The Annals of Statistics
1990, Vol. 18, No. 4, 1851–1867

EXPONENTIAL FAMILIES AND REGRESSION IN THE MONTE CARLO STUDY OF QUEUES AND RANDOM WALKS[1]

BY SØREN ASMUSSEN

University of California, Santa Barbara

An importance sampling method studied by Siegmund and Asmussen in the case of waiting time probabilities is extended to the mean and other functionals. The ideas involve exponential families of queues and control variates (regression), and it is found both from theory and practice that the method is dramatically better than standard tools like regenerative simulation or sample-mean estimation, not least under heavy traffic conditions.

1. Introduction. First passage time probabilities for a random walk $\{S_n\}$ come up in a variety of contexts in applied probability and statistics. In sequential analysis, a basic problem is to evaluate quantities associated with $\tau(a, b) = \inf\{n \geq 1\colon\ S_n < a \text{ or } S_n > b\}$, where $a < 0 < b$. In insurance risk, the emphasis is on the one-barrier problem and the ruin probabilities $\mathbb{P}(\tau(u) < \infty)$, where $\tau(u) = \inf\{n \geq 1\colon\ S_n > u\}$. In queueing theory, the quantities of main interest are the waiting time distribution $\mathbb{P}(M \leq u) = \mathbb{P}(\tau(u) = \infty)$, where $M = \max_{n \geq 0} S_n$ and (even more) functionals like the mean waiting time $\mathbb{E}M$.

Explicit expressions are, however, typically not available and simulation evaluation has therefore become a widespread and popular tool. However, the methods which are in use in practice tend to be crude and not always efficient. For example, in the queueing setting, it has been noted repeatedly (e.g., [13], [4] and [22]) that heavy traffic conditions present a particular challenge and that the standard methods may here require exceedingly long simulation runs to perform satisfactorily.

One frequent argument against more sophisticated variance reduction methods is that the ideas may work beautifully in oversimplified settings like the Monte Carlo evaluation of simple integrals, but it is most often not clear how to adapt such lines in more complex and realistic situations. There may often be some truth in this, nevertheless, some investigations have appeared which show that a more careful analysis of the specific structure of the problem in question can sometimes be exploited with great advantage. We are here concerned with one such particular technique, which was introduced by Siegmund [17] in the sequential analysis setting and was further studied by the author [1] for compound Poisson ruin probabilities. The method is based upon likelihood ratio identities for exponential families of stopped random walks and yields a very substantial variance reduction. From the queueing

Received April 1988; revised January 1990.
[1]Research supported by NSF Grant DMS-89-01556.
AMS 1980 *subject classifications*. Primary 65C05; secondary 62J05, 62M05.
Key words and phrases. Monte Carlo method, variance reduction, importance sampling, control variates, regression, heavy traffic, waiting time, random walk.

point of view, a main drawback is, however, that it is not apparent how to deal with quantities like the mean waiting time rather than the waiting time distribution, and this seems a must in that framework.

The purpose of the present paper is to present one possible approach for resolving this problem. The additional ideas involve a certain randomisation (which avoids estimating infinite integrals) as well as the method of linear control variates, which in turn is closely related to standard linear regression (also further regression schemes will be shown to be of possible relevance).

Though, essentially, we are dealing with a random walk problem, some of the main applications seem to be in queueing and the details will therefore be worked out in that setting. We start in Section 2 by introducing the basic setup and suggesting our Monte Carlo approach. Section 3 contains some empirical investigations of the efficiency of the method, which show that the variance reduction compared to standard methods may be very substantial (for example, it is indicated that in some particular settings the gain amounts to a factor of several thousands compared to the regenerative method). Section 4 then contains a theoretical study of the estimation method. The basic performance measures are evaluated, it is shown that in marked contrast to established methods the efficiency increases under heavy traffic conditions and we give some results helpful to find the most efficient way to carry out the randomisation step. In Section 5, we discuss the idea of analysing the simulation output by means of a weighted regression with normally distributed errors. Finally, Section 6 contains a reinspection of the empirical material of Section 3 in light of the theoretical insight obtained, as well as some concluding remarks, and in particular we outline an example of the applicability of the method outside the simple random walk setting.

2. The Monte Carlo method. We use throughout the notation of [2], so that $U_0, U_1, \ldots$ are the service times and $B(u) = \mathbb{P}(U \le u)$ is the service time distribution. Similarly, the interarrival times are $T_0, T_1, \ldots$, the interarrival distribution is $A(t) = \mathbb{P}(T \le t)$ and thus the traffic intensity is $\rho = \mathbb{E}U/\mathbb{E}T$. We assume $\rho < 1$ so that a steady-state limit W of the waiting-time process exists, and it is a standard fact that W is distributed as the maximum $M = \max_{n \ge 0} S_n$ of a random walk $S_n = X_0 + \cdots + X_{n-1}$, where $X_k = U_k - T_k$. That is,

$$\mathbb{P}(W > u) = \mathbb{P}(M > u) = \mathbb{P}(\tau(u) < \infty),$$

where $\tau(u) = \inf\{n \ge 1\colon S_n > u\}$. The moment generating functions are defined by $\hat{A}(s) = \mathbb{E}e^{sT}$ and $\hat{B}(s) = \mathbb{E}e^{sU}$ and we assume that B has enough exponential moments to ensure that a solution $\gamma > 0$ of $\hat{A}(-\gamma)\hat{B}(\gamma) = 1$ exists (this equation is of fundamental importance in random walk theory and we use notation corresponding to the name *Lundberg equation* in common use in risk theory). Under heavy traffic conditions, which are the typical setup of this paper, γ is close to zero. For example, for M/M/1 with arrival intensity ρ and service intensity 1, which will be used for illustration in much of the paper, $\gamma = 1 - \rho$. We define new distributions B_L and A_L of service and interarrival

times by

$$B_L(dx) = \frac{e^{\gamma x}}{\hat{B}(\gamma)} B(dx), \qquad A_L(dx) = \frac{e^{-\gamma x}}{\hat{A}(-\gamma)} A(dx).$$

Letting $\mathbb{P}_L$ and $\mathbb{E}_L$ refer to a queue with these distributions of service and interarrival times, we have ([2], Chapter 12) that $\mathbb{P}_L(\tau(u) < \infty) = 1$ and

$$\mathbb{P}(W > u) = e^{-\gamma u} \mathbb{E}_L e^{-\gamma B(u)}$$

where $B(u) = S_{\tau(u)} - u$. This relation is exploited for simulation purposes in [17] and [1] by noting that an unbiased simulation estimate of $\mathbb{P}(W > u)$ can be obtained as the sample mean of replications of $e^{-\gamma u} e^{-\gamma B(u)}$ drawn according to $\mathbb{P}_L$. The reason that this is efficient is that $e^{-\gamma B(u)}$ is close to 1 and, in particular, has a small variance which at least for large or moderate u is even further damped by the factor $e^{-\gamma u}$. For example, for M/M/1 with $\rho = 0.9$ and u chosen such that $\mathbb{P}(W > u) = 5\%$, a test run reported in [1] yielded a variance reduction by a factor of 432 compared to the regenerative method.

