the waiting time (since x^m appears in $\psi_k(x)$). It seems reasonable then, to recursively replace x^m by $\psi_m(x)$, so that we obtain an estimating function of the form

$$\phi_k(x) = b(k,1) + \sum_{j=2}^{k+1} b(k,j)h_j(x) \quad (2.14)$$

where $b(k,j)$ $(1 \leq j \leq k+1)$ are coefficients that are constructed from the recursion. We now make explicit (2.14) by providing two examples.

Example 2.1 *For $k = 1$ we have*

$$\phi_1(x) = \psi_1(x) = -c_1(2)h_2(x) + c_1(1)EX + c_1(2)EX^2.$$

But $c_1(1) = 0$ (by definition) so that the estimator based on $\phi_1(x)$ is exactly (2.11).

Example 2.2 *Taking $k = 2$,*

$$\begin{aligned}\psi_2(x) &= -c_2(2)h_2(x) - c_2(3)h_3(x) + \sum_{j=1}^{3} c_2(j)EX^j \\ &+ c_2(2)\binom{2}{1}EXx + c_2(3)\binom{3}{2}EX^2x.\end{aligned}$$

Replacing x in this expression with $\psi_1(x)$ gives (after simplification)

$$\phi_2(x) = \frac{EX^3}{-3EX} + \frac{1}{2}\left(\frac{EX^2}{EX}\right)^2 + \frac{h_3(x)}{3EX} - \frac{EX^2h_2(x)}{2(EX)^2}.$$

The resulting estimator for EW^2 is then given by $n^{-1}\sum_{i=0}^{n-1}\phi_2(W_i)$.

As was the case for the first moment, we see that a sufficient condition for these estimators to be consistent is that $EV_0^{k+2} < \infty$.

2.3.5 Other Estimators

In addition to the standard estimators of moments of the steady-state waiting time, other estimators have been proposed. For example, Asmussen (1990), proposed a technique based upon exponential twisting of the interarrival and service time distributions. (The technique presented there generalizes to other functionals of the steady-state waiting time as well). Another estimator was proposed by Minh and Sorli (1983), which is based upon identities due to Marshall (1968). To the best of our knowledge, the Minh-Sorli estimators are the most efficient (statistically) estimators of steady-state waiting time moments in heavy traffic, so that we will be interested in comparing their performance to the performance of the AMP estimators. Therefore, we include a brief discussion of the Minh-Sorli estimators next.

Let $Q(t)$ be the number of customers in the system at time t, and let T_k $(k \geq 0)$ be the the times when a customer arrives to an empty system, ie. $T_0 = 0$, and for $k \geq 1$, $T_k = \inf\{t > T_{k-1} : Q(t) = 1, Q(t-) = 0\}$. For $k \geq 1$, let $T'_k = \sup\{t < T_k : Q(t) = 0, Q(t-) = 1\}$, so that T'_k is the time that the kth busy period ends. Define the kth idle period I_k by $I_k = T_k - T'_k$, so that $(I_k : k \geq 1)$ is an iid sequence of rv's.

Marshall (1968) related the steady-state waiting time distribution to the idle period distribution by proving an identity relating the Laplace Steiltjes transforms (LST's) of the steady-state waiting time $(\tilde{w}(s) \stackrel{\triangle}{=} Ee^{-sW})$, the idle period $(\tilde{h}(s))$, and the increment random variable X_1 $(\tilde{k}(s))$. We remark that this identity does not easily generalize to the case for multiple servers. In particular, Marshall showed that

$$(1 - \tilde{k}(s))\tilde{w}(s) = a_0(1 - \tilde{h}(-s)), \ \mathrm{Re}(s) = 0, \tag{2.15}$$

for a certain constant a_0. We may compute the value of the constant a_0 as follows. Differentiating both sides of (2.15), and evaluating at $s = 0$, we obtain $EX_1 = -a_0 EI_1$. Now, if $C = C_1$ is the number of customers served in the first busy period, we have the relation $I_1 = -\sum_{i=1}^{C} X_i$. Wald's equation then gives

$$EI_1 = -ECEX_1. \tag{2.16}$$

Thus, $a_0 = 1/EC$.

By further differentiation we may then obtain relations between the moments of the waiting time and idle time. For example, if π is the distribution of the steady-state waiting time then

$$E_\pi W = \frac{EX_1^2}{-2EX_1} - \frac{EI_1^2}{2EI_1}. \tag{2.17}$$

Of course, the first term can be easily computed. Minh and Sorli (1983) proposed that an estimator for $E_\pi W_0$ be constructed by estimating the second term in (2.17). In particular, they proposed that $E_\pi W_0$ be estimated by

$$\alpha_n \stackrel{\triangle}{=} \frac{EX_1^2}{-2EX_1} - \frac{\sum_{i=1}^{N_n} I_i^2}{2\sum_{i=1}^{N_n} I_i},$$

where $N_n \stackrel{\triangle}{=} |\{k \leq n : W_k = 0\}| - 1$ is the number of regenerative cycles completed by time n.

This idea generalizes to higher moments, with the resulting estimators depending only on moments of the increment rv's X_i, and the idle times I_i. To obtain an estimator for $E_\pi W_0^k$, the idea is to differentiate (2.15) $k+1$ times, substitute $s = 0$, rearrange, and then recursively substitute the lower order Minh-Sorli estimators wherever $E_\pi W^j$ $(j < k)$ appears.

2.3.6 Performance of AMP Estimators

In this section we present three theorems which establish the order of the TAVC's for the standard, AMP and Minh-Sorli estimators of the kth moment of the steady-state waiting time. We also report results for two examples.

Theorem 2.2 *Suppose that* $E\bar{U}_1^{8k+9} < \infty$ *and* $E\bar{V}_0^{8k+9} < \infty$. *Then the standard*

estimator (2.5) of EW^k satisfies the CLT

$$n^{1/2}\left(\frac{1}{n}\sum_{j=0}^{n-1} W_j^k - EW^k\right) \Rightarrow \sigma_k N(0,1)$$

as $n \to \infty$. Furthermore, its TAVC $\sigma_k^2 = \sigma_k^2(\rho) = O((1-\rho)^{-2k-2})$ as $\rho \uparrow 1$.

For a proof of this result, see Theorem 5.1 of Asmussen (1992).

Theorem 2.3 *Suppose that $E\bar{V}_0^{k+2} < \infty$, $\bar{U}_1$ has an exponential tail (see Section 2.7.1 for a definition), and $\bar{U}_1$ has a continuous distribution. Consider the estimator*

$$\alpha_k(n) = \frac{1}{n}\sum_{i=0}^{n-1} \phi_k(W_i)$$

of the kth moment of the steady-state waiting time, where ϕ_k is defined via (2.14). The estimator $\alpha_k(n)$ satisfies the CLT

$$n^{1/2}(\alpha_k(n) - EW^k) \Rightarrow \sigma_k N(0,1)$$

as $n \to \infty$. Furthermore, its TAVC $\sigma_k^2 = \sigma_k^2(\rho) = O((1-\rho)^{-2k})$ as $\rho \uparrow 1$.

The proof is based on regenerative process arguments, and is given in Section 2.7.1.

Theorem 2.4 *Suppose that $E(\bar{U}_1^{2k+3} + \bar{V}_0^{2k+3}) < \infty$, and $\bar{U}_1$ has a continuous distribution. If $\alpha_k(n)$ is the Minh-Sorli estimator of EW^k, then $\alpha_k(n)$ satisfies the CLT*

$$n^{1/2}(\alpha_k(n) - EW^k) \Rightarrow \sigma_k N(0,1)$$

as $n \to \infty$. Furthermore, its TAVC $\sigma_k^2 = \sigma_k^2(\rho) = O((1-\rho)^{-2k+1})$ as $\rho \uparrow 1$.

The proof of this result is given in Section 2.7.1.

The above theorems allow us to conclude that in heavy traffic, the estimators, in order of efficiency, are Minh-Sorli, AMP, and the standard estimator. It is conceivable however, that this ordering is only valid for ρ *very* close to 1, so that it is of interest to examine a few examples.

Example 2.3 Our first example is the M/M/1 queue. In this case, it is easy to compute the mean steady-state waiting time, but it serves as an example where the variance constants can be explicitly computed. Law (1975) (see (2.1) in that paper) allows us to compute the TAVC for the standard estimator as

$$\frac{\rho(2+5\rho-4\rho^2+\rho^3)}{\mu^2(1-\rho)^4}. \tag{2.18}$$

Now consider the AMP estimator. For the M/M/1 queue,

$$h_2(x) = 2e^{-\lambda x}/(\lambda^2(1+\rho)).$$

Thus, from (2.11), the AMP estimator is based on the function

$$\phi_1(x) = \frac{EX^2}{-2EX} - \frac{e^{-\lambda x}}{\lambda(1-\rho^2)}.$$

To compute the TAVC for an estimator based on this function, we can ignore the constant term. Let $f(x) = e^{-\lambda x}/(\lambda(1-\rho^2))$. If π is the stationary distribution for the customer waiting time, given by

$$\pi[0,x] = 1 - \rho e^{-\mu(1-\rho)x},$$

$\pi f = 1/\lambda$. A solution g to Poisson's equation for this f can be computed from the results in Glynn (1994) as $-x/(1-\rho)$. Results from Glynn (1994) also allow us to conclude that the TAVC is given by

$$\begin{aligned} \sigma_1^2 &= \int_{[0,\infty)} (f(x)-\pi f)(2g(x)-(f(x)-\pi f))\,\pi(dx) \\ &= \frac{\rho^3(2+9\rho+5\rho^2)}{\lambda^2(1+\rho)^3(1-\rho)^2}. \end{aligned} \tag{2.19}$$

Notice the factors $(1-\rho)^{-4}$ and $(1-\rho)^{-2}$ appearing in (2.18) and (2.19), demonstrating the validity of Theorem 2.3. Furthermore, a direct computation shows that (2.19) is smaller than (2.18) for all $0 < \rho < 1$, and not just for ρ near 1, so that the AMP estimator is statistically more efficient than the standard estimator for every $\rho \in (0,1)$.

From (2.46), the TAVC for the Minh-Sorli estimator is

$$\frac{E(I_1^2 - 2rI_1)^2 EC}{4(EI_1)^2},$$

where $r = EI_1^2/EI_1$. Now I_1 is an exponential(λ) rv, and EC may be found from (2.16) and the fact that $EI_1 = \lambda^{-1}$. Hence $EC = (1-\rho)^{-1}$ (this same result holds for all M/G/1 queues). Therefore, the TAVC above simplifies to $2\lambda^{-2}(1-\rho)^{-1}$. The minimum value of ρ for which this is smaller than (2.19) is approximately $\rho = 0.67$.

Example 2.4 There are no closed form results for the U/U/1 queue, (where both the interarrival and service time distributions are uniform), so to compare the various estimators, we simulated this system. In particular, the service time distributions we chose were uniform on $[0,2]$, and the interarrival time distributions were uniform on $[0, 2/\rho]$, so that the traffic intensity was ρ. It is straight-forward to calculate

$$\frac{EX^2}{-2EX} = \frac{2\rho^2 - 3\rho + 2}{3\rho(1-\rho)},$$

and

$$h_2(x) = \begin{cases} 0 & \text{if } x > 2/\rho, \\ \rho(2/\rho - x)^4/48 & \text{if } 2/\rho - 2 \le x \le 2/\rho, \\ \rho((2/\rho - x)^4 - (2/\rho - 2 - x)^4)/48 & \text{if } 0 \le x < 2/\rho - 2. \end{cases}$$

The estimator of the steady-state waiting time is then given by (2.11), using the above expressions. We simulated 2000 busy cycles for various values of ρ, and computed the standard estimate, AMP estimate, the Minh-Sorli estimate, and estimates of their TAVC's. In view of our results on the heavy traffic behaviour of the AMP and

ρ	Standard			AMP			Minh-Sorli		
	Est	TAVC	CI	Est	TAVC	CI	Est	TAVC	CI
0.1	0.035	0.0014	2E-4	0.038	0.037	0.005	0.060	5.70	0.06
0.3	0.13	0.026	0.003	0.13	0.16	0.02	0.14	0.71	0.01
0.5	0.30	2	0.4	0.30	0.36	0.02	0.30	0.37	0.006
0.7	0.70	7	1	0.71	1.18	0.09	0.71	0.32	0.006
0.9	2.9	400	100	2.9	13	1	2.9	0.59	0.01
0.95	6.1	12000	9000	6.2	51	4	6.2	1.01	0.02
0.99	33	6E+6	4E+6	33	1100	100	33	4.8	0.1

Table 2.1: Simulation results for estimating the mean waiting time in the U/U/1 queue.

standard estimators, it is clear that as $\rho \uparrow 1$, we should simulate more cycles to obtain equally accurate estimates. However, the accuracy obtained through 2000 cycles was sufficient for our purposes.

The results are given in Table 2.1, and the TAVC's are plotted against ρ in Figure 2.1. The first number in each column of Table 2.1 is the point estimate, the second is an estimate of the TAVC, and the third number provides a 95% confidence interval for the TAVC. For example, when $\rho = 0.9$, a 95% confidence interval for the TAVC of the AMP estimator is 13 ± 1.

For $\rho > 0.5$, the estimators in order of statistical efficiency are Minh-Sorli, AMP, and the standard estimator, while for $\rho < 0.5$, the Minh-Sorli estimator is the least efficient.

2.4 Waiting Time Tails

In this section we consider the problem of estimating the quantity $P(W > w)$, where W is the steady-state waiting time rv. This problem has received a great deal of attention in the literature, partly due to the importance of the "dual" problem, of

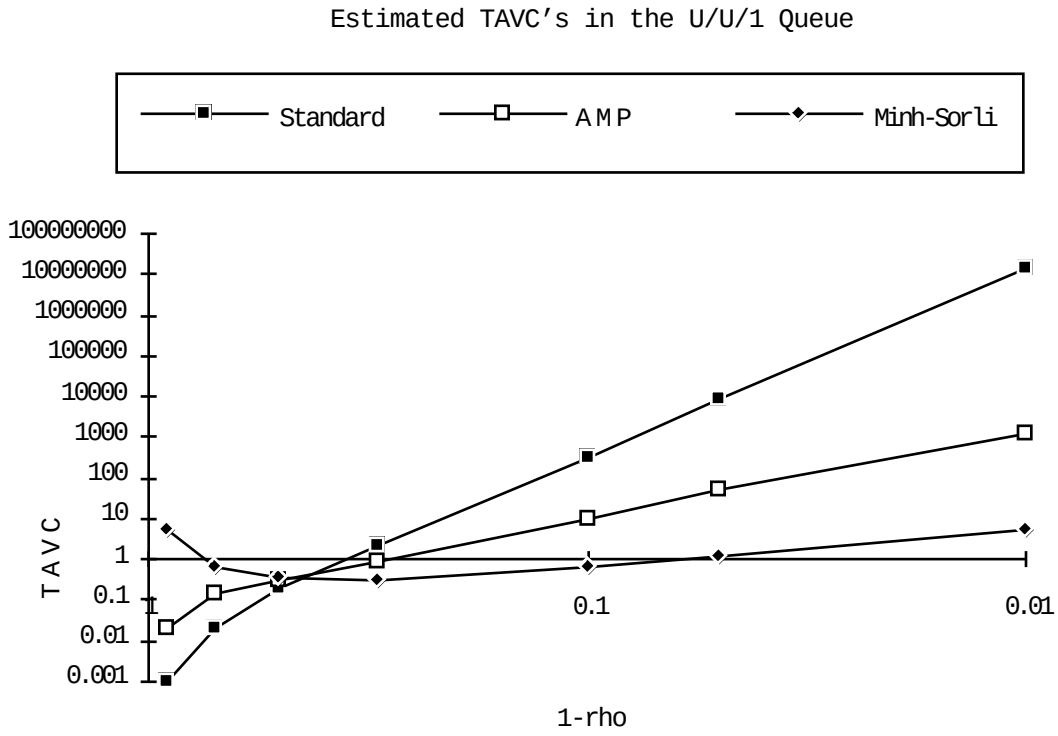

Figure 2.1: Log/log plot of TAVC estimates for the mean waiting time in the U/U/1 queue.

determining quantiles of the rv W; i .e., determining the value w such that $P(W > w) = p$ for some probability p .