The objective of the present paper is to make these ideas work also for the evaluation of functionals like the mean, the second moment, exponential moments and so on. As a simple and fundamental example, consider the mean

$$\mu = \mu_W = \mathbb{E}W = \int_0^\infty \mathbb{P}(W > u)\, du = \int_0^\infty e^{-\gamma u} \mathbb{E}_L e^{-\gamma B(u)}\, du. \tag{2.1}$$

Our main idea is now to avoid estimating an infinite integral not by truncating at a fixed large value but rather at a moderate random value L which is independent of the queueing process and which for the moment we take to be exponential, $\mathbb{P}(L > u) = e^{-\alpha u}$. To avoid introducing bias, we then must correct by $1/\mathbb{P}(L > a)$ and arrive at

$$Y = \int_0^L e^{\alpha u} e^{-\gamma u} e^{-\gamma B(u)}\, du \tag{2.2}$$

as an unbiased (w.r.t. $\mathbb{P}_L$) estimator of μ.

This idea is not particularly useful in itself because the variance of Y is rather large. In fact, in test runs the standard method of regenerative simulation was somewhat superior to simple Monte Carlo application of (2.2). However, inspection of (2.2) indicates that the variability of Y might to a large extent be explained by the fluctuations in L. For example, if we let $\alpha = \gamma$ and note that typically γ is small, it is suggested that Y is close to L. This suggests applying the method of control variates (see, for example, [16] and also [8], which has extensive references to queueing applications as well as discussion of nonlinear control variates which become of importance in Section 5). We apply here what are sometimes called regression-adjusted control variates. That is, we define C by replacing $\gamma B(u)$ by zero in (2.2),

$$C = \int_0^L e^{\alpha u} e^{-\gamma u}\, du, \tag{2.3}$$

and perform n replications to produce observations $(Y_1, C_1), \ldots, (Y_n, C_n)$. Let

$$\Sigma = \begin{pmatrix} \sigma_Y^2 & \sigma_{YC} \\ \sigma_{YC} & \sigma_C^2 \end{pmatrix}$$

be the covariance matrix of Y, C, $z = \sigma_{YC}/\sigma_Y/\sigma_C$ the correlation coefficient, S the empirical covariance matrix and $\overline{Y}, \overline{C}$ the empirical means. Then $\overline{Y} - \sigma_{YC}/\sigma_C^2(\overline{C} - \mu_C)$ is asymptotically the minimum-variance linear unbiased estimator of μ based on $\overline{Y}, \overline{C} - \mu_C$, and if we replace the unknown σ_{YC} and σ_C^2 by the empirical estimates, we arrive at an estimator with the same asymptotic properties,

$$(2.4) \quad \hat{\mu}_n = \overline{Y} - s_{YC}/s_C^2(\overline{C} - \mu_C) \approx N(\mu, \nu^2/n) \quad \text{where } \nu^2 = \sigma_Y^2(1 - z^2).$$

Note that, formally, this is equivalent to a regression of Y upon C and that the control variate estimate $\hat{\mu}_n$ in (2.4) is the level of the regression line at μ_C. However, (2.4) is valid also without assumptions like normality of $\mathbb{E}(Y|C)$ being linear in C (but of course the efficiency of the method hinges on the presence of such linear structure).

From the practical programming point of view, it is necessary to rewrite the definitions of Y, C in a more tractable form. Consider for simplicity the case $\alpha = \gamma$, which will turn out to be of paramount importance. Then C just reduces to L. For Y, note that the paths of $\{B(u)\}$ are piecewise linearly decreasing with jumps of sizes, say, $b(0), b(1), \ldots$ at $u(0) = 0 < u(1) < \cdots$, that is,

$$B(u) = b(k) - (u - u(k)), \qquad u(k) \le u < u(k+1).$$

From this it follows easily that

$$(2.5) \qquad Y = \sum_{k:\, u(k) < L} \gamma^{-1}(1 - e^{-\gamma b(k)}) - \gamma^{-1}(1 - e^{-\gamma B(L)}).$$

3. Some empirical examples. A single test evaluation of the estimator $\hat{\mu}_n$ in (2.4) in the M/M/1 case with $n = 100$ and traffic intensity 0.9 (so that $\mu = 9.0$) produced some highly encouraging results. We obtained $\hat{\mu}_n = 9.030$, $\hat{\sigma}_Y^2 = 102.7$ and $\hat{z} = 0.999395$. Thus, an asymptotic 95% confidence interval is

$$\hat{\mu}_n \pm 2\hat{\nu}/n^{1/2} = 9.030 \pm 2\big(102.7(1 - 0.999395^2)/100\big)^{1/2} = 9.030 \pm 0.071.$$

The observations are plotted in Figure 1, which shows a very strong linear dependence between Y and C, corresponding nicely to the rather remarkable high value of z.

Since $\gamma = 1 - \rho = 0.1$, $\mathbb{P}_L$ corresponds to arrival intensity $\rho + \gamma = 1$ and service intensity $1 - \gamma = 0.9$, i.e., a transient M/M/1 queue with traffic intensity 1.111111. The estimate $\hat{z}$ of $z = \sigma_{YC}/\sigma_Y/\sigma_C$ was computed using the estimate s_C of σ_C even though σ_C can be computed explicitly in this case. Our

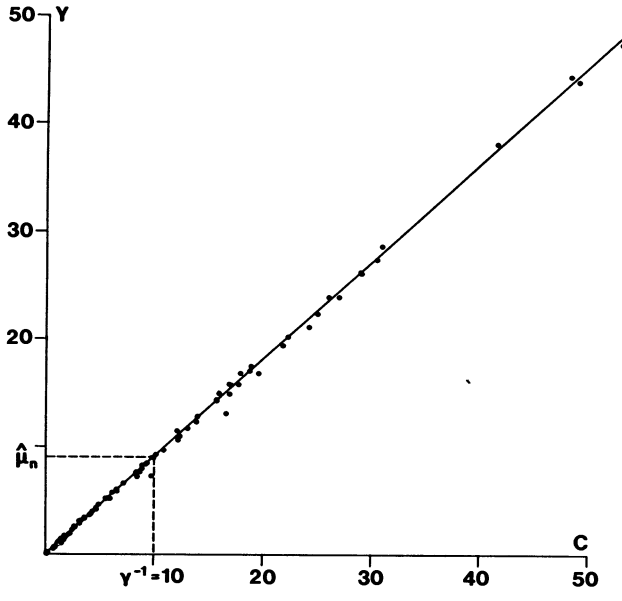

FIG. 1.

main reason for this is avoiding values greater than 1 which would otherwise have occurred in a number of our test runs.

For further illustration and comparison with established methods, a number of further test runs were performed for some standard simple queues with traffic intensity 0.9. In addition to (2.4), also the standard regenerative estimate and the Minh–Sorli estimate ([13]; see Section 6 for more detail) were recorded. In each table, the parameters were chosen so as to make expected execution times (defined as the number of customers taken through the system) roughly equal. All computations were done on a Macintosh™ SE 20 using Lightspeed™ Pascal and a Pascal version of the portable random number generator on page 215 of [5] (no attempt was made to optimize the programs in terms of execution time).

We first once more considered M/M/1 and obtained the results of Tables 1 and 2 (the only difference is that the sample size in Table 2 is 9 times as large).

TABLE 1

Method	Parameters	Confidence interval	Execution time (min)
(2.4)	$\alpha = \gamma/5$, $n = 20$	8.967 ± 0.102	8
(2.4)	$\alpha = \gamma$, $n = 100$	9.030 ± 0.071	9
(2.4)	$\alpha = 5\gamma$, $n = 500$	9.029 ± 0.119	17
Regenerative	1000 cycles	8.723 ± 2.543	8
Minh–Sorli	1000 cycles	9.004 ± 0.087	7

TABLE 2

Method	Parameters	Confidence interval	Execution time (min)
(2.4)	$\alpha = \gamma/5, n = 180$	8.992 ± 0.040	80
(2.4)	$\alpha = \gamma, n = 900$	8.993 ± 0.025	83
(2.4)	$\alpha = 5\gamma, n = 4500$	9.002 ± 0.138	130
Regenerative	9000 cycles	9.491 ± 1.983	81
Minh–Sorli	9000 cycles	8.993 ± 0.033	67

TABLE 3

Method	Parameters	Confidence interval	Execution time (min)
(2.4)	$\alpha = \gamma, n = 100$	4.511 ± 0.007	4
Regenerative	1000 cycles	3.501 ± 1.131	4
Minh–Sorli	1000 cycles	4.593 ± 0.071	4

For M/D/1 with service periods of unit lengths and arrival intensity 0.9, one has $\mu = 4.500$ and $\gamma = 0.207147$. Thus $\mathbb{P}_L$ corresponds to arrival intensity 1.107147 (and unit service periods). The simulation results are in Table 3.

For D/M/1 with interarrival periods of lengths $1/0.9 = 1.111111$ and service intensity 1, one has $\mu = 4.179$ and $\gamma = 0.193100$. Thus, $\mathbb{P}_L$ corresponds to service intensity 0.918011 (and unit interarrival periods). The simulation results are in Tables 4 and 5 (the difference is as before a factor 9 in the sample size).

It is seen from Tables 1–5 that the estimator (2.4) in all cases is much superior to regenerative simulation. For example, for D/M/1, Table 5 indicates a variance reduction by a factor of $0.557^2/0.032^2 = 303$, corresponding to a reduction in computer time by the same factor to obtain a given precision.

TABLE 4

Method	Parameters	Confidence interval	Execution time (min)
(2.4)	$\alpha = \gamma, n = 100$	4.213 ± 0.067	2
Regenerative	1000 cycles	5.280 ± 1.907	4
Minh–Sorli	1000 cycles	4.186 ± 0.009	3

TABLE 5

Method	Parameters	Confidence interval	Execution time (min)
(2.4)	$\alpha = \gamma, n = 900$	4.166 ± 0.032	25
Regenerative	9000 cycles	4.513 ± 0.557	27
Minh–Sorli	9000 cycles	4.181 ± 0.003	21

For M/M/1 in Table 2, the corresponding figure is 6292, and for M/D/1 in Table 3 it is 26105! In some cases the method performs also somewhat better than the Minh–Sorli method, in some cases it is somewhat inferior. For the moment, we take these observations as sufficient motivation for a closer study of the estimator (2.4) and return to a more detailed discussion of Tables 1–5 in Section 6.

4. Performance evaluation. We first generalize the setup somewhat. In the same way as in [18], [19], [1] and [2] (Chapter XII) we think of the given queue imbedded corresponding to $\theta = \theta_0$ in an exponential family $\{\mathbb{P}_\theta\}$ of queues, where $\mathbb{P}_0$ corresponds to traffic intensity 1. That is, if A_θ, B_θ are the interarrival and service time distributions corresponding to $\mathbb{P}_\theta$, then $\mathbb{E}_0 T = \mathbb{E}_0 U$ and

$$B_\theta(dx) = \frac{e^{\theta x}}{\hat{B}_0(\theta)} B_0(dx), \qquad A_\theta(dx) = \frac{e^{-\theta x}}{\hat{A}_0(-\theta)} A_0(dx).$$

Further, $\gamma = \theta_L - \theta_0$, where $\theta_L > 0$ is the solution of

$$\hat{A}_0(-\theta_0)\hat{B}_0(\theta_0) = \hat{A}_0(-\theta_L)\hat{B}_0(\theta_L).$$

Limit theorems as $\theta_0 \uparrow 0$ (or, equivalently, $\gamma \downarrow 0$) then provide approximations for the given queue. As a main example of fundamental importance for the following, note that γW is approximately exponentially distributed with rate 1. In particular, $m \approx \gamma^1 = \mu_C$ and this provides one way of explaining the efficiency of the Monte Carlo method: In some sense we are using the heavy-traffic approximation as control and are simulating only the correction.