Recently Glynn and Torres (1996), and Torres and Glynn (1996) have investigated the simulation run lengths required to estimate tail probabilities in the single-server queue. They identify two very different behaviours depending on whether the tail probability is estimated parametrically, or nonparametrically . Parametric estimation corresponds to the situation where a particular structure is assumed of the system (for instance, assuming a queueing system is an M/M/1 queue), so that an expression for the tail probability in terms of certain parameters can be obtained, and then the *parameters* of the model are estimated, thereby indirectly estimating the tail probability. The nonparametric case corresponds to estimating the tail probability without assuming such structure . In the parametric case, their asymptotic analysis

(in w) suggests that in order for the relative error in the estimate to be negligible, the run length must be large compared with w^2, while in the nonparametric case, the run lengths must increase exponentially in w.

We shall consider the nonparametric case, as there are no closed form expressions for tail probabilities (in the case of general interarrival and service time distributions) that would admit a parametric estimation. Specifically, we define the *standard estimator* as

$$\frac{1}{n}\sum_{k=0}^{n-1} I(W_k > w), \tag{2.20}$$

which is based on the *estimating function* $f(x) \stackrel{\triangle}{=} I(x > w)$. We shall derive an AMP estimator, again using a heavy traffic approximation to the waiting time sequence.

Recall from Section 2.3.1 that the waiting time sequence $(W_n : n \geq 0)$ may be approximated by an RBM $\tilde{W}$ with drift $\gamma \stackrel{\triangle}{=} -(1-\rho)/\lambda$, diffusion coefficient (or infinitesimal variance) $\sigma^2 \stackrel{\triangle}{=} \text{var}\,\bar{U}_1 + \text{var}\,\bar{V}_1$, and initial state $\tilde{W}(0) = 0$.

Glynn and Torres (1996) determine a solution to Poisson's equation (for the RBM)

$$\frac{\sigma^2}{2}g''(x) + \gamma g'(x) = -(I(x > w) - e^{-\theta w}) \qquad g(0) = g'(0) = 0,$$

where $\theta \stackrel{\triangle}{=} 2\gamma/\sigma^2$ is the parameter (i.e., the inverse of the mean) of the stationary exponential distribution for $\tilde{W}$. They obtain

$$g(x) = \begin{cases} \frac{e^{-\theta w}}{\gamma\theta}(e^{\theta x} - \theta x - 1) & \text{if } 0 \leq x \leq w, \\ c + \frac{1-e^{-\theta w}}{\gamma}(x - w) & \text{if } x > w, \end{cases} \tag{2.21}$$

where

$$c \stackrel{\triangle}{=} \frac{1 - e^{-\theta w} - \theta w e^{-\theta w}}{\gamma\theta}.$$

The next step is to compute the shadow function Ag, where A is the generator of the waiting time sequence defined by (2.2). We find that

$$Ag(x) \;=\; Eg([x + X]_+) - g(x)$$

$$= E(g(x+X); 0 \leq x+X \leq w) + E(g(x+X); x+X > w) - g(x). \quad (2.22)$$

The AMP estimator is then

$$\alpha_2(t) = \frac{1}{n}\sum_{k=0}^{n-1} I(W_k > w) + Ag(W_k).$$

2.4.1 Some Implementation Issues

The expected values in (2.22) may be somewhat difficult to compute analytically. Even for the M/M/1 queue a nonnegligible amount of effort is required. This impediment to implementation may be avoided if we are willing to calculate the appropriate quantities numerically.

Consider again the function underlying the AMP estimator,

$$I(x > w) - g(x) + Eg([x+X]_+).$$

Clearly, the only difficult term to compute is the last one, and so it is this term that we focus on computing numerically.

Let $\varphi(x) \stackrel{\triangle}{=} Eg([x+X]_+)$. Then φ is increasing and convex. To see why, note that both $g(x)$ and $[x+y]_+$ are increasing in x, so that $g([x+y]_+)$ is increasing in x. Therefore φ is increasing in x. Furthermore, it is a standard result in convex analysis (see for example, Exercise 3.12 of Bazaraa, Sherali and Shetty (1993) p. 119) that if both $h : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ are convex functions, and g is increasing, then $g(h(x))$ is also convex. Applying this result twice, we see that $g([x+y]_+)$ is a convex function, and it immediately follows that φ is convex. The convexity of φ then implies that φ is continuous, and differentiable almost everywhere (Bazaraa, Sherali and Shetty pp. 81–83).

Now that we have derived some of the properties of the function φ, consider how

a particular value $\varphi(x)$ may be computed. Our aim is merely to show that the numerical approach is feasible, so we will restrict ourselves to suggesting two reasonable alternatives. In any implementation, it would obviously be important to seek out many alternatives, and evaluate each of them carefully in terms of computational cost, and accuracy.

One approach would be to replace the expectation $\varphi(x)$ by a r.v. with the same mean. For example, instead of using an estimator of the form

$$n^{-1} \sum_{k=0}^{n-1} I(W_k > w) - g(W_k) + \varphi(W_k)$$

one might use the same estimator with $\varphi(W_k)$ replaced by

$$E(g(W_k + X_{k+1})|W_k) + g(0)I(W_k + X_{k+1} \leq 0) - g(W_k + X_{k+1})I(W_k + X_{k+1} \leq 0).$$

A second approach would be through numerical integration. For ease of exposition, suppose that U and V have densities f_U and f_V, so that X has a density f_X say. Since $g(0) = 0$, we may write

$$\begin{aligned} \varphi(x) &= \int_0^\infty g(t) f_X(t-x)\, dt \\ &= \int_0^\infty \int_{t-x}^\infty g(t) f_U(u) f_V(u+t-x)\, du\, dt. \end{aligned}$$

The integrand in this double integral is nonnegative, so we may use (for example) quadrature rules in two dimensions to numerically evaluate the integral in a stable fashion (see for example, Stroud (1971)).

So a possible implementation of the AMP method might consist of a lookup table of values of $\varphi(x)$, constructed dynamically as follows. To compute $\varphi(x)$, first determine two values of x already in the table, such that $x_1 \leq x < x_2$. If no such values exist, then compute $\varphi(x)$ and store this value in the table. Since φ is increasing, a reasonable approximation of $\varphi(x)$ will be $(\varphi(x_1) + \varphi(x_2))/2$, with an error bound of $(\varphi(x_2) - \varphi(x_1))/2$. If this error bound is acceptable, then we are done, otherwise compute $\varphi(x)$ and store the result in the table.

The question of what is an acceptable error is obviously related to the precision one requires of the tail probability estimate, and how accurate we must be to ensure that variance reduction is obtained. Any implementation would have to consider this question very carefully.

Of course, the approximation to $\varphi(x)$ and the corresponding error bound discussed above do not take advantage of the convexity of φ. By using this additional information one would hope to improve the approximations and error bounds returned by the (very) simple algorithm discussed above.

We should mention at this point, that such an implementation of the AMP method would be quite expensive relative to the standard estimator, so that the computational costs might outweigh the statistical efficiencies gained. The methodology discussed in Glynn and Whitt (1992), and Glynn (1994a) might then be useful in determining which estimator to use.

2.4.2 Performance Analysis

We first provide conditions under which the AMP estimator is consistent.

Theorem 2.5 *Suppose that* $EV_0^2 < \infty$, *and* $EU_1 < \infty$. *Then the AMP estimator of the tail probability* $P(W > w)$ *is consistent.*

For the proof, see Section 2.7.1.

To compare the efficiencies of the AMP and standard estimators, we present the following example, which shows that in heavy traffic, the standard estimator has a TAVC that is of the order $O((1-\rho)^{-2})$, while the AMP estimator has a TAVC that is $O(1)$.

Example 2.5 Consider the M/M/1 queue with arrival rate λ and service rate μ, where $\rho \stackrel{\triangle}{=} \lambda/\mu < 1$. Of course, for this model, a closed form expression for the tail probability is known, but it serves as an example where the TAVC's for the standard and AMP estimators can be explicitly computed. It is well known that the exact value of the tail probability is

$$\alpha \stackrel{\triangle}{=} P(W > w) = \rho e^{-\mu(1-\rho)w}.$$

From results in Glynn (1994), we may compute the TAVC of the standard estimator by first solving Poisson's equation

$$Ag(x) = -(I(x > w) - \alpha),$$

where A is the generator of the waiting time process defined by (2.2). We find that a solution is given by

$$g(x) = \frac{-\alpha}{(1-\rho)^2} - \frac{\lambda\alpha x}{1-\rho} + \frac{\alpha e^{\mu(1-\rho)x}}{(1-\rho)^2}$$

if $x \leq w$, and

$$g(x) = 1 - \frac{\lambda\alpha w}{1-\rho} + \frac{\rho - \alpha}{(1-\rho)^2} + \frac{\lambda(1-\alpha)}{1-\rho}(x-w)$$

if $x > w$. The TAVC σ_1^2 may be computed via $\sigma_1^2 = \pi(2f_c g - f_c^2)$ (Glynn (1994)), where π is the stationary distribution of the waiting time, given by

$$\pi[0,x] = 1 - \rho e^{-\mu(1-\rho)x}$$

for $x \geq 0$. Straightforward calculations then yield

$$\sigma_1^2 = \frac{4\alpha^2(\alpha-1)\lambda w}{1-\rho} + \frac{4\rho\alpha^3 - (3+2\rho+3\rho^2)\alpha^2 + (1+\rho)^2\alpha}{(1-\rho)^2}. \tag{2.23}$$

For the AMP method we proceed as follows. Straightforward (but laborious!) calculations determine that the AMP estimator of the tail probability $P(W > w)$ is given by $n^{-1}\sum_{k=0}^{n-1} f(W_k)$, where

$$\begin{aligned} f(x) &= e^{-\theta w} - c_1 e^{-\theta w - \lambda x} + c_1 e^{-\lambda(x-w)} I(x > w) \\ &\quad + (c_2 e^{-\theta(w-x)} - c_3 e^{-\mu(w-x)}) I(x \leq w), \end{aligned} \tag{2.24}$$

and the constants c_1, c_2 and c_3 are given by

$$\begin{aligned} c_1 &= \theta(\lambda\gamma(1+\rho)(\lambda+\theta))^{-1}, \\ c_2 &= (\lambda-\mu+\theta)(\gamma(\mu-\theta)(\lambda+\theta))^{-1} \text{ and} \\ c_3 &= \lambda\theta(\gamma\mu(\lambda+\mu)(\mu-\theta))^{-1}. \end{aligned}$$

Once again we may refer to Glynn (1994) to calculate the TAVC of this estimator. The calculations involved are somewhat formidable, but if we choose θ carefully, the problem becomes manageable. Recall that θ is a parameter associated with the approximating RBM, and as such, we may choose it in any way we see fit. Large deviations theory (Asmussen (1987) p. 269) determines that in great generality, the tail probability $P(W > w)$ in a GI/G/1 queue is approximately exponential, i.e., $P(W > w) \approx ae^{-\theta^* w}$. For the M/M/1 queue, the large deviations results suggest that $\theta^* = \mu - \lambda$. (This matches the known tail probability $\rho e^{-(\mu-\lambda)x}$ exactly.) Taking $\theta = \theta^*$, we see that the constants c_1, c_2 and c_3 defined above simplify to

$$\begin{aligned} c_1 &= (1+\rho)^{-1}, \\ c_2 &= 0 \text{ and} \\ c_3 &= \rho(1+\rho)^{-1}. \end{aligned} \tag{2.25}$$

The AMP estimator is now a little more manageable! With f defined by (2.24) and the constants (2.25), define $f_c(x) = f(x) - \alpha$. A solution to Poisson's equation $Ag = -f_c$, where A is the generator (2.2), is given by

$$g(x) = \lambda e^{-\theta w} x + I(x > w).$$

The TAVC may then be computed to be of the form $a_\rho + b_\rho(1-\rho)w$, for constants a_ρ and b_ρ which remain bounded as $\rho \uparrow 1$ (and the bounds do not depend on w). We do not provide expressions for these constants because they are rather complicated, and their exact form is not important.

Let W_ρ denote the steady-state waiting time in the ρth system. We would now like to compare the behaviour of the TAVC's of the standard and AMP estimators as $\rho \uparrow 1$.

To do so for a fixed value of w would be meaningless, since $P(W_\rho > w) \to 1$ as $\rho \uparrow 1$. We therefore examine the asymptotic value of the TAVC's in estimating $P(W_\rho > w_\rho)$, where $w_\rho = w/(\mu(1-\rho))$. (We then see that $P(W_\rho > w_\rho) = \rho e^{-w} \to e^{-w}$ as $\rho \uparrow 1$.)

From (2.23), it is easy to see that the TAVC for the standard estimator σ_1^2 is asymptotically

$$\frac{4(e^{-w} - 3e^{-2w} + e^{-3w})}{(1-\rho)^2},$$

while the TAVC for the AMP estimator is $O(1)$ as $\rho \uparrow 1$. Hence, for the M/M/1 queue, the AMP estimator is more effective than the standard estimator in heavy traffic.

2.5 Queue Size Moments

In this section, we will again use a heavy traffic approximation to develop AMP estimators for the first two moments of the steady state queue size. In principle, our analysis could be extended to higher moments, but the calculations involved quickly become rather forbidding. We should emphasize that by the term *queue size*, we mean the number of customers in the entire system, and not just the number waiting for service in the queue. This is in contrast to the previous sections where we analyzed waiting times in the queue itself.

Of course, we could indirectly estimate the first moment of the queue size via Little's law and an estimator of the mean waiting time in the system (Glynn and Whitt (1989)). If $\hat{W}$ is an estimator of EW, where W is the steady-state waiting time in the system (excluding service), then an indirect estimator $\hat{Q}$ of the steady-state queue size is given by $\hat{Q} = \lambda(\hat{W} + \mu^{-1})$. Glynn and Whitt (1989) show that in great generality, the indirect estimator has a smaller TAVC than a direct estimator (i.e., an estimator based on observing the queue size process alone).