We shall consider functionals more general than the mean, which, in view of the heavy traffic limit theorem, it will be convenient to represent as $m_\Phi = \mathbb{E}\Phi(\gamma W)$. Thus, for example, $\mu = \gamma^{-1} m_\Phi$ with $\Phi(x) = x$. We assume without loss of generality that $\Phi(0) = 0$ and let ϕ be the derivative of Φ. Then,

$$\begin{aligned} m_\Phi = \mathbb{E}\Phi(\gamma W) &= \int_0^\infty \phi(u)\mathbb{P}(\gamma W > u)\, du \\ &= \int_0^\infty \phi(u) e^{-u} \mathbb{E}_L e^{-\gamma B(u/\gamma)}\, du. \end{aligned} \tag{4.1}$$

We also allow L to have a general distribution by writing $L = V/\gamma$, where $g(t) = \mathbb{P}(V > t)$ is arbitrary [except that $g(t) > 0$ for all $t < \infty$]. The definitions (2.2) and (2.3) are then in an obvious manner generalised to

$$Y = \int_0^V \phi(u) g(u)^{-1} e^{-u} e^{-\gamma B(u/\gamma)}\, du, \qquad C = \int_0^V \phi(u) g(u)^{-1} e^{-u}\, du. \tag{4.2}$$

We have $\mu_Y = m_\Phi$ and, provided the relevant second moments exist,

$$\hat{m}_\Phi = \overline{Y} - s_{YC}/s_C^2(\overline{C} - \mu_C) \approx N\big(m_\Phi, \nu^2/n\big), \tag{4.3}$$

where $\nu^2 = \sigma_Y^2(1 - z^2)$ with $z = \sigma_{YC}/\sigma_Y/\sigma_C$.

We are now ready to state some of our main theoretical results. They involve constants β, τ^2 defined in the proof of Proposition 2 (which can be expressed in terms of the $\mathbb{P}_0$-ascending ladder height distribution but the expressions are of less importance in the present context; an example is in Section 6). The following regularity conditions are assumed throughout:

A1. There exists $\alpha > 0$ such that $\phi(u)g(u)^{-1} \le e^{\alpha u}$.

A2. The distribution of $X = U - T$ is strongly nonlattice.

This seems a little restrictive: Since W has an exponential tail, moments corresponding to $\phi(u)$ increasing faster than exponentially are infinite, and if $g(u)$ decreases slower than exponentially, then it is easily seen that the variances of Y and C are infinite. A2 is needed for uniform exponential convergence rates in the key renewal theorem (see [18] for details) and holds, for example, if either U or T is spread out ([2], Chapter VI.2).

The first result gives the heavy-traffic behavior of the asymptotic variance in (4.3):

THEOREM 1. *As* $\theta_0 \uparrow 0$, *it holds that*

$$\nu^2 = \sigma_Y^2(1 - z^2) \approx \gamma^3\tau^2\int_0^\infty \phi(u)^2 g(u)^{-1} e^{-2u}\, du.$$

The implication that the variance ν^2/n decreases at rate γ^3 for a fixed number n of replications is somewhat deceiving since it is also relevant to note that on the contrary, the expected computer time i needed to create one replication increases. As in [1] we define i as the expected number of customers needed for one replication and take $\pi^2 = i\nu^2$ rather than ν^2 as the main performance measure. Up to a constant, the interpretation is as the variance on the estimator per unit computer time, and the following result shows a decrease at rate γ.

THEOREM 2. *For some constant* c_1, $i \approx c_1\gamma^{-2}\mu_V$, *and hence*

$$\pi^2 = i\nu^2 \approx \gamma c_1 \kappa(g)\tau^2 \text{ where } \kappa(g) = \int_0^\infty g(t)\, dt \int_0^\infty \phi(u)^2 g(u)^{-1} e^{-2u}\, du.$$

In contrast, it is shown in [4] that the similar performance measure for the regenerative method increases at rate γ^{-2} so that indeed the empirical comparisons of the two methods in Section 3 are confirmed by theory. The present estimator also performs better than the one obtained by averaging $\tilde{Y} = \phi(U)e^{-\gamma B(U/\gamma)}$, where U is exponential with unit rate and $\{B(t)\}$ is simulated from $\mathbb{P}_L$ [this estimator is suggested from (4.1) by analogies to the area of Monte Carlo evaluation of infinite integrals]. Indeed,

$$\tilde{\nu}^2 = \operatorname{Var} \tilde{Y} \approx \operatorname{Var} \phi(U)(1 - \gamma B(U/\gamma)) \approx \gamma^2 \operatorname{Var} \phi(U) \operatorname{Var} B(\infty),$$

and thus $\tilde{\pi}^2 = i\tilde{\nu}^2 \approx c_2$.

In the proofs of Theorems 1 and 2 we concentrate on the steps needed to check the form of the constants and expressions and omit the verification of regularity conditions like uniform integrability since arguments of this type follow established lines of the references of this section. A fundamental observation is the following:

PROPOSITION 1. *As $\theta_0 \uparrow 0$, it holds that the conditional distribution of Y given C or, equivalently, V is asymptotically normal with mean $(1 - \gamma\beta)C$ and variance*

$$\gamma^3\tau^2\int_0^V \phi(u)^2 g(u)^{-2} e^{-2u}\, du.$$

PROOF. Let $b_i(\theta) = \lim_{t\to\infty} \mathbb{E}_\theta B(t)^i$, $\beta = b_1(0)$,

$$J_i = \int_0^V \phi(u) g(u)^{-1} e^{-u}\big(B(u/\gamma)^i - b_i(\theta_L)\big)\, du,$$

and let τ^2 be the limiting $\mathbb{P}_0$-variance of $t^{-1/2}\int_0^t B(u)\, du$ which exist according to the CLT for cumulative processes (cf. [2], Chapter V.3). Letting ξ denote standard Brownian motion, it follows from a suitable version of the functional CLT (e.g., [7] and references therein) that

$$\begin{aligned}\left\{\int_0^T (B(u/\gamma) - b_1(\theta_L))\, du\right\}_{T\ge 0} &= \left\{\gamma\int_0^{T/\gamma} (B(u) - b_1(\theta_L))\, du\right\}_{T\ge 0}\\ &\approx \{\gamma\tau\xi(T/\gamma)\}_{T\ge 0}\\ &= \{\tau\gamma^{1/2}\xi(T)\}_{T\ge 0}.\end{aligned}$$