It is clear that the TAVC of $\hat{Q}$ is a scalar multiple (λ^2) of the TAVC of $\hat{W}$. Thus, if $\hat{W}$ is the Minh-Sorli estimate, the TAVC of $\hat{Q}$ will typically be of the order $O((1-\rho)^{-1})$ as $\rho \uparrow 1$, whereas if we use the AMP estimator, the TAVC will be $O((1-\rho)^{-2})$ (see Theorems 2.3 and 2.4). We will see (Theorem 2.7) that the AMP estimator of the mean queue size is of the order $O((1-\rho)^{-2})$ as $\rho \uparrow 1$. Thus it seems that in operational terms, to estimate the mean queue size for a system in moderate to heavy traffic, Little's law coupled with the Minh-Sorli estimate of the mean waiting time should be used.

In contrast to the waiting time sequence, the queue size process is *not* Markov unless the interarrival and service time distributions are exponential. We shall see that this has a marked effect on the ease with which we may obtain efficient estimators.

Recall that if $\rho < 1$ and U_1 has a nonlattice distribution, then the steady-state queue size rv Q exists. Miyazawa (1979) showed that under the additional hypothesis that $EV_0^{k+1} < \infty$, $EQ^k < \infty$.

We define the *standard estimator* of EQ^k by

$$\frac{1}{t}\int_0^t Q(s)^k\,ds. \tag{2.26}$$

As in the waiting time case, we appeal to the regenerative structure of the GI/G/1 queue to show that the standard estimator is consistent and satisfies a central limit theorem. Let $T_0 = 0$, and for $k \geq 1$, let $T_k = \inf\{t > T_{k-1} : Q(t-) = 0, Q(t) = 1\}$. The sequence $(T_j : j \geq 0)$ denotes the times when a customer arrives to an empty queue and defines regeneration times for the process $X(t) = (Q(t), A(t), S(t))$. If we define $Y_n(f) = \int_{T_{n-1}}^{T_n} f(X(s))\,ds$ for real valued functions f, the same results we quoted for the waiting time sequence also hold for $X(t)$. Namely, if $ET_1 < \infty$ then the estimator of EQ^k given in (2.26) is consistent, and in addition, if $0 < E[(Y_1(f_k) - T_1 EQ^k)^2] < \infty$, (where $f_k(q, t_1, t_2) = q^k$), then there exists a finite constant η_k^2 such that

$$\sqrt{t}\left(\frac{1}{t}\int_0^t Q(s)^k\,ds - EQ^k\right) \Rightarrow \eta_k N(0,1).$$

To the best of our knowledge, there is no general result to show that η_k^2 is of the order $(1-\rho)^{-2k-2}$, although we have observed this rate empirically. In the case $k = 1$, the result follows from that for the first moment of the waiting time (as proved by Asmussen (1992)), because of the direct relationship between sample path integrals of the form $\int_0^t Q(s)\,ds$ and $\sum_{j=0}^n W_j$.

Proposition 2.1 *If the interarrival and service times have exponential moments, then the TAVC* η_1^2 *is* $O((1-\rho)^{-4})$ *as* $\rho \uparrow 1$.

The proof of this result is given in Section 2.7. Furthermore, as might be expected, the result for general k holds for the M/M/1 queue.

Proposition 2.2 *For the M/M/1 queue, the TAVC* η_k^2 *is of the order* $O((1-\rho)^{-2k-2})$.

The proof of this result is given in Section 2.7.

Thus, there is evidence to indicate that η_k^2 is typically of the order $(1-\rho)^{-2k-2}$, so that in heavy traffic the standard estimators of the moments of the queue size will require large run lengths to return reliable estimates.

2.5.1 Heavy Traffic Theory

As in the waiting time case, we will be using a heavy traffic approximation to the queue size process to obtain an AMP estimator, and so we will need to embed our particular GI/G/1 queue in a parameterized family of such queues.

Define $\bar{U}_n$, $\bar{V}_n$, $U_n(\rho)$ and $V_n(\rho)$ as in Section 2.3.5, and let $(Q_\rho(t) : t \geq 0)$ be the queue size process in the ρth system, assuming that a customer arrives at time 0 to an empty system.

Let $\sigma^2 = \mu^3 \text{var}\, \bar{U}_1 + \mu^3 \text{var}\, \bar{V}_1$. Note that σ^2 does not depend on ρ. Define the process $R^{(\rho)} = (R^{(\rho)}(t) : t \geq 0)$ by

$$R^{(\rho)}(t) = (1-\rho) Q_\rho \left(\frac{t}{(1-\rho)^2} \right),$$

and let $R = (R(t) : t \geq 0)$ be a reflected (or regulated) Brownian motion (RBM) with drift $-\mu$, variance σ^2 and starting state $R(0) = 0$. The following result is adapted from Lemoine (1978).

Theorem 2.6 *Suppose that* $E\bar{U}_1^3 + E\bar{V}_1^3 < \infty$. *Then* $R^{(\rho)} \Rightarrow R$ *in* $D[0,1]$ *as* $\rho \to 1$, *where* $D[0,1]$ *is the space of real-valued processes on* $[0,1]$ *with sample paths that are right-continuous and have left limits.*

Theorem 2.6 suggests the following approximation for the stochastic process $Q_\rho(\cdot)$.

$$Q_\rho(\cdot) \overset{\mathcal{D}}{\approx} \frac{R((1-\rho)^2 \cdot)}{1-\rho} \overset{\mathcal{D}}{=} \tilde{Q}(\cdot) \tag{2.27}$$

where $\tilde{Q}(\cdot)$ is a reflected Brownian motion with drift $-\mu(1-\rho)$ and variance σ^2.

2.5.2 Solving Poisson's Equation for the RBM

Equation (2.27) furnishes the required approximation for $Q(\cdot)$. The first step in applying the AMP method is to solve Poisson's equation for the approximating system $\tilde{Q}(\cdot)$ to obtain the surrogate function. The analysis required is identical to Section 2.3.2, and so we merely state the results.

The surrogate function g for cost functions of the form $f(x) = x^k$ is a polynomial of order $k+1$, with no linear or constant terms. Specifically, for $k = 1$, $g(x) = x^2/(2\mu(1-\rho))$, and for $k = 2$,

$$\frac{x^3}{3\mu(1-\rho)} + \frac{\sigma^2 x^2}{2\mu^2(1-\rho)^2}.$$

The next step is to calculate the shadow function Ag, where A is the generator of the process $X(\cdot)$ described in Section 2.2.2.

Since g is a polynomial, the first step in calculating Ag is to find an expression for $Af_k(q,t_a,t_s)$, where $f_k(q,t_a,t_s) = q^k$. From (2.4), we find that for $q > 0$

$$Af_k(q,t_a,t_s) = r_a(t_a)\sum_{j=0}^{k-1}\binom{k}{j}q^j + r_s(t_s)\sum_{j=0}^{k-1}\binom{k}{j}(-1)^{k-j}q^j, \tag{2.28}$$

and for $q = 0$, $Af_k(0,t_a,0) = r_a(t_a)$. Dependence on ρ is exhibited through the functions r_a and r_s.

2.5.3 The First Moment

For the first moment, the cost function is $f(x) = x$, and the surrogate function (for the approximating RBM $\tilde{Q}$) is $g_0(q,t_a,t_s) = q^2$ (modulo a multiplicative constant). Then, from (2.28), we compute (for $q > 0$)

$$Ag_0(q,t_a,t_s) = 2q(r_a(t_a) - r_s(t_s)) + r_a(t_a) + r_s(t_s), \tag{2.29}$$

so that the AMP estimator (based on $f(q,t_a,t_s) + Ag_0(q,t_a,t_s)/(2\mu(1-\rho)))$ is given by

$$\begin{aligned}\alpha_1(t) &= t^{-1}\int_0^t Q(u)\left(1 + \frac{r_a(A(u)) - r_s(S(u))}{\mu(1-\rho)}\right)\\ &\quad + \frac{r_a(A(u)) + I(Q(u) > 0)r_s(S(u))}{2\mu(1-\rho)}\,du,\end{aligned} \tag{2.30}$$

For the M/M/1 queue, (2.30) simplifies to

$$\alpha_1(t) = \frac{\rho}{2(1-\rho)} + \frac{1}{2(1-\rho)}t^{-1}\int_0^t I(Q(s) > 0)\,ds.$$

From (2.39), we see that $\alpha_1(t)$ has a TAVC given by $\rho/(2\mu(1-\rho)^2)$. This compares favourably with the standard estimator (see Proposition 2.2). However, the M/M/1

ρ	Standard			AMP		
	Pt Est	TAVC	CI	Pt Est	TAVC	CI
0.1	0.10	0.085	$\pm$ 0.003	0.10	0.023	$\pm$ 0.002
0.3	0.34	0.46	$\pm$ 0.02	0.34	0.22	$\pm$ 0.02
0.5	0.65	1.8	$\pm$ 0.2	0.65	1.6	$\pm$ 0.2
0.7	1.2	16	$\pm$ 4	1.2	19	$\pm$ 2
0.9	3.4	900	$\pm$ 400	3.5	1500	$\pm$ 500
0.95	6.9	2E+4	$\pm$ 1E+4	6.8	2.6E+4	$\pm$ 8E+3

Table 2.2: Simulation results for estimating the mean queue size in the U/U/1 queue.

queue size process does not require the age processes $A(s)$ and $S(s)$ to make it Markov. We therefore simulated an instance of the U/U/1 queue to get some idea of the estimator's performance.

Example 2.6 Let $Q(t)$ be the queue size process in a U/U/1 queue. The service time distribution is uniform on $(0, 2)$, and the interarrival time distribution is uniform on $(0, 2/\rho)$, so that the traffic intensity is ρ. It is straight-forward to calculate

$$r_a(t) = \begin{cases} \rho/(2 - \rho t) & \text{if } 0 \leq t < 2\rho^{-1} \\ 0 & \text{otherwise} \end{cases}$$

with a similar expression for $r_s(t)$.

We simulated 20 repetitions of 2000 busy cycles for various values of ρ, and computed estimates of the TAVC's of the standard estimator $(t^{-1} \int_0^t Q(s)\, ds)$, and the AMP estimator (2.30). The results are given in Table 2.2. The first number in each column of Table 2.2 is the point estimate, the second is an estimate of the TAVC, and the third provides a 95% confidence interval for the TAVC.

It is clear from these results that something is wrong. Because the RBM becomes a good approximation of the queue size process, we would expect that as $\rho \uparrow 1$, the AMP estimator should become extremely effective, just as we saw for waiting time moments (Figure 2.1), but this is not the case. In fact, the AMP estimator does worse than the standard estimator for $\rho \geq 0.7$!

The problem stems from the supplementary variables we had to adjoin to the state space of the queue size process to make it Markov. To see why, we reason intuitively as follows. Recall that the AMP estimator will be effective if the surrogate function g is a good approximation to the solution of Poisson's equation, in the sense that

$$Ag(q, t_a, t_s) \approx -(q - EQ).$$

From (2.29) we see that this is far from the case when $g(q, t_a, t_s) = q^2$. The problem is that g ignores the effect of the age processes $A(s)$ and $S(s)$. In the waiting time case this problem did not arise, because the process of interest $(W_n : n \geq 0)$ was already Markov.

There are perhaps two ways to proceed at this point. One method would be to attempt to alter g in some way to take account of the effect of the age variables. We will discuss this approach in great detail shortly. A second method would be to use an estimator of the form

$$\alpha_2(t) \stackrel{\triangle}{=} \frac{1}{t}\int_0^t f(X(s))\,ds + \frac{\beta}{t}\int_0^t Ag(X(s))\,ds$$

where we choose β so as to attempt to minimize the TAVC of $\alpha_2(t)$. This second method is the classical control variates approach. We would expect that such an approach would achieve useful variance reductions, but not the order of magnitude reductions achieved previously for waiting time performance measures. This expectation is borne out in our continuation of Example 2.6.

Returning to our U/U/1 example, consider an application of the control variates approach. Once again we simulated 20 repetitions of 2000 regenerative cycles for various values of ρ. The standard estimator is computed exactly as before, but the computation of the AMP estimator is now somewhat different. The AMP estimator is now of the form

$$\frac{\sum_{i=1}^{2000} Y_i + \beta Z_i}{\sum_{i=1}^{2000} \tau_i}$$

where $Y_i = \int_{C_i} f(X(s))\,ds$, $Z_i = \int_{C_i} Ag(X(s))\,ds$, C_i is the ith regenerative cycle, and τ_i is the length of C_i. We chose β to be the sample estimate of

$$-\mathrm{cov}(Y_1 - \alpha\tau_1, Z_1)/\mathrm{var}\, Z_1,$$

ρ	AMP		
	Pt Est	TAVC	CI
0.1	0.10	0.021	$\pm$ 0.0009
0.3	0.34	0.11	$\pm$ 0.004
0.5	0.65	0.40	$\pm$ 0.03
0.7	1.2	3.7	$\pm$ 0.7
0.9	3.5	240	$\pm$ 70
0.95	6.7	3200	$\pm$ 600

Table 2.3: Simulation results for the control variate approach to estimating the mean queue size in the U/U/1 queue.

(Loh (1994) p. 29) where $\alpha = EY_1/E\tau_1$. The results for the AMP estimator are presented in Table 2.3 (results for the standard estimator are the same as before).

Comparing these results with those in Table 2.2 we observe that useful variance reductions are obtained for all values of ρ, but the order of magnitude variance reductions (in heavy traffic) that we saw for the waiting time examples do not appear. Notice that even in light traffic, the AMP estimator does very well, when we would expect that the RBM approximation is very poor.

The above example demonstrates that by using the standard control variates approach we may obtain useful variance reductions when a direct application of the AMP method (taking $\beta = 1$ instead of estimating it) does not perform well.

Notice that in heavy traffic the TAVC's of the standard estimator and the AMP estimator are of the same order, even when we estimate β as above. So let us now consider the second approach mentioned earlier, namely modifying the surrogate function to account for the extra variables. (We return now to general interarrival and service time distributions under the assumptions discussed in Section 2.2.2.)

A direct result of g being a poor approximation to the solution to Poisson's equation is the non-zero multiplier of $Q(u)$ in (2.30). We would prefer this coefficient to be zero, so that the estimator would depend mainly on the age processes $A(\cdot)$ and

$S(\cdot)$. The reason for this is that, as $\rho \uparrow 1$, the queue size $Q_\rho(t)$ begins to fluctuate wildly, while the age processes do not change their behaviour in any fundamentally significant manner. Thus, a reasonable way to proceed would be to attempt to find a function that depends principally on the age processes, and *not* on the queue size.

It seems reasonable to attempt to "cancel" the term involving $Q(u)$ in (2.30), while not introducing bias. According to the discussion in Chapter 1, this may be accomplished by modifying the estimator by adding functions of the form Ag.

Consider the function $g_1(q, t_a, t_s) = q\varphi_1(t_a)$, for some function φ_1. From (2.4), we find that for $q > 0$,

$$Ag_1(q, t_a, t_s) = q(\varphi_1'(t_a) - r_a(t_a)(\varphi_1(t_a) - \varphi_1(0))) + r_a(t_a)\varphi_1(0) - \varphi_1(t_a)r_s(t_s).$$

Suppose we solve the ordinary differential equation (ODE)

$$\varphi_1'(t_a) - r_a(t_a)(\varphi_1(t_a) - 1) = r_a(t_a) - \lambda, \qquad \varphi_1(0) = 1,$$

to obtain $\varphi_1(t) = \lambda E(U - t|U > t)$ (recall that U is a generic rv representing an interarrival time). Similarly, if we set $g_2(q, t_a, t_s) = q\varphi_2(t_s)$, and solve the ODE

$$\varphi_2'(t_s) - r_s(t_s)(\varphi_2(t_s) - 1) = r_s(t_s) - \mu, \quad \varphi_2(0) = 1,$$

to obtain $\varphi_2(t) = \mu E(V - t|V > t)$, we find that if $g_2(q, t_a, t_s) = q\varphi_2(t_s)$,

$$\begin{aligned}(A(g_0/2 - g_1 + g_2))(q, t_a, t_s) &= q(\lambda - \mu) - r_a(t_a)/2 - r_s(t_s)/2 \\ &+ \varphi_1(t_a)r_s(t_s) + r_a(t_a)\varphi_2(t_s).\end{aligned}$$

So let

$$\begin{aligned}\psi_1(q, t_a, t_s) &= q + (\mu - \lambda)^{-1}A(g_0/2 - g_1 + g_2)(q, t_a, t_s) \\ &= (\mu - \lambda)^{-1}(\varphi_1(t_a)r_s(t_s) + r_a(t_a)\varphi_2(t_s) - (r_a(t_a) + r_s(t_s))/2)\end{aligned}$$

for $q > 0$, and for $q = 0$, $\psi_1(0, t_a, 0) = (\mu - \lambda)^{-1}r_a(t_a)/2$. An estimator of the steady-state mean of $Q(\cdot)$, is given by $t^{-1}\int_0^t \psi_1(X(t))$. We will see shortly that it will prove

useful to remove the "interaction terms" $\varphi_1(t_a)r_s(t_s)$ and $r_a(t_a)\varphi_2(t_s)$. So we alter the estimator by adding $(\mu-\lambda)^{-1}Ag_3$, where $g_3(q,t_a,t_s)=\varphi_1(t_a)\varphi_2(t_s)$. This yields the function

$$\begin{aligned}\psi_2(q,t_a,t_s) &= q+(\mu-\lambda)^{-1}A(g_0/2-g_1+g_2-g_3)(q,t_a,t_s)\\ &= \frac{\mu\varphi_1(t_a)+\lambda\varphi_2(t_s)-(r_a(t_a)+r_s(t_s))/2}{\mu-\lambda}\end{aligned}$$

for $q>0$, and for $q=0$, $\psi_2(0,t_a,0)=(\mu-\lambda)^{-1}(\lambda-r_a(t_a)/2)$.

The only dependence on the queue size exhibited by ψ_2 is whether q is positive or zero. We therefore expect that an estimator based on this function will be very effective compared with the standard estimator in heavy traffic.

But we can do more. The marginal stationary distribution of the age process $A(t)$ has a density given by $\pi_a(dt)=\lambda\bar{F}_a(t)\,dt$ for $t\geq 0$ (Asmussen (1987) p. 116). Therefore, terms which involve the age process $A(t)$ alone may be replaced by their steady-state mean. Let us rewrite $\psi_2(q,t_a,t_s)$ as

$$\begin{aligned}\psi_2(q,t_a,t_s) &= \frac{\varphi_1(t_a)-(2\mu)^{-1}r_a(t_a)}{1-\rho}+\frac{I(q>0)}{1-\rho}(\rho\varphi_2(t_s)-\frac{r_s(t_s)}{2\mu})\\ &\quad+\frac{I(q=0)}{1-\rho}(\rho-\varphi_1(t_a))\end{aligned}\tag{2.31}$$

We may replace the first term in (2.31) by $(1-\rho)^{-1}\pi_a(\varphi_1-(2\mu)^{-1}r_a)$ which is easily computed to be $(2(1-\rho))^{-1}(\lambda^2EU^2-\rho)$. A similar approach may be used to deal with the second term in (2.31). In this case, the age process $S(\cdot)$ is not a renewal process, because of the "boundary effects" when the queue size is 0. Let $\tilde{S}=(\tilde{S}(t):t\geq 0)$ be the renewal process obtained from the sequence of service times. Then $\tilde{S}$ may be thought of as the renewal process which results from $S(\cdot)$ if we "cut out" the periods when $Q(s)=0$. Let $B(t)=\int_0^t I(Q(s)>0)\,ds$ be the amount of time the server is busy in $(0,t)$. Then, if φ is a given non-negative function, we find that

$$\begin{aligned}t^{-1}\int_0^t\varphi(S(u))I(Q(u)>0)\,du &= \frac{B(t)}{t}B(t)^{-1}\int_0^{B(t)}\varphi(\tilde{S}(u))\,du\\ &\to P(Q>0)\bar{\pi}_s\varphi\end{aligned}$$

as $t \to \infty$, where $\bar{\pi}_s(du) = \mu \bar{F}_s(u)\, du$ for $u \geq 0$. Since $P(Q > 0) = \rho$, we obtain (after simplification) our final function

$$\phi_1(q, t_a, t_s) = \rho + \frac{\lambda E X^2}{-2EX} - \frac{\varphi_1(t_a) I(q = 0)}{1 - \rho} \tag{2.32}$$

and an estimator of the mean steady-state queue size is given by

$$\frac{1}{t} \int_0^t \phi_1(Q(u), A(u), S(u))\, du. \tag{2.33}$$

Note the similarity of ϕ_1 to (2.11). This estimator depends on the queue size process only on the boundary, and so it seems reasonable that such an estimator will be efficient in heavy traffic. This expectation is borne out by our analysis and examples in Section 2.5.5.

2.5.4 The Second Moment

We now turn to developing an AMP estimator for EQ^2, where Q is the steady-state queue size. The procedure is very similar to the analysis for the mean, so we will skip several of the details.

Let

$$\begin{aligned}
g_0(q, t_a, t_s) &= q^3, \\
g_1(q, t_a, t_s) &= q^2 \varphi_1(t_a), \\
g_2(q, t_a, t_s) &= q^2 \varphi_2(t_s), \\
g_3(q, t_a, t_s) &= q \varphi_1(t_a) \varphi_2(t_s), \\
g_4(q, t_a, t_s) &= q \varphi_1(t_a) \text{ and} \\
g_5(q, t_a, t_s) &= q \varphi_2(t_s)
\end{aligned}$$

Then, setting $g_6 = g_0/3 - g_1 + g_2 - 2g_3 + g_4 + g_5$, we find that for $q > 0$,

$$\begin{aligned}
q^2 + (\mu - \lambda)^{-1} A g_6(q, t_a, t_s) &= \frac{q}{\mu - \lambda}(2\lambda \varphi_2(t_s) + 2\mu \varphi_1(t_a) - \lambda - \mu) \\
&+ \frac{r_a(t_a) - r_s(t_s)}{3(\mu - \lambda)}.
\end{aligned}$$

The next step is to "remove" the $q\varphi_i$ terms. Let $\alpha \stackrel{\triangle}{=} \lambda^2 EU^2/2$, and define the function $g_7(q, t_a, t_s) = q\varphi_3(t_a)$, where

$$\begin{aligned}\varphi_3(t) &= \bar{F}_a(t)^{-1} \int_0^t \bar{F}_a(u)(\varphi_1(u) - \alpha)\, du \\ &= \alpha E(U - t|U > t) - \frac{\lambda}{2} E((U-t)^2|U > t).\end{aligned}$$

Then for $q > 0$,

$$Ag_7(q, t_a, t_s) = q(\varphi_1(t_a) - \alpha) - \varphi_3(t_a) r_s(t_s).$$

It is easy to calculate $\pi\varphi_3 = \lambda^3 u_2^2/4 - \lambda^2 u_3/6$, where $u_k = EU^k$ for $k = 2, 3$. Similarly, define $g_8(q, t_a, t_s) = q\varphi_4(t_s)$, where

$$\varphi_4(t) = \beta E(V - t|V > t) - \frac{\mu}{2} E((V-t)^2|V > t),$$

and choose $\beta = \mu^2 EV^2/2$. For future reference, let $v_k \stackrel{\triangle}{=} EV^k$, for $k = 2, 3$. Setting $g_9 = g_6 - 2\mu g_7 - 2\lambda g_8$, we find that for $q > 0$,

$$\begin{aligned}q^2 + \frac{Ag_9(q, t_a, t_s)}{\mu - \lambda} &= \frac{(2\mu\alpha + 2\lambda\beta - \mu - \lambda)q}{\mu - \lambda} + \frac{1}{\mu - \lambda}\{(r_a(t_a) - r_s(t_s))/3 \\ &+ 2\mu\varphi_3(t_a) r_s(t_s) - 2\lambda r_a(t_a)\varphi_4(t_s)\} \qquad (2.34)\end{aligned}$$

At this point we may substitute our estimator (2.33) in for q, and therefore have an estimator which relies on the queue size q only according to whether q is positive or zero. However, as in the previous section, it will be useful to remove the interaction terms present in (2.34), so that we may replace terms involving only one of the age processes ($A(s)$ or $S(s)$) with their known stationary expected values. Define $g_{10} = \varphi_3(t_a)\varphi_2(t_s)$ and $g_{11} = \varphi_1(t_a)\varphi_4(t_s)$, and let $g_{12} = g_9 - 2\mu g_{10} + 2\lambda g_{11}$. The function g_{12} is our final surrogate function. Next, calculate $q^2 + Ag_{12}/(\mu - \lambda)$, and substitute in the estimator (2.33) for q. After substituting in the known stationary expected values we obtain

$$\begin{aligned}\phi_2(q, t_a, t_s) &= \frac{I(q=0)}{(1-\rho)^2}(\varphi_1(t_a)(3\rho - 1 - \lambda^2(u_2 + v_2)) - 2\mu(1-\rho)\varphi_3(t_a)) \\ &\quad -2\varphi_1(t_a)\varphi_2(t_s) I(q > 0) + k\end{aligned}$$

where

$$k = \frac{\lambda^2\mu(\rho v_3 - u_3)}{3(1-\rho)} + \frac{\lambda^3(\mu u_2 + \lambda v_2)(u_2 + v_2)}{2(1-\rho)^2} + \frac{\lambda^2 u_2(1-5\rho) - \lambda^2 v_2(1+3\rho)}{2(1-\rho)^2} + \frac{\rho^2(1+\rho)}{(1-\rho)^2}.$$

The AMP estimator for EQ^2 is then

$$\frac{1}{t}\int_0^t \phi_2(X(s))\,ds. \tag{2.35}$$

2.5.5 Performance Analysis

The goal of this section is to demonstrate that the estimators we derived for the first two moments of the steady-state queue size Q are consistent, and have lower TAVC's than the standard estimators in heavy traffic.

The following theorem establishes that the TAVC for the AMP estimator of EQ^k is $O((1-\rho)^{-2k})$ as $\rho \to 1$, for $k = 1, 2$. This compares favourably with $O((1-\rho)^{-2k-2})$, which Proposition 2.2 suggests is the order of the TAVC of the standard estimator. For a proof of this result, see Section 2.7.2.

Theorem 2.7 *Suppose that $E\bar{U}_1^5 + E\bar{V}_0^5 < \infty$, the functions φ_1 and φ_2 are bounded, and $\bar{U}_1$ and $\bar{V}_0$ both possess bounded and continuous densities (with respect to Lebesgue measure) and bounded hazard rate functions. Then the estimators (2.33) and (2.35) satisfy CLT's of the form*

$$t^{1/2}\left(\frac{1}{t}\int_0^t \phi_k(X(u))\,du - EQ^k\right) \Rightarrow \sigma_k N(0,1)$$

as $t \to \infty$, for $k = 1, 2$. Furthermore, the TAVC $\sigma_k^2 = \sigma_k^2(\rho) = O((1-\rho)^{-2k})$ as $\rho \uparrow 1$.

We remark that the conditions of this theorem are *sufficient* conditions, and should not be considered necessary. In particular, the assumption that the hazard rate

functions be bounded is somewhat restrictive, since it precludes the use of bounded interarrival and service time random variables! However, it seems reasonable that this assumption could be removed in tighter versions of these theorems.

The following examples provide concrete demonstrations of the above results. The second example reinforces our view that the sufficient conditions stated in the above theorems are stronger than strictly required. In fact, for that example, the densities are not continuous, and the hazard rate functions are unbounded!

Example 2.7 Consider the M/M/1 queue with arrival rate λ, service rate μ and traffic intensity ρ. Let $Q(s)$ be the number of customers in the system at time s, and suppose that at time 0 a customer arrives to an empty system. As before, the M/M/1 queue is a system for which we may explicitly compute the variance constants for the standard and AMP estimators of the moments of the steady-state system size Q. We shall do so for the first moment.

Let π be the distribution of Q. Both the standard estimator, and the AMP estimator are of the form $t^{-1}\int_0^t f(Q(s))\,ds$, for some suitably defined f. We may compute the TAVC of such an estimator using Theorem 4.3 of Glynn and Meyn (1993). The procedure is to first solve Poisson's equation $Ag = -f_c$, and then compute the TAVC $\sigma_f^2 = \pi(2f_c g)$. The function $f_c(x) \stackrel{\triangle}{=} f(x) - \pi f$, and the generator A is, of course, the rate matrix

$$A = \begin{bmatrix} -\lambda & \lambda & & & \\ \mu & -(\lambda+\mu) & \lambda & & \\ & \mu & -(\lambda+\mu) & \lambda & \\ & & & \ddots & \end{bmatrix}.$$

For the standard estimator, $f(x) = x$ and $f_c(x) = x - \rho(1-\rho)^{-1}$. It is easy to check that a solution to Poisson's equation is given by

$$g(n) = \frac{n(n+1)}{2\mu(1-\rho)}$$

and that the resulting TAVC is

$$\sigma_1^2 = \frac{2\rho(1+\rho)}{\mu(1-\rho)^4}.$$

As for the AMP estimator, the estimating function ϕ_1 simplifies to

$$f(q) = \frac{1 - I(q=0)}{1-\rho}.$$

For this f, a solution to Poisson's equation is given by $g(n) = n/(\mu(1-\rho))$, and the resulting TAVC σ_2^2 is given by

$$\sigma_2^2 = \frac{2\rho}{\mu(1-\rho)^2}.$$

Comparing the two variance constants σ_1^2 and σ_2^2 we see that the AMP estimator is statistically more efficient for all $\rho \in (0,1)$.