Hence,

$$\begin{aligned}Y &\approx \int_0^V \phi(u) g(u)^{-1} e^{-u}(1 - \gamma B(u/\gamma))\, du + O(\gamma^2)\\ &\approx \int_0^V \phi(u) g(u)^{-1} e^{-u}(1 - \gamma b_1(\theta_L))\, du - \gamma J_1\\ &\approx (1 - \gamma\beta)C - \gamma\int_0^V \phi(u) g(u)^{-1} e^{-u}\tau\gamma^{1/2}\, d\xi(u),\end{aligned}$$

which has the asserted conditional distribution. □

PROOF OF THEOREM 1. By an obvious extension of the proof of Proposition 1, it follows that

$$Y = p(\gamma)C + \sum_{i=1}^{3} (-1)^i \gamma^i / i! J_i + o(\gamma^3) \tag{4.4}$$

where

$$p(\gamma) = 1 + \sum_{i=1}^{3} (-1)^i \gamma^i / i! b_i(\theta_L).$$

Now A2 implies that $|\mathbb{E}B(u/\gamma)^i - b_i(\theta)| \le ce^{-\varepsilon u}$ for some ε, c ([18], Lemma 5), and thus

$$
\begin{aligned}
|\mathbb{E}(C-\mu_C)J_i| &= \left|\mathbb{E}(C-\mu_C)\int_0^V \phi(u)g(u)^{-1}e^{-u}\big(\mathbb{E}B(u/\gamma)^i - b_i(\theta_L)\big)\,du\right| \\
&= \left|\mathbb{E}(C-\mu_C)\int_V^\infty \phi(u)g(u)^{-1}e^{-u}\big(\mathbb{E}B(u/\gamma)^i - b_i(\theta_L)\big)\,du\right| \\
&\le \mathbb{E}|C-\mu_C|\int_V^\infty \phi(u)g(u)^{-1}e^{-u}ce^{-\varepsilon u}\,du \\
&\le (1+\varepsilon/\gamma-\alpha)^{-1}\mathbb{E}\big(|C-\mu_C|ce^{-\alpha-\varepsilon V/\gamma}\big) \\
&= O(\gamma)o(1)
\end{aligned} \tag{4.5}
$$

(using A1 in the fourth step). Hence, up to $o(\gamma^3)$ terms,

$$
\begin{aligned}
\sigma_Y^2 &\approx p(\gamma)^2\sigma_C^2 + \gamma^3 \operatorname{Var} J_1 - 2\gamma p(\gamma)\mathbb{E}(C-\mu_C)J_1, \\
\sigma_{YC} &\approx p(\gamma)\sigma_C^2 - \gamma\mathbb{E}(C-\mu_C)J_1, \\
\sigma_{YC}^2 &\approx \sigma_C^2\big[p(\gamma)^2\sigma_C^2 - 2\gamma p(\gamma)\mathbb{E}(C-\mu_C)J_1\big] \approx \sigma_C^2\big[\sigma_Y^2 - \gamma^3\operatorname{Var}J_1\big],
\end{aligned} \tag{4.6}
$$

$$
z^2 \approx 1 - \gamma^3 \operatorname{Var} J_1/\sigma_Y^2 \approx 1 - \gamma^3\tau^2\int_0^\infty \phi(u)^2 g(u)^{-1}e^{-2u}\,du/\sigma_Y^2. \tag{4.7}
$$

From this the desired estimate for ν^2 follows. □

PROOF OF THEOREM 2. The number of customers needed to generate Y, C is $\tau(V/\gamma)$. It is a standard fact (an easy consequence of Wald's identity) that, for each v, $\mathbb{E}_L\tau(v/\gamma) \approx c_1\gamma^{-2}v$. Hence

$$
i = \mathbb{E}_L\tau(V/\gamma) \approx \mathbb{E}c_1\gamma^{-2}V = c_1\gamma^{-2}\mu_V,
$$

and from this the estimate for π^2 follows by combining with Theorem 1. □

An obvious question is to look for an optimal V, i.e., for the form $g(t)$ of the tail probability which minimises $\kappa(g)$ in Theorem 2. To this end, we first note the following.

PROPOSITION 2. *For any strictly positive function g on $(0,\infty)$ it holds that $\kappa(g) \ge c_3^2$, where $c_3 = \int_0^\infty h(u)\,du$, with $h(u) = \phi(u)e^{-u}$. Equality holds if and only if $g(u)/h(u)$ is constant a.e. on the set $\{h > 0\}$.*

PROOF. Assume first that $c_3 < \infty$, let H be a random variable with density $h(u)/c_3$ and define $K = g(H)/h(H)$. Then, by Jensen's inequality,

$$
1 \le \mathbb{E}K\mathbb{E}K^{-1} = \int_0^\infty g(t)/c_3\,dt\int_0^\infty \phi(u)^2 g(u)^{-1}e^{-2u}/c_3\,du = \kappa(g)/c_3^2,
$$

with equality if and only if K is a.s. constant, which shows the claim in this

case. The case $c_3 = \infty$ is treated by truncating ϕ and using monotone convergence. □

EXAMPLE 1. Consider simulation of the mean so that $\phi(u) = 1$. It then follows from Proposition 2 that the choice $g(u) = h(u) = e^{-u}$, i.e., V exponential with rate 1, is asymptotically optimal in heavy traffic.

EXAMPLE 2. Consider simulation of an exponential moment $\mathbb{E}e^{\beta W}$ which (e.g., [2], page 269) exists if and only if $\beta < \gamma$. Then $\mathbb{E}e^{\beta W} = \alpha m_\Phi + 1$, where $\Phi(x) = (e^{\alpha x} - 1)/\alpha$, with $\alpha = \beta/\gamma$. Corresponding to this Φ we have $h(u) = \phi(u)e^{-u} = e^{(\alpha-1)u}$, and it is suggested to simulate by taking V to be exponential with rate $1 - \alpha$.

EXAMPLE 3. Consider simulation of a higher order moment $\mathbb{E}W^k$, $k \geq 2$, or equivalently, $\mathbb{E}W^k/k$, which corresponds to $h(u) = \phi(u)e^{-u} = u^{k-1}e^{-u}$. Taking g proportional to h is not feasible since h is not nonincreasing, and the problem of minimizing $\kappa(g)$ in the class of nonincreasing functions appears much more complicated. Nevertheless, Proposition 2 may guide the choice of a suitable (though not optimal) g. In fact, if we look for a g with the same tail behavior as h, an obvious candidate is the Erlang(k) distribution where

$$g(u) = e^{-u}(1 + u + u^2/2 + \cdots + u^{k-1}/(k-1)!).$$

Consider $k = 2, 3, 4$. Then the global minimum c_3^2 of $\kappa(g)$ is 1, 4 and 36, respectively. If, motivated by Example 1, we try $g(u) = e^{-u}$, we get $\kappa(g) = 2$, 24 and 720, whereas the Erlang(k) distribution leads to 1.2, 5.7 and 58.8 (these figures were computed by numerical integration). Obviously, this is a substantial improvement of the exponential case and of the same order of magnitude as c_3^2.

Note that in these examples computationally convenient forms like (2.5) can be found for Y and C and that the computer generation of random variables distributed as V is straightforward. In general, this may of course present an added complication in the choice of g.

5. Weighted regression and ratio control. For simplicity, we shall confine the discussion of the rest of the paper to the case of the mean μ and let Y, C be defined as in Section 2 with $\alpha = \gamma$ so that $C = L$ (that is, we do not normalise by γ as was done for mathematical convenience in Section 4). Some results of Section 4 then need some slight translation. In particular, note that $\nu^2 \approx \gamma\tau^2$ and that $\mathrm{Var}(Y|C) \approx \gamma^2\tau^2 C$.