Example 2.8 Our second example is the U/U/1 queue, where we once again estimated the mean queue size. We chose the interarrival and service time distributions to be uniform on $[0, 2/\rho]$ and $[0,2]$ respectively, so that the traffic intensity was ρ. The function $\varphi_1(t) = 1 - \rho t/2$ for $0 \le t \le 2/\rho$. To compare the standard and AMP estimators, we simulated 20 repetitions of 2000 busy cycles for various values of ρ, and computed the point estimates, estimates of their TAVC's, and 95% confidence intervals for the TAVC's. The results are given in Table 2.4, and the TAVC estimates are plotted against $1-\rho$ in Figure 2.2. The first number in each column of Table 2.4 is the point estimate, the second is an estimate of the TAVC, and the third provides a 95% confidence interval for the TAVC.

As discussed in the introduction to this section, an indirect estimator $\hat{Q}$ of the mean queue size can be obtained by setting $\hat{Q} = \lambda(\hat{W} + \mu^{-1})$, where $\hat{W}$ is an estimator of the mean steady-state waiting time EW (excluding service). Table 2.5 provides data for such an estimator when the standard, AMP and Minh-Sorli estimators of EW are employed. These values are derived from the results given in Table 2.1.

ρ	Standard			AMP		
	Pt Est	TAVC	CI	Pt Est	TAVC	CI
0.1	0.10	0.085	± 0.003	0.10	0.12	± 0.002
0.3	0.34	0.46	± 0.02	0.34	0.19	± 0.003
0.5	0.65	1.8	± 0.2	0.65	0.48	± 0.01
0.7	1.2	16	± 4	1.2	1.4	± 0.04
0.9	3.4	1000	± 400	3.5	13	± 1
0.95	6.9	2E+4	± 1E+4	6.9	53	± 5
0.99	35	9E+6	± 4E+6	34	1200	± 200

Table 2.4: Simulation results for estimating the mean queue size in the U/U/1 queue: the improved AMP estimator.

ρ	Standard			AMP			Minh-Sorli		
	Est	TAVC	CI	Est	TAVC	CI	Est	TAVC	CI
0.1	0.10	1.4E-5	2E-6	0.10	3.7E-4	5E-5	0.11	0.057	6E-4
0.3	0.34	2.3E-3	3E-4	0.34	0.014	2E-3	0.34	0.064	9E-4
0.5	0.65	0.5	0.1	0.65	0.09	0.005	0.65	0.093	0.002
0.7	1.2	3.4	0.5	1.2	0.58	0.04	1.2	0.16	0.003
0.9	3.5	320	80	3.5	11	0.8	3.5	0.48	0.008
0.95	6.7	11000	8000	6.8	46	4	6.8	0.91	0.02
0.99	34	6E+6	4E+6	34	1100	100	34	4.7	0.1

Table 2.5: Simulation results for indirect estimators of the mean queue size based on the standard, AMP, and Minh-Sorli estimators of the mean waiting time.

We see that using the Minh-Sorli estimator is by far the most efficient in heavy traffic (just as we anticipated in the introduction to this section). Comparing the results for the standard and AMP estimators in the direct (above) and indirect case we observe that the indirect estimators are far more effective than the direct estimators, but that the difference becomes less pronounced as we move in to heavy traffic.

2.6 Queue Size Tails

In this section we consider the problem of estimating the tail probability $P(Q \geq q)$, where Q is the steady-state queue size rv.

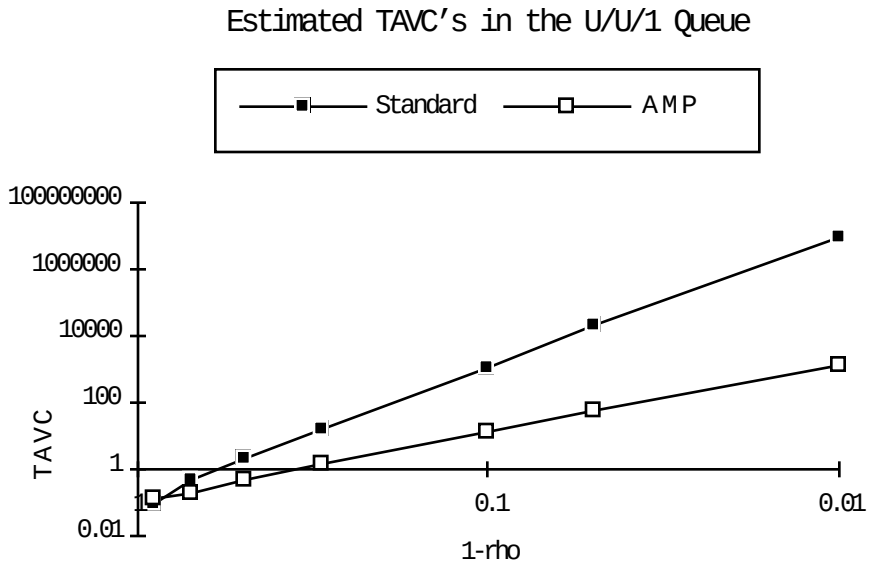

Figure 2.2: Log/log plot ofTAVC estimates for the mean queue size in the U/U/1 queue.

By the *standard* estimator, we mean

$$\frac{1}{t}\int_0^t I(Q(s) \geq q)\, ds, \tag{2.36}$$

which is based on the *estimating function* $f(x) \stackrel{\triangle}{=} I(x \geq q)$. We shall derive an AMP estimator, again using a heavy traffic approximation to the queue size process.

Recall from Theorem 2.6 that the queue size process $(Q(t) : t \geq 0)$ may be approximated by an RBM $(\tilde{Q}(t) : t \geq 0)$ with drift $-\gamma < 0$, infinitesimal variance $\sigma^2 > 0$, and initial state 0. Define $\theta = 2\gamma/\sigma^2$. The first step is to solve Poisson's equation for the RBM, given by

$$\frac{\sigma^2}{2} g_1''(x) + \gamma g_1'(x) = -(I(x \geq q) - e^{-\theta q}) \qquad g_1(0) = g_1'(0) = 0.$$

The solution to this equation is

$$g_1(x) = \begin{cases} \frac{e^{-\theta q}}{\gamma\theta}(e^{\theta x} - \theta x - 1) & \text{if } 0 \leq x < q, \\ c + \frac{1-e^{-\theta q}}{\gamma}(x - q) & \text{if } x \geq q, \end{cases} \tag{2.37}$$

where

$$c \stackrel{\triangle}{=} \frac{1 - e^{-\theta q} - \theta q e^{-\theta q}}{\gamma\theta}.$$

A naive application of the AMP method results in an estimator of the form

$$\frac{1}{t}\int_0^t I(Q(s) \geq q) + Ag_1(X(s))\, ds,$$

where A is the generator of the process $X(t) = (Q(t), A(t), S(t))$, defined in Section 2.2.2.

Just as in Section 2.5, a test of this estimator on the U/U/1 queue reveals that it does not dramatically out-perform the standard estimator in heavy traffic, while it does very well on the M/M/1 queue. An inspection of the estimator reveals that we have the same problem as we did for queueing moments, namely that the surrogate function g_1 ignores the effect of the age processes $A(\cdot)$ and $S(\cdot)$, and is therefore a poor approximation to the solution to Poisson's equation. But by taking account of the age processes we may improve the performance of the estimator in heavy traffic, just as we did in Section 2.5. Recall that we added extra functions of the form Ag for some g to obtain an improved AMP estimator. But how do we choose the extra functions? There seems to be some art to this, but the following rule of thumb appears to be useful. Since the estimator works very well for the M/M/1 queue, it seems reasonable to attempt to add functions which yield a shadow function that resembles the M/M/1 shadow function. This approach served us well in Section 2.5, and we shall see that it also appears to work in the current context. However, we emphasize that we do not know of any *automatic* method for choosing these functions.

Define the functions $\varphi_1(t_a)$ and $\varphi_2(t_s)$ as in Section 2.5, and set

$$\begin{aligned}
g_2(n, t_a, t_s) &= \frac{e^{-\theta q} - I(n \geq q)}{\gamma}(\varphi_1(t_a) - \varphi_2(t_s)), \\
g_3(n, t_a, t_s) &= \frac{e^{-\theta(q-n)}(1 - e^{-\theta})}{\gamma\theta}(\varphi_1(t_a) - e^{\theta}\varphi_2(t_s))I(n < q) \text{ and} \\
g_4(n, t_a, t_s) &= \frac{e^{-\theta}(e^{\theta} - 1)^2}{\gamma\theta} e^{-\theta(q-n)}\varphi_1(t_a)\varphi_2(t_s)I(n < q).
\end{aligned}$$

Now, let $g = g_1 + g_2 - g_3 - g_4$, and $\phi = f + Ag$. The function value $\phi(n, t_a, t_s)$ is given by

$$\begin{cases} \frac{\lambda e^{-\theta q}}{\gamma\theta}(e^{\theta} - \theta - 1) & \text{if } n = 0, \\ e^{-\theta q} + \frac{(\lambda e^{-\theta} - \mu)(e^{\theta} - 1)}{\gamma\theta} e^{-\theta(q-n)} & \\ + \frac{e^{-\theta}(e^{\theta} - 1)^2}{\gamma\theta} e^{-\theta(q-n)}(\lambda\varphi_2(t_s) + \mu\varphi_1(t_a)) & \text{if } 1 \le n \le q-2, \\ \frac{(\lambda e^{-\theta} - \mu)(1 - e^{-\theta})}{\gamma\theta} + \frac{r_a(t_a)(1 - \varphi_2(t_s))}{\gamma\theta}(1 - \theta - e^{-\theta}) & \\ + \frac{(1 - e^{-\theta})^2}{\gamma\theta}(\lambda\varphi_2(t_s) + \mu\varphi_1(t_a)) + e^{-\theta q} & \text{if } n = q-1, \\ e^{-\theta q} + \frac{\varphi_1(t_a) r_s(t_s)}{\gamma\theta}(e^{-\theta} + \theta - 1) & \text{if } n = q, \text{ and} \\ e^{-\theta q} & \text{if } n > q. \end{cases}$$

The AMP estimator of the tail probability $P(Q \ge q)$ is then given by

$$t^{-1} \int_0^t \phi(X(s))\, ds \tag{2.38}$$

This rather fearsome looking estimator is nevertheless quite effective in heavy traffic, as the examples in the next section demonstrate.

2.6.1 Performance Analysis

We first provide conditions under which (2.38) is a consistent estimator of $P(Q \ge q)$.

Theorem 2.8 *Suppose that $\rho < 1$, U_1 and V_0 have bounded and continuous densities, and bounded hazard rate functions. If $EV_0^2 < \infty$, and the functions φ_1 and φ_2 are bounded, then the AMP estimator of the tail probability $P(Q \ge q)$ is consistent.*

The proof of this result is given in Section 2.7.2. We remark that just as in the case for computing queue size moments, we believe that the conditions given in this theorem may be weakened. Our second example provides empirical evidence for this belief.

Example 2.9 Consider the M/M/1 queue with arrival rate λ, service rate μ, and traffic intensity $\rho = \lambda/\mu$.

Theorem 4.2 of Glynn and Torres determines the TAVC for the standard estimator (2.36) to be

$$\sigma_1^2 = \frac{2(1+\rho)(1-\rho^q)\rho^q}{\mu(1-\rho)^2} - \frac{4q\rho^{2q}}{\mu(1-\rho)}. \tag{2.39}$$

Just as for the waiting time tails, to compare the standard and AMP estimators in heavy traffic, we must allow the tail point q to change with ρ, so that we consider the problem of estimating $P(Q_\rho \geq q_\rho)$ where $q_\rho = q/(1-\rho)$. (The rv Q_ρ represents the steady-state queue size in an M/M/1 queue with service rate μ and arrival rate $\lambda_\rho \stackrel{\triangle}{=} \mu\rho$). Taking $q = q_\rho$ in (2.39) and noting that $\rho^{q_\rho} \sim e^{-q}$ as $\rho \uparrow 1$, we see that

$$\sigma_1^2 \sim \frac{4e^{-q} - 4(1+q)e^{-2q}}{\mu(1-\rho^2)}$$

as $\rho \uparrow 1$, which is the same behaviour observed for the standard estimator of the waiting time tail in the M/M/1 queue.

Even for the M/M/1 queue, computing the TAVC for an estimator based on the function ϕ is a daunting calculation, and so we instead present numerical results from a simulation. We chose $\mu = 1$, and simulated 200 replicates of 2000 busy cycles for various values of ρ, computing the AMP estimator of the tail probability $P(Q_\rho \geq q_\rho)$ where $q_\rho = \lceil 2(1-\rho)^{-1} \rceil$. The results are given in Table 2.6, where we have tabulated ρ, q_ρ, the known $P(Q_\rho \geq q_\rho)$, the known value of the TAVC of the standard estimator (calculated from (2.39)), and for the AMP estimator, the point estimate, an estimate of the TAVC, and a 95% confidence interval for the TAVC.

We see the same behaviour exhibited by these estimators as in the waiting time tail case, i.e., the standard estimator's TAVC is growing rapidly as $\rho \uparrow 1$, while the TAVC for the AMP estimator is not.

Example 2.10 Our second example is the U/U/1 queue, with service times uniform on $(0, 2)$ and interarrival times uniform on $(0, 2/\rho)$. We simulated 200 repetitions of

ρ	q_ρ	$P(Q_\rho \geq q_\rho)$	Standard TAVC	AMP Pt Est	AMP TAVC	AMP CI
0.1	3	0.001	0.0027	9.9E-4	6E-4	± 1E-4
0.3	3	0.027	0.13	0.027	0.015	± 0.001
0.5	4	0.063	0.58	0.062	0.035	± 0.003
0.7	7	0.082	2.2	0.083	0.044	± 0.003
0.9	20	0.12	29	0.12	0.054	± 0.004
0.95	40	0.13	120	0.13	0.047	± 0.006
0.99	200	0.13	3200	0.13	0.03	± 0.01

Table 2.6: Simulation results for estimating the tail probabilities of the queue size in the M/M/1 queue.

ρ	q_ρ	Standard Pt Est	Standard TAVC	Standard CI	AMP Pt Est	AMP TAVC * CI	
0.1	2	0.0035	0.0042	± 0.0004	0.0036	0.0076	± 0.0008
0.3	2	0.036	0.060	± 0.003	0.036	0.069	± 0.005
0.5	2	0.12	0.30	± 0.02	0.12	0.18	± 0.008
0.7	4	0.044	0.46	± 0.09	0.043	0.10	± 0.01
0.9	10	0.049	4.1	± 2	0.057	0.080	± 0.02
0.95	20	0.053	20	± 10	0.053	0.061	± 0.01
0.99	100	0.053	300	± 100	0.050	0.041	± 0.02

Table 2.7: Simulation results for estimating tail probabilities of the queue size in the U/U/1 queue.

2000 busy cycles, and estimated $P(Q_\rho \geq q_\rho)$ where $q_\rho = \lceil (1-\rho)^{-1} \rceil$. The results are given in Table 2.7. The first number in each column of Table 2.7 is the point estimate, the second is an estimate of the TAVC, and the third provides a 95% confidence interval for the TAVC.

2.7 Proofs

2.7.1 The Waiting Time Process

In proving our results on the waiting time process, we will make frequent use of the results in the following proposition. Recall that we defined a sequence of systems parameterized by ρ, where the ρth system has interarrival times $\bar{U}_n/\rho$ and service times $\bar{V}_n$.

Proposition 2.3 *Let $C_\rho = \inf\{n \geq 1 : W_n(\rho) = 0\}$, and let $I_n(\rho)$ be the idle period (as defined in Section 2.3.5. Suppose that $W_0(\rho) = 0$ for all ρ.*

1. *Suppose that $E\bar{U}_1^3 < \infty$ and $E\bar{V}_0^3 < \infty$, and that either $\bar{U}_1$ or $\bar{V}_0$ has a continuous distribution. Then $I_1(\rho)$ converges in distribution and $EI_1(\rho)$ converges to a finite constant as $\rho \uparrow 1$. Furthermore, $(1-\rho)EC_\rho$ is bounded away from 0 and ∞.*
2. *Let $k \geq 1$. If $E\bar{U}_1^{2k+1} < \infty$ and $E\bar{V}_0^{2k+1} < \infty$ then $EC_\rho^k = O((1-\rho)^{1-2k})$ as $\rho \uparrow 1$.*

For proofs of these results, see Asmussen (1992), Lemma 4.1 and Corollary 5.1.

Lemma 2.1 *Suppose that $E(\bar{V}_0^{k+2} + \bar{U}_1^{k+2}) < \infty$. Then, for $2 \leq j \leq k+1$, the coefficient $b(k,j) = O((1-\rho)^{-(k+2-j)})$.*

Proof. The proof is by induction on k. From (2.11),

$$\phi_1(x) = \frac{EX^2}{-2EX} + \frac{1}{2EX} h_2(x).$$

Since $EX = -(1-\rho)/\lambda$, the result is true for $k = 1$. So suppose the result is true for $1 \leq j \leq k$. From our expression for $\psi_k(x)$, we see that

$$\phi_{k+1}(x) = -\sum_{j=2}^{k+2} c_{k+1}(j) h_j(x) + \sum_{m=0}^{k} \sum_{j=1}^{k+2-m} c_{k+1}(m+j) \binom{m+j}{j} EX^j \phi_m(x).$$

By the inductive hypothesis, for $m \leq k$, the coefficients $b(m,\ell)$ of $h_\ell(x)$ appearing in the expression (2.14) for $\phi_m(x)$ are of the order $O((1-\rho)^{-(m+2-\ell)})$. Now, the coefficent of $h_\ell(x)$ in $\phi_{k+1}(x)$ (for $2 \leq \ell \leq k+2$) is

$$-c_{k+1}(\ell) + \sum_{m=0}^{k} \sum_{j=1}^{k+2-m} c_{k+1}(m+j) \binom{m+j}{j} EX^j b(m,\ell).$$

From (2.12), we see that $c_k(j) = O((1-\rho)^{-(k+2-j)})$, so that the coefficient of $h_\ell(x)$ in $\phi_{k+1}(x)$ is

$$\begin{aligned}
& O((1-\rho)^{-(k+3-\ell)}) + \sum_{m=0}^{k} \sum_{j=1}^{k+2-m} O((1-\rho)^{-(k+3-m-j)+I(j=1)-(m+2-\ell)}) \\
= \;& O((1-\rho)^{-(k+3-\ell)}) + \sum_{m=0}^{k} \sum_{j=1}^{k+2-m} O((1-\rho)^{-(k+5-I(j=1)-j+\ell)}) \\
= \;& O((1-\rho)^{-(k+3-\ell)}) + \sum_{j=1}^{k+2} O((1-\rho)^{-(k+5-I(j=1)-j+\ell)}) \\
= \;& O((1-\rho)^{-(k+3-\ell)}).
\end{aligned}$$

Thus, the claim is true for $k+1$, and we are done. ■

Henceforth, let π_ρ denote the stationary distribution of the waiting time.

Definition 2.1 *The rv X has an* exponential tail *if for some constants $c, \gamma > 0$, $P(X > x) \leq ce^{-\gamma x}$ for all $x \geq 0$.*

Lemma 2.2 *If $\bar{U}_1$ has an exponential tail, then there exists a $\gamma > 0$ such that the following results hold.*

1. *As* $\rho \uparrow 1$, $E_{\pi_\rho} e^{-\gamma W_0(\rho)} = O(1-\rho)$.

2. *For any* $\rho_0 > 0$, *there exists* $d_k = d_k(\rho_0) < \infty$ *such that*

$$\sup_{\rho_0 \le \rho \le 1} |h_k(x;\rho)| \le d_k e^{-\gamma x}.$$

Proof. The first statement in the lemma asserts that the Laplace Stieltjes transform (LST) of the stationary waiting time evaluated at γ is $O(1-\rho)$. In Section 2.3.5 we noted that

$$(1-\tilde{k}_\rho(s))\tilde{w}_\rho(s) = \frac{1-\tilde{h}_\rho(-s)}{EC},$$

where $\tilde{k}_\rho$, $\tilde{w}_\rho$ and $\tilde{h}_\rho$ are the LST's of the increment rv X, the stationary waiting time, and the idle time in the ρth system respectively. Our assumption that the interarrival times $\bar{U}_n/\rho$ satisfy $P(\bar{U}_n/\rho > x) \le ce^{-\gamma'\rho x}$ for some $c, \gamma' > 0$ guarantees that $\tilde{k}_\rho(s)$ and $\tilde{h}_\rho(-s)$ are finite for $0 \le s < \gamma'$. Now, pick $\gamma_1 \in (0, \gamma')$ so that $\tilde{k}_\rho(\gamma_1)$ is bounded away from 1 as $\rho \to 1$. (We can pick such a γ_1 since $\tilde{k}_\rho(0) = 1$, $\tilde{k}'_\rho(0) = -EX_1(\rho) \ge 0$ for all ρ, and $\tilde{k}''_\rho(0) = EX_1(\rho)^2$ which is bounded away from 0 as $\rho \to 1$.) We then find that

$$\begin{aligned} Ee^{-\gamma_1 W} &= \tilde{w}_\rho(\gamma_1) \\ &= \frac{1-\tilde{h}_\rho(-\gamma_1)}{EC(1-\tilde{k}_\rho(\gamma_1))}, \end{aligned}$$

and since $(1-\rho)EC$ is bounded away from 0 and ∞ (Proposition 2.3), we have the first result.

The second result of the lemma follows by a direct calculation as follows. For $y \le 0$,

$$\begin{aligned} P(X < y) &= P(V - \bar{U}/\rho < y) \\ &\le P(-\bar{U}/\rho < y) \\ &= P(\bar{U} > \rho|y|) \end{aligned}$$

Therefore X has an exponential left tail, i.e., $P(X < y) \leq ce^{\gamma' \rho y}$ for all $y < 0$. Let $x \geq 0$ and suppose that $\rho \geq \rho_0 > 0$. Using integration by parts,

$$\begin{aligned} |h_k(x)| &= (-1)^k \int_{-\infty}^{-x} (x+y)^k P(X \in dy) \\ &= (-1)^k [(x+y)^k P(X < y)]_{-\infty}^{-x} \\ &\quad + (-1)^{k+1} \int_{-\infty}^{-x} k(x+y)^{k-1} P(X < y)\, dy \\ &\leq k(-1)^{k+1} \int_{-\infty}^{-x} (x+y)^{k-1} c e^{\gamma' \rho y}\, dy \\ &= \frac{ck! e^{\gamma' \rho x}}{(\gamma' \rho)^k} \\ &\leq d_k e^{-\gamma_2 x}, \end{aligned}$$

where $d_k = ck!/(\gamma_2 \rho_0)^k$ and $\gamma_2 = \gamma' \rho_0$. Taking $\gamma = \min\{\gamma_1, \gamma_2\}$ yields the result. ■

Lemma 2.3 *If $\bar{U}_1$ has an exponential tail and a continuous distribution, then as $\rho \uparrow 1$, $\pi_\rho h_k(\cdot, \rho) = O(1-\rho)$, and*

$$E\left(\sum_{i=0}^{C-1} h_k(W_i)\right)^2 = O((1-\rho)^{-1}).$$

Proof. From Lemma 2.2,

$$|E h_k(W)| \leq d_k E e^{-\gamma W}$$

so that the first result follows. For the second result, note that

$$\begin{aligned} E\left(\sum_{i=0}^{C-1} h_k(W_i)\right)^2 &= 2E\left(\sum_{n=0}^{C-1} h_k(W_n) \sum_{m=n}^{C-1} h_k(W_m)\right) \\ &\quad - E\left(\sum_{i=0}^{C-1} h_k^2(W_i)\right), \end{aligned} \tag{2.40}$$

(the representation (2.40) has been fruitfully exploited previously; see Asmussen (1992)). The second term on the right-hand side of (2.40) is bounded by

$$E\sum_{i=0}^{C-1} d_k^2 e^{-2\gamma W_n} = d_k^2 EC E e^{-2\gamma W} \leq d_k^2 EC$$

which is $O((1-\rho)^{-1})$ as $\rho \uparrow 1$ (Proposition 2.3).

The first term in (2.40) is equal to

$$2E\left(\sum_{n=0}^{C-1} h_k(W_n) E\left[\sum_{m=n}^{C-1} h_k(W_m) \middle| W_n, C > n\right]\right),$$

which is bounded (in absolute value) by

$$2d_k^2 E\left(\sum_{n=0}^{C-1} e^{-\gamma W_n} E\left[\sum_{m=n}^{C-1} e^{-\gamma W_m} \middle| W_n, C > n\right]\right). \tag{2.41}$$

Let us turn our attention to the conditional expectation in (2.41). Observe that

$$E\left[\sum_{m=n}^{C-1} e^{-\gamma W_m} \middle| W_n, C > n\right] \leq E_{W_n} T,$$

where $T = \inf(m \geq 1 : W_m = 0) = \inf(m \geq 1 : S_m \leq 0)$ and $S_m = W_0 + X_1 + \cdots + X_m$. Suppose that $W_0 = x$, and let us determine $E_x T$. Since $S_0 = x$, $x - S_T = \sum_{i=1}^{T} X_i$. Wald's identity then gives $x - E_x S_T = -EXE_x T$, and hence $W_n - E_{W_n} S_T = -EXE_{W_n} T$. But $-E_{W_n} S_T = EI$, where I is the length of the first idle period. From Proposition 2.3 (which requires the continuity assumption on the distribution of $\bar{U}_1$), $EI \to c$ as $\rho \to 1$, where c is a finite constant, and so

$$E_{W_n} T_0 \leq \frac{W_n}{-EX} + \frac{2c}{-EX}$$

for ρ sufficiently close to 1.

Thus $E_{W_n} T_0 \leq (W_n + 2c)/(-EX)$, and so (2.41) is bounded by

$$\frac{2d_k^2}{-EX} E \sum_{n=0}^{C-1} e^{-\gamma W_n} (W_n + 2c).$$

But $(x + 2c)e^{-\gamma x} \leq e^{-\gamma x/2}$ for x larger than some constant a_1 say. Thus, (2.41) is bounded by

$$\begin{aligned} & \frac{a_2}{-EX} E\left(\sum_{n=0}^{C-1} (e^{-\gamma W_n/2} + a_3 I(W_n \leq a_1)\right) \\ = \; & \frac{1}{1-\rho}(a_4 EC E e^{-\gamma W/2} + a_5 EC P(W < a_1)), \end{aligned} \tag{2.42}$$

for some deterministic constants a_i $(1 \leq i \leq 5)$.

The second result now follows by noting that $Ee^{-\gamma W/2} = O(1-\rho)$, $EC = O((1-\rho)^{-1})$, and

$$\begin{aligned} P(W < a_1) &= P(e^{-\gamma W} > e^{-\gamma a_1}) \\ &\leq e^{\gamma a_1} E e^{-\gamma W} \\ &= O(1-\rho). \end{aligned}$$

■

Proof of Theorem 2.3. Recall from Chapter 1 that a sufficient condition for the estimator (2.11) to be consistent is that $\pi(Ag) = 0$. By Proposition 1.4 this holds if $\pi|g| < \infty$, and $g \in D(A_\rho)$. The condition $E\bar{V}_0^{k+2} < \infty$ ensures that the steady state waiting time has moments up to order $k+1$ and since $g(x)$ is a polynomial of order $k+1$, $\pi|g| < \infty$. To see that $g \in D(A_\rho)$ note that

$$\begin{aligned} E_x|g(W_1)| &= \sum_{i=0}^{k+1} a_i [x + X_1]_+^k \\ &\leq \sum_{i=0}^{k+1} |a_i| |x + X_1|^k \end{aligned}$$

where a_i are deterministic constants $(0 \leq i \leq k+1)$. Our assumptions imply that $E|X_1|^{k+1} < \infty$, which ensures that $E_x|g(W_1)| < \infty$ for every $x \geq 0$, and so $g \in D(A_\rho)$.

For $\rho < 1$, W_n is a regenerative sequence (regenerative state 0), so that a central limit theorem for $\alpha_k(n)$ follows from the regenerative central limit theorem (Wolff (1989) pp. 122–124), and we may obtain an expression for β^2 from that result. Note that ϕ_k is a bounded function (since the h_j's are all bounded functions) for any given $\rho < 1$. Furthermore, $EC^2 < \infty$ where $C = \inf\{ n \geq 1 : W_n = 0 \}$ (Proposition 2.3). These two results together imply that the conditions of the regenerative CLT are satisfied, hence

$$\beta^2 = \frac{E\left(\sum_{n=0}^{C-1} (\phi_k(W_n) - \pi_\rho \phi_k)\right)^2}{EC} < +\infty. \tag{2.43}$$

We will now determine the order of the numerator in (2.43). From (2.14)

$$\phi_k(x) - \pi_\rho \phi_k = \sum_{j=2}^{k+1} b(k,j)(h_j(x) - \pi_\rho h_j)$$

so that

$$\begin{aligned} \beta^2 EC &= E\left(\sum_{j=2}^{k+1} b(k,j) \sum_{n=0}^{C-1} (h_j(W_n) - \pi_\rho h_j)\right)^2 \\ &\leq k \sum_{j=2}^{k+1} b(k,j)^2 E\left(\sum_{n=0}^{C-1} (h_j(W_n) - \pi_\rho h_j)\right)^2. \end{aligned} \tag{2.44}$$

This inequality follows from the real variables result that $(a_1 + \cdots + a_m)^2 \leq m(a_1^2 + \cdots + a_m^2)$. A second application of this result yields

$$E\left(\sum_{n=0}^{C-1} (h_j(W_n) - \pi_\rho h_j)\right)^2 \leq 2E\left(\sum_{n=0}^{C-1} h_j(W_n)\right)^2 + 2(\pi_\rho h_j)^2 EC^2. \tag{2.45}$$

Now, $EC^2 = O((1-\rho)^{-3})$ (Proposition 2.3) and Lemma 2.2 shows that $\pi_\rho h_j = O(1-\rho)$, so that the second term on the right hand side of (2.45) is $O((1-\rho)^{-1})$. Lemma 2.3 proves that the first term on the right hand side of (2.45) is also $O((1-\rho)^{-1})$.