When applying the estimation method discussed so far under heavy traffic conditions, the number n of replications could in many cases be rather small. This is so not only because each replication is time-consuming, but also because a high precision is obtained already for quite small values of n. Assume, for example, that in the M/M/1 example with $\rho = 0.95$ we want an estimate with a relative precision of 1%. For $\rho = 0.9$, Table 1 indicates that

this is achieved for $n = 100$, and Theorem 1 therefore suggests that $n = 100/2^3 = 13$ should be sufficient for $\rho = 0.95$. However, the considerations leading to confidence intervals are based upon large-sample properties, for example, the CLT for Y, C, and that the variance on the control coefficient does not contribute substantially to the variance on the estimator. One might therefore ask whether in such cases it would not be more appropriate to base the choice of method on the fact that $1 - \rho$ is small rather than on n being large.

To this end, note that, according to Proposition 1, it holds asymptotically in heavy traffic that

$$Y_i = aC_i + C_i^{1/2}\varepsilon_i, \tag{5.1}$$

where the ε_i are i.i.d. $N(0, \lambda^2)$ for some λ^2 (in fact, $a = 1 - \gamma\beta$ and $\lambda^2 = \gamma^2\tau^2$). Thus, $\mu = \mu_Y \approx \mu_C a$, and one might consider estimating a by maximum likelihood estimation in the weighted regression (5.1) (with the C_i treated as constants), which leads to $\hat{a} = \overline{Y}/\overline{C}$, and use $\hat{\mu}_n^R = \mu_C \overline{Y}/\overline{C}$ as a heavy-traffic estimator of μ. This procedure is of course also motivated by empirical findings like the beautiful linear dependence in Figure 1 (it is of some interest to note that $\mu \approx \mu_C a = \gamma^{-1} - \beta$ is Siegmund's [18] corrected heavy-traffic approximation, in the form incorporating the first-order correction term, see further the discussion in Section 6).

For confidence intervals, we may treat the C_i either as random variables, see to this end Proposition 3, or as constants. In the constant case we have $\hat{a} \approx N(a, \lambda^2/(n\overline{C}))$, where the standard estimator for λ^2 is $(n-1)\hat{\lambda}^2 = \Sigma(Y_i - aC_i)^2/C_i$, and we are lead to the asymptotic 95% confidence interval

$$\hat{\mu}_n^R \pm 2\hat{\lambda}\mu_C/(n\overline{C})^{1/2}. \tag{5.2}$$

To investigate whether these ideas present any improvement, we simulated $\hat{\mu}_n^R$ and (5.2) 100 times in the preceding M/M/1 situation with $\rho = 0.95$ and $n = 13$, evaluating also $\hat{\mu}_n$ and the control variate confidence interval. In 57 cases $\hat{\mu}_n^R$ were closer to the true value 19 than $\hat{\mu}_n$ (however, the difference was in most cases minor) and the coverage figures for the two confidence intervals were 88 and 79, respectively.

Treating the C_i as random variables, one may consider $\hat{\mu}_n^R$ as an instance of ratio estimation (e.g., [15], [9]), but the estimator is probably more naturally viewed within the framework of nonlinear control variates (see, e.g., [11] and [8] and references therein. The following result gives the large-sample properties and shows that these, for all practical purposes, seem equivalent to those of $\hat{\mu}_n$:

PROPOSITION 3. (a) *As $n \to \infty$ with θ_0 fixed, it holds that $\hat{\mu}_n^R = \mu_C \overline{Y}/\overline{C}$ is asymptotically normal with mean $\mu_Y = \mu$ and variance ω^2/n, where*

$$\omega^2 = \sigma_Y^2 + \mu_Y^2 \mu_C^{-2}\sigma_C^2 - 2\mu_Y \mu_C^{-1}\sigma_{YC}.$$

(b) *Let ν^2 be defined as in Theorem* 1. *Then $\nu^2 \le \omega^2$.*

(c) *As* $n \to \infty$ *with* θ_0 *fixed, it holds that* $\mu_n - \mu_n^R \approx \delta(\overline{C} - \mu_C) \approx N(0, \delta^2\sigma_C^2/n)$, *where* $\delta = \mu_Y\mu_C^{-1} - \sigma_{YC}/\sigma_C^2$.

(d) *The constant* δ *in* (b) *satisfies* $\delta = O(\gamma^2)$, $\theta_0 \uparrow 0$. *In particular,* $\nu^2/\omega^2 \approx 1 + O(\gamma)$.

(e) *If* $n \to \infty$ *and* $\theta_0 \uparrow 0$ *at the same time, then* $\hat{\lambda}\mu_C/\overline{C}^{1/2} \approx \gamma\tau^2$, $\hat{\nu}^2 \approx \gamma\tau^2$.

PROOF. Letting $f(y, c) = \mu_C y/c$ and noting that

$$f(\mu_Y, \mu_C) = \mu, \qquad f_y(\mu_Y, \mu_C) = 1, \qquad f_c(\mu_Y, \mu_C) = -\mu_Y\mu_C^{-1},$$

it follows by Taylor expansion that

$$\begin{aligned} \hat{\mu}_n^R &= f(\overline{Y}, \overline{C}) \\ &\approx f(\mu_Y, \mu_C) + f_y(\mu_Y, \mu_C)(\overline{Y} - \mu_Y) + f_C(\mu_Y, \mu_C)(\overline{C} - \mu_C) \\ &\approx \overline{Y} - \mu_Y\mu_C^{-1}(\overline{C} - \mu_C), \end{aligned} \tag{5.3}$$

from which (a) follows and also (c) by noting that $s_{YC}/s_C^2 \approx \sigma_{YC}/\sigma_C^2$. For (b), just note that by (a) and Theorem 1, $\omega^2 - \nu^2 = \sigma_C^2\delta^2$ (alternatively, as noted in [8] in a more general setting, (b) is a consequence of (5.3), the optimality of the control coefficient σ_{YC}/σ_C^2 and $s_{YC}/s_C^2 \approx \sigma_{YC}/\sigma_C^2$). For (d), we proceed by small variants of the proofs of Theorems 1 and 2. As in (4.5), $\mathbb{E}J_i = O(\gamma)$ and thus, by (4.4), $\mu_Y \approx p(\gamma)\mu_C + O(\gamma^2)$, whereas (4.6) yields $\sigma_{YC} \approx p(\gamma)\sigma_C^2 + O(\gamma^2)$, from which $\delta = O(\gamma^2)$ follows. The estimate for ν^2/ω^2 is then obtained by noting that $\nu^2 = O(\gamma^3)$, $\omega^2 - \nu^2 = O(\delta^2) = O(\gamma^4)$. Finally, (e) is immediate from earlier estimates. □

The content of (d) is that the loss of efficiency given by (b) for the ratio control method is asymptotically negligible, whereas (e) indicates that in situations like $\rho = 0.95$ and $n = 500$, (5.2) should be close to the control variate confidence interval. This was nicely confirmed by a test run for M/M/1 (where $\mu = 19$ when $\rho = 0.95$), which gave 18.997 ± 0.026, respectively, 18.996 ± 0.029, for the two methods.