Combining these results and Lemma 2.1 with (2.44), we find that

$$\begin{aligned} \beta^2 EC &\leq k \sum_{j=2}^{k+1} b(k,j)^2 O((1-\rho)^{-1}) \\ &= \sum_{j=2}^{k+1} O((1-\rho)^{-2(k+2-j)}) O((1-\rho)^{-1}) \\ &= O((1-\rho)^{-2k-1}). \end{aligned}$$

But $(1-\rho)EC$ is bounded away from 0 and ∞ (Proposition 2.3), so that $\beta^2 = O((1-\rho)^{-2k})$. ■

Proof of Theorem 2.4. Once again, the proof that the estimator is consistent and satisfies a CLT can be shown using a regenerative analysis. We show this for the first moment. The case for higher moments is similar.

Define $Z_i = I_i^2 - 2rI_i$, where $r = EI_1^2/(2EI_1)$. The Z_i's are iid, and if $\sigma^2 = EZ_1^2$, then by the standard CLT for iid rv's,

$$n^{-1/2}\sum_{i=1}^{n} Z_i \Rightarrow \sigma N(0,1)$$

as $n \to \infty$. (Note that $\sigma^2 < \infty$ since $EV_0^5 + EU_1^5 < \infty$ implies that I_1 has finite 4th moments (Wolff (1989) p. 415). Now, by Anscombe's theorem (Chung (1974) p. 216), it follows that

$$N_n^{-1/2}\sum_{i=1}^{N_n} Z_i \Rightarrow \sigma N(0,1),$$

as $n \to \infty$, or equivalently (since $n/N_n \to EC$ a.s.),

$$n^{-1/2}\sum_{i=1}^{N_n} Z_i \Rightarrow \frac{\sigma}{\sqrt{EC}} N(0,1)$$

as $n \to \infty$ by the converging together lemma (see Billingsley (1986) Example 25.8). Now, since $n^{-1}\sum_{i=1}^{N_n} I_i \to EI/EC$ a.s., it follows by a second application of the converging together lemma that

$$n^{-1/2}\frac{\sum_{i=1}^{N_n} Z_i}{2\sum_{i=1}^{N_n} I_i} \Rightarrow \frac{\sigma EC}{2EI\sqrt{EC}} N(0,1),$$

or equivalently,

$$\sqrt{n}(\alpha_n - EW) \Rightarrow \eta N(0,1)$$

as $n \to \infty$, where

$$\eta^2 = \sigma^2 EC/(4(EI)^2). \tag{2.46}$$

We have thus proved the required CLT for the Minh-Sorli estimator of EW.

We now prove that the TAVC of the Minh-Sorli estimator of EW^k is of the order $O((1-\rho)^{-2k+1})$ as $\rho \uparrow 1$ by induction on k. First note that EI is bounded away from 0 as $\rho \to 1$ (Proposition 2.3). Now $EC = O((1-\rho)^{-1})$, and σ^2 remains bounded as $\rho \uparrow 1$ (see Lemma 4.1 of Asmussen (1992), and note that the argument given there for boundedness of EI_ρ as $\rho \uparrow 1$ extends easily to the case of higher moments). The TAVC for the Minh-Sorli estimator of the first moment is therefore $O((1-\rho)^{-1})$, and the result is true for $k = 1$.

Suppose the result is true for $1 \leq j < k$. We may differentiate (2.15) $k+1$ times, and evaluate the result at $s = 0$ to obtain

$$\sum_{j=1}^{k+1} \binom{k+1}{j} (-1)^k EX^j EW^{k+1-j} = -EI^{k+1}/EC_1.$$

Rearranging, we obtain

$$EW^k = \frac{1}{(k+1)EX} \left[(-1)^{k+1} \frac{EI^{k+1}}{EC_1} - \sum_{j=2}^{k+1} \binom{k+1}{j} EX^j EW^{k+1-j} \right]. \tag{2.47}$$

The Minh-Sorli estimator of EW^k is obtained by estimating the term EI^{k+1} and substituting in the Minh-Sorli estimators for the moments EW^{k+1-j} $(j \geq 2)$. By the inductive hypothesis, the dominating term in (2.47) (in terms of TAVC) is

$$\frac{EX^2 EW^{k-1}}{EX}.$$

Since $EX(1) = -(1-\rho)/\lambda$, the estimator of this term has a TAVC that is $O((1-\rho)^{(-2(k-1)+1)-2})$ and so the proof is complete. ■

Proof of Theorem 2.5. Recall from Chapter 1 that the AMP estimator will be consistent if $\pi Ag = 0$, where π is the distribution of the steady-state waiting time. From Proposition 1.4 it suffices to show that $g \in D(A)$ and $\pi|g| < \infty$. Since $g(x) \leq a_1 + a_2 x$ for some deterministic constants a_1 and a_2, the π-integrability of g will follow if $EW < \infty$. A sufficient condition for this is that $EV_0^2 < \infty$ and $\rho < 1$. We must also show that $g \in D(A)$, but for $x \geq 0$,

$$\begin{aligned} E_x g(W_1) &\leq a_1 + a_2 E_x W_1 \\ &\leq a_1 + a_2(x + E|X_1|) \end{aligned}$$

so that $g \in D(A)$ follows immediately from $E|X_1| < \infty$. ■

2.7.2 The Queue Size Process

Proof of Proposition 2.1.

Let $Q(s)$ be the number of customers in the system at time s, $\tilde{W}_k$ be the waiting time (including service) of the kth customer to arrive, C be the number of customers served in the first busy period, and τ be the end of the busy period (the first time after time 0 that a customer arrives to an empty system). Define q, w and $\tilde{w}$ to be the mean steady-state queue size, waiting time excluding service, and waiting time including service respectively. From Little's law, $q = \lambda\tilde{w}$ (Wolff (1989) p. 235). Let η^2 be the TAVC for the standard estimator (2.26) of the queue size, σ^2 be the TAVC for the standard estimator (2.5) of the expected waiting time excluding service, and let $\tilde{\sigma}^2$ be the TAVC for the estimator of the waiting time including service, given by

$$\frac{1}{n}\sum_{k=0}^{n-1}\tilde{W}_k.$$

Expressions for the above TAVC's may be obtained from the regenerative CLT. We obtain

$$\begin{aligned}
\eta^2 &= \frac{E\left(\int_0^\tau Q(s)\,ds - \tau q\right)^2}{E\tau}, \\
\sigma^2 &= \frac{E\left(\sum_{k=0}^{C-1} W_k - Cw\right)^2}{EC}, \text{ and} \\
\tilde{\sigma}^2 &= \frac{E\left(\sum_{k=0}^{C-1} \tilde{W}_k - C\tilde{w}\right)^2}{EC}.
\end{aligned}$$

Note that $\tau = \sum_{i=1}^{C} U_i$, so that by Wald's first moment identity, $E\tau = EC/\lambda$. Furthermore, $\int_0^\tau Q(s) = \sum_{k=0}^{C-1}\tilde{W}_k$, so that

$$\begin{aligned}
\eta^2 &= \frac{\lambda}{EC}E\left(\sum_{k=0}^{C-1}(\tilde{W}_k - qU_k)\right)^2 \\
&= \frac{\lambda}{EC}E\left(\sum_{k=0}^{C-1}(\tilde{W}_k - \tilde{w}) + (\tilde{w} - \lambda\tilde{w}U_k)\right)^2 \\
&= \lambda\tilde{\sigma}^2 + \frac{\lambda^3\tilde{w}^2}{EC}E\left(\sum_{k=1}^{C}\lambda^{-1} - U_k\right)^2 \\
&\quad + \frac{2\lambda^2\tilde{w}}{EC}E\left(\sum_{k=0}^{C-1}(\tilde{W}_k - \tilde{w})\sum_{j=0}^{C-1}(\lambda^{-1} - U_j)\right)
\end{aligned}$$

$$= \lambda(\sigma^2 + \text{var}\, V_0) + \lambda^3 \tilde{w}^2 \text{var}\, U_1 + \frac{2\lambda^2 \tilde{w}}{EC} E\left(\sum_{k=0}^{C-1} (\tilde{W}_k - \tilde{w}) \sum_{j=0}^{C-1} (\lambda^{-1} - U_j) \right)$$

where the last equality follows from Wald's second moment identity.

Now consider the magnitudes of the terms in this expression for η^2. By Theorem 2.2, the first term is of the order $O((1-\rho)^{-4})$ as $\rho \uparrow 1$. Since $\tilde{w}$ is of the order $O((1-\rho)^{-1}$ as $\rho \uparrow 1$ (Asmussen (1992)), the second term is $O((1-\rho)^{-2})$. The Cauchy-Schwartz inequality shows that the last term is $O((1-\rho)^{-3})$ as $\rho \uparrow 1$, and so the proof is complete. ■

Proof of Proposition 2.2.

Define the kth factorial power of x ($x^{\underline{k}}$) by

$$x^{\underline{k}} = x(x-1)\cdots(x-k+1),$$

and let $f(x) = x^{\underline{k}}$. We will show that if $Q(s)$ is the queue size at time s, $Q(\infty)$ is the stationary queue size ($Q(s) \Rightarrow Q(\infty)$) and σ_k^2 is the TAVC for the estimator

$$\frac{1}{t} \int_0^t f(Q(s))\, ds$$

of $Ef(Q(\infty))$, then $\sigma_k^2 = O((1-\rho)^{-(2k+2)})$. This is of course equivalent to showing that the TAVC for the estimator

$$\frac{1}{t} \int_0^t Q(s)^k\, ds$$

of $EQ(\infty)^k$ is $O((1-\rho)^{-(2k+2)})$. The reason we choose to work with factorial moments is that the computations work out far more cleanly than if we worked with standard moments.

Let π be the stationary distribution of the M/M/1 queue size, so that

$$\begin{aligned} \pi f &= (1-\rho) \sum_{n=k}^{\infty} n^{\underline{k}} \rho^n \\ &= k! \rho^k (1-\rho)^{-k} \stackrel{\triangle}{=} \alpha \text{ say.} \end{aligned}$$

Define $f_c(x) = f(x) - \alpha$, and let g be a solution to Poisson's equation $Ag = -f_c$, where A is the the rate matrix for the M/M/1 queue size CTMC. To be precise, $-\lambda g_0 + \lambda g_1 = \alpha$, and for $n \geq 1$,

$$\mu g_{n-1} - (\lambda + \mu) g_n + \lambda g_{n+1} = \alpha - n^{\underline{k}} \tag{2.48}$$

Set $g_0 = 0$ and $G(s) \stackrel{\triangle}{=} \sum_{n=0}^{\infty} g_n s^n$. Summing (2.48) from 1 to ∞ we obtain (after some manipulation)

$$\begin{aligned} G(s) &= \frac{k! s(\rho^k (1-s)^k - (1-\rho)^k s^k)}{\mu (1-\rho)^k (1-s)^{k+2} (\rho - s)} \\ &= \frac{k!}{\mu} \sum_{j=0}^{k-1} \frac{\rho^j}{(1-\rho)^{j+1}} s^{k-j} (1-s)^{-(k+2-j)}. \end{aligned} \tag{2.49}$$

Now, Theorem 4.3 of Glynn and Meyn (1993) provides us with an expression for σ_k^2, which in our context yields

$$\begin{aligned} \sigma_k^2 &= 2 \sum_{n=0}^{\infty} (1-\rho) \rho^n (n^{\underline{k}} - \alpha) g_n \\ &= 2(1-\rho) \rho^k G^{(k)}(\rho) - 2\alpha (1-\rho) G(\rho) \end{aligned}$$

where $G^{(k)}(\rho)$ is the kth derivative of $G(s)$ evaluated at $s = \rho$. From (2.49),

$$G(\rho) = \frac{k(k!)\rho^k}{\mu (1-\rho)^{k+3}}$$

so that $2\alpha(1-\rho)G(\rho) = O((1-\rho)^{-2k-2})$. The proof will therefore be complete if we show that $G^{(k)}(\rho) = O((1-\rho)^{-2k-3})$.

The kth derivative of a product $h_1(s)h_2(s)$ is given by

$$\frac{d^k (h_1(s) h_2(s))}{ds^k} = \sum_{i=0}^{k} \binom{k}{i} h_1^{(i)}(s) h_2^{(k-i)}(s)$$

so that from (2.49)

$$\begin{aligned} G^{(k)}(s) &= \frac{k!}{\mu} \sum_{j=0}^{k-1} \frac{\rho^j}{(1-\rho)^{(j+1)}} \sum_{i=0}^{k-j} \binom{k}{i} (-1)^{k-i} \times \\ &\quad \frac{(2k+1-i-j)!}{(k+1-j)(k-j-i)!} s^{k-j-i} (1-s)^{-(2k+2-j-i)}. \end{aligned}$$

Therefore,

$$\begin{aligned} G^{(k)}(\rho) &= \frac{k!}{\mu}\sum_{j=0}^{k-1}\sum_{i=0}^{k-j}\binom{k}{i}(-1)^{k-i}\frac{(2k+1-i-j)!}{(k+1-j)(k-j-i)!}\frac{\rho^{k-i}}{(1-\rho)^{2k+3-i}} \\ &= O((1-\rho)^{-(2k+3)}) \end{aligned}$$

as required. ■

Our principal tool in proving Theorem 2.7 will be regenerative process theory. Recall that the process X is a regenerative process with respect to the regeneration times $(T_n : n \geq 0)$ (the times when a customer arrives to an empty system). Since we are assuming that at time 0 a customer arrives to an empty system, the regenerative process is non-delayed. We emphasize that the process X depends on ρ, but we suppress that dependence in our notation. (Recall that in the heavy traffic formulation that is appropriate here, the ρth process has interarrival times $\bar{U}_n/\rho$ and service times $\bar{V}_n$, where $E\bar{U}_n = E\bar{V}_n = \mu^{-1}$).

Let us restate our assumptions on the interarrival and service time rv's from Section 2.2.2. First, the distribution functions of $\bar{U}_n$ and $\bar{V}_n$ possess densities (so that we may define their hazard rate functions). Further, we assume that in estimating EQ^k, $E\bar{V}_0^{k+1} < \infty$, so that the moment exists (for $k = 1, 2$). We now assume these conditions throughout the remainder of this section.

Under the above assumptions, $X(t) \Rightarrow X(\infty)$ as $t \to \infty$ (Asmussen (1987) p. 126). Let $X(\infty)$ have distribution π, and let $\mathcal{S} \subseteq \mathbb{N} \times \mathbb{R}^+ \times \mathbb{R}^+$ be the state space of X.

In determining the sizes of the TAVC's associated with the estimators (in heavy traffic), we will first prove some slightly more general results, and then specialize to our particular situation. Let $\tau_\rho = T_1$ be the length of the first regenerative cycle, and let $Y_\rho(\phi) \stackrel{\triangle}{=} \int_0^{\tau_\rho} \phi(X(s))\, ds$, for each $\phi : \mathcal{S} \to \mathbb{R}$. Suppose that $E(Y_\rho(|\phi|)^2 + \tau_\rho^2) < \infty$. Then by the regenerative CLT (Wolff 1989 p. 124),

$$\sqrt{t}\left(t^{-1}\int_0^t \phi(X(s))\, ds - r_\rho\right) \Rightarrow \sigma_\rho N(0,1),$$

where

$$\sigma_\rho^2 = \frac{E\left(\int_0^{\tau_\rho} \phi(X(s))\, ds - r_\rho \tau_\rho\right)^2}{E\tau_\rho}, \tag{2.50}$$

and $r_\rho = \pi\phi$. We will use this representation of the TAVC σ_ρ^2 to obtain our results.

Let C_ρ be the number of customers served in the first cycle. Our first result gives the order of $E\tau_\rho^k$, for $k = 1, 2$.

Proposition 2.4 *If $E(\bar{U}_1^5 + \bar{V}_0^5) < \infty$, then the first two moments of τ_ρ have the same order as those of C_ρ. Specifically, $E\tau_\rho^k = O((1-\rho)^{-2k+1})$ for $k = 1, 2$. In addition, $(1-\rho)E\tau_\rho$ is bounded away from 0 and ∞ as $\rho \uparrow 1$.*

Proof. From Proposition 2.3 (which requires the hypothesis that $E(\bar{U}_1^5 + \bar{V}_0^5) < \infty$), $EC_\rho^k = O((1-\rho)^{-2k+1})$, for $k = 1, 2$, and furthermore, $(1-\rho)EC_\rho$ is bounded away from 0 and ∞ as $\rho \uparrow 1$.

Now $\tau_\rho = \sum_{n=1}^{C_\rho} U_n$. Since C_ρ is a stopping time with respect to $\mathcal{F}_n = \sigma\{(U_i, V_{i-1}) : 1 \le i \le n\}$, C_ρ is a *randomized* stopping time with respect to $\sigma\{U_i : 1 \le i \le n\}$. We may therefore apply Wald's first moment identity, to obtain $E\tau_\rho = \lambda^{-1} EC_\rho$. This relation establishes our result for $k = 1$. For $k = 2$ we may invoke Wald's second moment identity (Chow, Robbins and Teicher (1965)), to establish that

$$\begin{aligned} E\tau_\rho^2 &= E\left(\sum_{i=1}^{C_\rho}(U_i - \lambda^{-1}) + \lambda^{-1}C_\rho\right)^2 \\ &\le 2E\left(\sum_{i=1}^{C_\rho}(U_i - \lambda^{-1})\right)^2 + 2\lambda^{-2}EC_\rho^2 \\ &= 2\mathrm{var}\, U EC_\rho + 2\lambda^{-2}EC_\rho^2 \\ &= O((1-\rho)^{-3}). \blacksquare \end{aligned}$$

For notational convenience, if c is a scalar, and ϕ is a function, we define $\phi - c$ to be the function $\bar{\phi}$ given by $\bar{\phi}(x) = \phi(x) - c$.

Lemma 2.4 *Suppose $\phi(q, t_a, t_s) = I(q = 0)f(t_a)$, where f is bounded, and let $y_\rho = EY_\rho(\phi)$. If $E\bar{U}_1^5 + \bar{V}_0^5 < \infty$, and $\bar{U}_1$ has a continuous distribution, then*

$$var(Y_\rho(\phi) - y_\rho \tau_\rho / E\tau_\rho) = O((1-\rho)^{-1}).$$

Proof. Since $\text{var}\,(X+Y) \le 2\text{var}\,X + 2\text{var}\,Y$,

$$\begin{aligned} \text{var}\,(Y_\rho(\phi) - y_\rho\tau_\rho/E\tau_\rho) &\le 2\text{var}\,Y_\rho(\phi) + 2y_\rho^2 \text{var}\,\tau_\rho/(E\tau_\rho)^2 \\ &\le 2\|f\|^2 EI_\rho^2 + 2\|f\|^2 (EI_\rho)^2 E\tau_\rho^2/(E\tau_\rho)^2 \end{aligned} \tag{2.51}$$

where I_ρ is the first idle period (the period at the end of the first regenerative cycle when $Q(s) = 0$). The first term on the right-hand side of (2.51) converges to a finite constant as $\rho \uparrow 1$ (Asmussen (1992), Lemma 4.1 shows that this holds for EI_ρ, but the argument extends easily to the second moment), and is therefore bounded in ρ. Similarly $(EI_\rho)^2$ is bounded in ρ as $\rho \uparrow 1$, and so Proposition 2.4 then allows us to conclude that the second term on the right-hand side of (2.51) is $O((1-\rho)^{-1})$. ■

Lemma 2.5 *If $\phi(q, t_a, t_s) = f(t_a, t_s)$, and f is bounded, then $var\, Y_\rho(\phi - \alpha_\rho) = O((1-\rho)^{-3})$, where $\alpha_\rho = EY_\rho(\phi)/E\tau_\rho$.*

Proof. Let k be the uniform bound on f. Then $|Y_\rho(\phi - \alpha_\rho)| \le 2k\tau_\rho$, and so $\text{var}\, Y_\rho(\phi - \alpha_\rho) \le 4k^2 E\tau_\rho^2$. Proposition 2.4 then yields the result. ■

Proof of Theorem 2.7. First consider the case $k = 1$. The estimating function is given by $f + Ag$, where $f(q, t_a, t_s) = q$, and $g = (\mu - \lambda)^{-1}(g_0/2 - g_1 + g_2 + g_3)$. Recall that π is the stationary distribution of X. Since $EV_0^2 < \infty$, $\pi f = EQ < \infty$, and so to show consistency, it suffices (by Proposition 1.4 of Chapter 1) to show that $g \in D(A)$, and both g and Ag are π-integrable, where A is the (extended) generator of X. Since the functions φ_1 and φ_2 are bounded,

$$|g(q, t_a, t_s)| \le a_0 + a_1 q + a_2 q^2$$

for some deterministic constants a_0, a_1 and a_2. But $EV_0^3 < \infty$, so $EQ^2 < \infty$ and hence g is π-integrable. Similarly,

$$|Ag(q, t_a, t_s)| \leq a_3 + a_4 r_a(t_a) + a_5 r_s(t_s) I(q > 0) + q,$$

and therefore Ag is π-integrable. It remains to show that $g \in D(A)$. But this follows immediately from the result quoted in Section 2.2.2.

It is easy to see that the AMP estimator satisfies a CLT. Under our assumptions, the estimating function ϕ_1 is bounded, and since $E\tau^2 < \infty$ (Proposition 2.4), the conditions of the regenerative CLT are satisfied.

Finally, Lemma 2.4 allows us to conclude that $\operatorname{var} Y_\rho(\phi_1 - EQ) = (1-\rho)^{-2} O((1-\rho)^{-1})$, or $O((1-\rho)^{-3})$, and since $\sigma_\rho^2 = \operatorname{var}(Y_\rho(\phi_1 - EQ))/E\tau_\rho$, the second result of Proposition 2.4 yields the result.

The proof for the case $k = 2$ is virtually identical to the proof for $k = 1$, except that we must show that the contribution to the time average variance by the term $I(q > 0)\varphi_1(t_a)\varphi_2(t_s)$ is $O((1-\rho)^{-4})$. This follows by Lemma 2.5. ■

Proof of Theorem 2.8. The estimator of the tail probability $P(Q \geq q)$ is given by

$$t^{-1} \int_0^t (f + Ag)(X(s))\, ds$$

where $f(n, t_a, t_s) = I(Q \geq q)$, A is the generator of the process X, and g is the function $g_1 + g_2 - g_3 - g_4$. Now, if π is the stationary distribution of X, then f is clearly π-integrable. Notice also that under our assumptions g is bounded by a function that is linear in n, the number of customers in the system. Therefore, since $EV_0^2 < \infty$, $EQ < \infty$ and so g is π-integrable. Furthermore

$$\begin{aligned} |Ag(n, t_a, t_s)| &\leq a_1 + a_2 r_a(t_a) I(n = q-1) + a_3 r_s(t_s) I(n = q) \\ &\leq a_1 + a_2 r_a(t_a) + a_3 r_s(t_s) I(n > 0) \end{aligned}$$

for some deterministic constants a_1, a_2 and a_3, and this bound is clearly π-integrable, so that Ag is also π-integrable. According to Proposition 1.4 of Chapter 1, to show

that the estimator (2.38) is consistent, it remains to show that $g \in D(A)$. But this follows immediately from the result quoted in Section 2.2.2. ■

Bibliography

Abate, J., Choudhury, G. L., Lucantoni, D. M., and Whitt, W. (1995). Asymptotic analysis of tail probabilities based on the computation of moments. *Ann. Appl. Probab.* **5**, 983–1007.

Andradóttir, S., Calvin, J. M., Glynn, P. W. (1995). Accelerated regeneration for Markov chain simulations. *Probab. Engrg. Inform. Sci.* **9** 497–523.

Asmussen, S. (1987). *Applied Probability and Queues.* John Wiley and Sons.

Asmussen, S. (1990). Exponential families and regression in the Monte Carlo study of queues and random walks. *Ann. Statist.* **18** 1851–1867.

Asmussen, S. (1992). Queueing simulation in heavy traffic. *Math. Oper. Res.* **17** 84–111.

Athreya, K. B. and Ney, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.* **245** 493–501.

Azema, J., Duflo, M. and Revuz, D. (1969). Proprietes relatives des processus de Markov. *Z. Wahrsch. verw. Gebeite* **13** 286–314.

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (1993). *Nonlinear Programming: Theory and Algorithms.* John Wiley and Sons.

Billingsley, P. (1986). *Probability and Measure, 2nd ed.* Wiley, New York.

Bratley, P., Fox, B. L., and Schrage, E. L. (1987). *A Guide to Simulation, 2nd Ed.* Springer-Verlag.

Breiman, L. (1968). *Probability.* Addison Wesley.

Burman, D. Y. (1981). Insensitivity in Queueing Systems. *Adv. Appl. Probab.* **13** 846–859.

Chow, Y. S., Robbins, H., and Teicher, H. (1965). Moments of randomly stopped

sums. *Ann. Math. Statist.* **36** 789–799.

Chung, K. L. (1967). *Markov Chains with Stationary Transition Probabilities.* Springer-Verlag, New York.

Chung, K. L. (1974). *A Course in Probability Theory, 2nd ed.* Academic Press.

Ethier, S. and Kurtz, T. (1986). *Markov Processes: Characterization and Convergence.* John Wiley and Sons.

Gaver, D. P. and Thompson, G. L. (1973). *Programming and Probability Models in Operations Research.* Wadsworth Publishing Co., Belmont California.

Glynn, P. W. (1982). *Simulation output analysis for general state space Markov chains.* Ph.D. Thesis, Dept. of Operations Research, Stanford University.

Glynn, P. W. (1988). A GSMP formalism for discrete event systems. *Proc. IEEE* **77** 14–23.

Glynn, P. W. (1990). Diffusion Approximations. In D. P. Heyman and M. J. Sobel, Eds., *Handbooks on OR & MS, Vol. 2.* Elsevier Science Publishers B.V. (North-Holland).

Glynn, P. W. (1994). Poisson's equation for the recurrent M/G/1 queue. *Adv. Appl. Prob.* **26** 1044–1062.

Glynn, P. W. (1994a). Efficiency improvement techniques. *Ann. Oper. Res.* **53** 175–197.

Glynn, P. W. and Iglehart, D. L. (1987). A joint central limit theorem for the sample mean and regenerative variance estimator. *Ann. Oper. Res.* **8** 41–45.

Glynn, P. W. and Iglehart, D. L. (1993). Conditions for the applicability of the regenerative method. *Man. Sci.* **39** 1108–1111.

Glynn, P. W. and Meyn, S. P. (1996). A Liapounov bound for solutions of the Poisson equation. *Ann. Probab.* **24** 916–931.

Glynn P. W. and Torres, M. (1996). Nonparametric estimation of tail probabilities for the single-server queue. In *Stochastic Networks: Stability and Rare Events.* Glasserman, P., Sigman, K. and Yao, D. D., Editors. Springer-Verlag.

Glynn, P. W. and Whitt, W. (1989). Indirect estimation via $L = \lambda W$. *Oper. Res.* **37** 82–103.

Glynn, P. W. and Whitt, W. (1992). The asymptotic efficiency of simulation estimators. *Oper. Res.* **40** 505–520.

Glynn, P. W. and Whitt, W. (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Probab.* **31A** 131–156.

Gross, D. and Harris, C. M. (1985). *Fundamentals of Queueing Theory, 2nd Ed.* John Wiley and Sons.

Harrison, J. M. (1990). *Brownian Motion and Stochastic Flow Systems, 2nd Ed.* John Wiley and Sons.

Ho, Y. C. (editor). (1991). *Discrete Event Dynamical Systems: Analyzing Complexity in the Modern World.* IEEE Press.

Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes.* Academic Press, New York.

Karlin, S. and Taylor, H. M. (1981). *A Second Course in Stochastic Processes.* Academic Press, New York.

Kiefer, J. and Wolfowitz, J. (1956). On the characteristics of the general queueing process, with applications to random walk. *Ann. Math. Statist.* **27** 147–161.

Kurtz, T. G. (1969). Extensions of Trotter's operator semigroup approximation theorems. *J. Functional Analysis.* **3** 354–375.

Law, A. M. (1975). Efficient estimators for simulated queueing systems. *Man. Sci.* **22** 30–41.

Law, A. M. and Kelton, W. D. (1991). *Simulation Modeling and Analysis, 2nd ed.* McGraw-Hill.

Lemoine, A. J. (1978). Networks of queues — A survey of weak convergence results. *Man. Sci.* **24** 1175–1193.

Lindvall, T. (1986). On coupling of renewal processes with use of failure rates. *Stochastic Process. Appl.* **22** 1–15.

Loh, W. W. (1994). *On the Method of Control Variates.* Ph.D. Thesis, Dept. of Operations Research, Stanford University.

Mandl, P. (1968). *Analytical Treatment of One-dimensional Markov Processes.* Springer-Verlag, New York.

Marshall, K. T. (1968). Some relationships between the distributions of waiting time,

idle time, and interoutput time in the GI/G/1 queue. *SIAM J. Appl. Math.* **16** 324–327.

Minh, D. L. (1987). Simulating GI/G/k queues in heavy traffic. *Man. Sci.* **33** 1192–1199.

Minh, D. L. and Sorli, R. M. (1983). Simulating the GI/G/1 queue in heavy traffic. *Oper. Res.* **31** 966–971.

Miyazawa, M. (1979). A formal approach to queueing processes in the steady state and their applications. *J. Appl. Probab.* **16**, 332–346.

Ross, S. M. (1983). *Stochastic Processes.* John Wiley and Sons.

Shedler, G. S. (1993). *Regenerative Stochastic Simulation.* Academic Press, Boston.

Sigman, K. (1990). One-dependent regenerative processes and queues in continuous time. *Math. Oper. Res.* bf 15 175–189.

Stroud, A. H. (1971). *Approximate Calculation of Multiple Integrals.* Prentice-Hall, Englewood Cliffs, New Jersey.

Torres, M. and Glynn, P. W. (1996). Parametric estimation of tail probabilities for the single-server queue. In *Frontiers in Queueing.* Dshalalow, E., Ed. CRC Press.

Wolff, R. W. (1989). *Stochastic Modeling and the Theory of Queues.* Prentice Hall, Englewood Cliffs, NJ.