We would tend to conclude that from a practical point of view the choice between the linear control variate method considered so far and ratio control [possibly with the confidence interval (5.2)] seldom matters much, but that there may be some special situations where one would feel that ratio control is based on more solid theoretical ground.

6. Concluding discussion. We first reinspect the empirical material of Section 3.

For the M/M/1 case, it is of interest to compute the constant τ^2 of Theorem 1 in the optimal case $\alpha = \gamma$. To this end, notice that $\mathbb{P}_0$ corresponds to both arrival rate and service rate equal to $(1 + \rho)/2$ which in heavy traffic is conveniently replaced by 1. It is then easily seen from the standard variance formula for cumulative processes that $\tau^2 = 2$. An alternative check of this is provided by an explicit calculation of the constants σ_Y^2, σ_C^2 and σ_{YC} of

Section 2, which gives

$$\sigma_C^2 = (1-\rho)^{-2},$$

$$\sigma_{YC} = \rho\sigma_Y^2 = \rho(1-\rho)^{-2},$$

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

$$z^2 = \left(1 + \frac{2(1-\rho)^3}{1+(1-\rho)-(1-\rho)^2-(1-\rho)^3}\right)^{-1}.$$

(The derivation of the expression for σ_Y^2 is tedious but elementary and is omitted.) Taylor expansion yields $1 - z^2 \approx 2(1-\rho)^3$ and comparison with (4.6) reconfirms that $\tau^2 = 2$. In particular, for $\rho = 0.9$, the exact value of $1 - z$ is 0.00091, the heavy-traffic approximation [as given by replacing σ_Y^2 by σ_C^2 in (4.7) and taking the square root] is 0.00100 and the empirical value reported for the first test run is $1 - 0.999395 \approx 0.00060$.

The implication of Example 1 that the choice $\alpha = \gamma$ is close to being optimal is also confirmed by Tables 1 and 2. Since the sample size in Table 2 is 9 times as large as in Table 1, we expect the confidence intervals to be 3 times as narrow. This also roughly holds with the exception for the case (2.4) with $\alpha = 5\gamma$. The explanation for this is simply that here all variances are infinite. Nevertheless, the estimator is still consistent and seems to perform reasonably well.

Siegmund's [18] corrected diffusion approximations are of course highly relevant in the present context for a number of reasons. For example, the underlying mathematical methods are the ones providing the foundation of the performance evaluation in Section 4. As noted earlier, they provide an intuitive motivation for the regression scheme (5.1) and, finally, they of course provide an alternative to simulation which is potentially attractive since the precision is high, not least when the $O(\gamma)$ term, say η, is included. As pointed out by an editor, it should be noted that the version $\mu \approx \gamma^{-1} - \beta$ including only the first-order correction is always easy to use for GI/G/1 queues since β can be evaluated by a simple numerical integration. For η this scheme has, however, not yet been implemented.

We next turn to the comparisons with the Minh–Sorli method, which is based on the relation ([2], page 186) $\mu = \mu_W = (\mu_{X^2} + \mu_X\mu_{I^2}/\mu_I)/2\mu_{-X}$, where I is the idle period and $X = U - T$. Here, only μ_{I^2}/μ_I needs to be simulated. This is an instance of ratio estimation, and for the M/M/1 case where I is exponentially distributed with rate ρ, it is easily seen as in (5.3) that the large-sample variance on the ratio estimate of μ_{I^2}/μ_I is ω_1^2/n, where

$$\omega_1^2 = \mu_I^{-2}\sigma_{I^2}^2 + \mu_{I^2}^2\mu_I^{-4}\sigma_I^2 - 2\mu_{I^2}\mu_I^{-3}\sigma_{I^2I} = 8\rho^{-2}.$$

Normalising by the correct value yields the sample variance $\mu_W^{-2}\omega_1^2/4n$ on the Minh–Sorli estimator. Since the mean number of customers needed to produce

one replication is just the mean number of customers $(1-\rho)^{-1}$ served in a busy cycle, it follows that the performance measure similar to π^2 is $(1-\rho)^{-1}\mu_W^{-2}\omega_1^2/4 \approx 2(1-\rho)$ (in the general case, we get similarly a behavior like $c_2\gamma$ for some constant c_2). For the method of the present paper, we get $\pi^2 \approx (1-\rho)\tau^2 = 2(1-\rho)$. *That is, in the M/M/1 case the present method and the Minh–Sorli method have equivalent heavy-traffic properties.* Nevertheless, for moderate values of ρ, it follows both from Tables 1 and 2 and from theory that the present method is mildly better. Thus, in Table 2 we observed a width of the confidence interval of 0.025, whereas the preceding explicit formulas lead to expecting 0.028. For the Minh–Sorli method we observed 0.033, whereas the exact formulae lead to expecting 0.037. The heavy-traffic approximation is 0.030 in both cases.

For M/D/1 in Table 3, the present estimator performs even better than the Minh–Sorli estimator, whereas for D/M/1 in Table 4 (confirmed by the more precise Table 5) the opposite is the case. We have not gone into theoretical calculations to confirm these observations, but consider the empirical findings to be intuitively reasonable. More precisely, much of the variance on $\hat{\mu}_n$ is due to the variability of $B(u)$, which, in turn, we roughly expect to increase as the variance of the service time distribution increases. Similarly, the variance on the Minh–Sorli estimate comes from the variability of I, which, in turn, we roughly expect to increase as the variance of the interarrival distribution increases.

More than from the fact that the present method is slightly superior to the Minh–Sorli method for some simple queues, we would, however advocate its advantages by the greater flexibility. As a first example, it has already been seen that the method is not restricted to integral moments. We have also performed some test runs for simulating the transient behavior, more specifically the mean waiting time of the Nth customer in the M/M/1 case with $\rho = 0.9$ (in the definition of Y one then has to replace the upper limit of the integral by $C \wedge M_N$, where M_N is the Nth partial maximum of the random walk). The variance reduction compared to crude simulation was substantial; a factor of 3.2, 5.2 and 12.5 for $N = 50$, 100 and 200, respectively. The reason that this, nevertheless, is less dramatic than for the steady-state case is simply that Y need not be close to C if M_N is substantially smaller than C (correspondingly, the correlation was only 0.60, 0.70 and 0.84). It would seem appealing to take the fluctuations in M_N into account by using S_N as a further control as in [6] but we have not carried this out.

Finally, but least, we should like to point out that also a number of models more complex than GI/G/1 queues can be treated. Some remarks on random walks with spatial inhomogeneity are given at the end of [1], and we shall outline here how to treat random walks with Markov-dependent increments (Markov additive processes), a class of models which is becoming increasingly important in applied probability (a brief introduction is in [2], Chapter X.4-5 and extensive bibliographies are in [3] and [14]; in the statistical literature, the closest reference we know of is [10]). Here the random walk $S_n = X_0 + \cdots + X_{n-1}$ is governed by a Markov chain $\{J_n\}_{n=0,1,\ldots}$ with state space

E in the sense that the X_k are independent given $\{J_n\}$ with X_k having distribution depending only on J_k, J_{k+1}, say, $\mathbb{P}(X_k \le x,\ J_{k+1} = j | J_k = i) = F^{(ij)}(x)$. We assume for simplicity that E is finite (though this is not crucial), write $\mathbb{P}_i = \mathbb{P}(\cdot | J_0 = i)$, $p(i,j) = \mathbb{P}_i(J_1 = j)$ and denote by $\hat{F}(\alpha)$ the matrix with ijth entry $\hat{F}^{(ij)}(\alpha)$. The Lundberg equation then becomes $\mathrm{spr}(\hat{F}(\gamma)) = 1$, and if $h = (h(i))_{i \in E}$ denotes the positive right eigenvector of $\hat{F}(\gamma)$ corresponding to the eigenvalue 1, the $\mathbb{P}_L$-distribution of the Markov additive process is given by

$$F^{(ij)}(dx) = \frac{h(j)}{h(i)} e^{\gamma x} F^{(ij)}(dx).$$

Letting $M = \max_n S_{n \ge 0}$, $\tau(u)$, $B(u)$, etc., be defined as before and $\mathbb{P}_{L;i} = \mathbb{P}_L(\cdot | J_0 = i)$, Wald's likelihood ratio identity becomes

$$\mathbb{P}_i(M > u) = h(i) e^{-\gamma u} \mathbb{E}_{L;i} \left(\frac{e^{-\gamma B(u)}}{J(\tau(u))} \right)$$

(see, e.g., [12] or [20]), and it is thus suggested to estimate $\mathbb{E}_i M$ by simulating

$$Y = h(i) \int_0^C \frac{e^{-\gamma B(u)}}{J(\tau(u))}\, du$$

(with C exponential with rate γ) from $\mathbb{P}_{L;i}$ and using C as control variate.

Acknowledgments. I would like to thank Allan Deis, Claus Holst and Henrik Stryhn for their patience in implementing an earlier unsuccessful attempt at resolving the problem of the present paper. I am also indebted to J. Deshpande, P. W. Glynn, J. B. Robertson, the Editor and the referee for a number of useful comments.

REFERENCES

[1] ASMUSSEN, S. (1985). Conjugate processes and the simulation of ruin problems. *Stochastic Process. Appl.* **20** 213–229.
[2] ASMUSSEN, S. (1987). *Applied Probability and Queues*. Wiley, New York.
[3] ASMUSSEN, S. (1989). Risk theory in a Markovian environment. *Scand. Actuar. J.* 69–100.
[4] ASMUSSEN, S. (1991). Queueing simulation in heavy traffic. *Math. Oper. Res.* To appear.
[5] BRATLEY, P., FOX, B. L. and SCHRAGE, L. E. (1987). *A Guide to Simulation*, 2nd ed. Springer, New York.
[6] GAVER, D. P. and THOMPSON, G. L. (1973). *Programming and Probability Models in Operations Research*. Brooks/Cole, Monterey, Calif.
[7] GLYNN, P. W. and WHITT, W. (1987). Sufficient conditions for functional-limit-theorem versions of $L = \lambda W$. *Queueing Systems Theory Appl.* **1** 279–287.
[8] GLYNN, P. W. and WHITT, W. (1989). Indirect estimation via $L = \lambda W$. *Oper. Res.* **37** 82–103.
[9] GRAY, H. L. and SCHUCHANY, W. R. (1972). *The Generalised Jackknife Statistics*. Dekker, New York.
[10] HÖGLUND, T. (1974). Central limit theorems and statistical inference for Markov chains. *Z. Wahrsche. Verw. Gebiete* **29** 123–151.
[11] KLEIJNEN, J. P. C. (1974). *Statistical Techniques in Simulation, Part 1*. Dekker, New York.
[12] MILLER, H. D. (1961). A generalization of Wald's identity with applications to random walks. *Ann. Math. Statist.* **32** 549–560.

[13] MINH, D. L. and SORLI, R. M. (1983). Simulating the GI/G/1 queue in heavy traffic. *Oper. Res.* **31** 966–971.
[14] NEUTS, M. F. (1989). *Structured Stochastic Matrices of the M/G/1 Type and their Applications*. Dekker, New York.
[15] RAO, J. N. K. (1969). Ratio and regression estimators. In *New Developments in Survey Sampling* (N. L. Johnson and H. Smith, Jr., eds.) 213–234. Interscience, New York.
[16] RUBINSTEIN, R. Y. (1981). *Simulation and the Monte Carlo Method*. Wiley, New York.
[17] SIEGMUND, D. (1976). Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.* **4** 673–684.
[18] SIEGMUND, D. (1979). Corrected diffusion approximations in certain random walk problems. *Adv. Appl. Probab.* **11** 701–719.
[19] SIEGMUND, D. (1985). *Sequential Analysis*. Springer, New York.
[20] TWEEDIE, M. C. K. (1960). Generalizations of Wald's fundamental identity of sequential analysis to Markov chains. *Proc. Cambridge Philos. Soc.* **56** 205–214.
[21] WALRAND, J. (1987). Quick simulation of rare events in queueing networks. *Proc. 2nd International Workshop on Applied Mathematics and Performance/Reliability Models of Computer/Communication Systems* (G. Iazeolla, P. J. Courtois and O. J. Boxma, eds.) 275–286. North-Holland, Amsterdam.
[22] WHITT, W. (1989). Planning queueing simulations. *Management Sci.* **35** 1341–1366.

DEPARTMENT OF MATHEMATICS
CHALMERS UNIVERSITY OF TECHNOLOGY
S-41296 GOTHENBURG
SWEDEN