Rare Event Simulation for a Slotted Time $M/G/s$ Model

Blanchet, J., Glynn, P., and Lam, H.

October 23, 2009

Abstract

This paper develops a rare-event simulation algorithm for a discrete time version of the $M/G/s$ loss system and a related Markov modulated variant of the same loss model. The algorithm is shown to be efficient in the many server asymptotic regime in which the number of servers and the arrival rate increase to infinity in fixed proportion. A key idea is to study the system as a measure-valued Markov chain and to steer the system to the rare event through a randomization of the time horizon over which the rare event is induced.

1 Introduction

A central contribution of A. K. Erlang was his development and analysis of multi-server loss models. In this paper, we offer a computational perspective on the loss models studied by Erlang, with a special focus on the calculation of loss probabilities suffered by the $M/G/s$ queue when a customer arrives to find all the servers busy. We have a particular interest in the many server setting in which the numbers of servers s is large, as well as the arrival rate to the system.

An extensive literature exists on rare-event simulation for computing tail probabilities for single-server and multi-server systems with infinite capacity buffers, or for calculating loss probabilities for such single and multi-server systems in the large buffer limit (some early references include Siegmund (1976), Chang et al. (1994) and Heidelberger (1995)). In such problem settings, one can often draw upon the extensive theory on large deviations of random walk to develop an appropriate change-of-measure for use in the associated importance sampler. On the other hand, in the many server loss context of the current paper, the rare event is induced in a fundamentally different way. For example, in the single-server setting the rare event is generated by an associated random walk undergoing unusual dynamics over a time horizon that is of the same order as the buffer size (see, Asmussen (1985)). On the other hand, as we shall see in this paper, the most efficient way of inducing loss in the many server context is over a finite horizon (independent of s) over which the arrival rate and mix of incoming traffic change so as to eventually saturate all the servers. Specifically, the mix of service requirements changes in such a way that longer service times are favored relative to the nominal service time distribution.

There is a small existing literature on related many-server simulation problems. For the $M/M/s$ queue, Cottrell et al. (1983) use the birth-death structure to show that the optimal importance distribution involves switching the birth and death rates to the right of the mode of the equilibrium distribution. Glynn (1995) studies large deviations for the infinite server queue when subjected to a high arrival rate, and Szechtman and Glynn (2002) exploit this large deviations result to develop an efficient importance sampler for computing equilibrium tail probabilities for this class of models. Both of the latter papers exploit the fact that the steady-state distribution of such

infinite server systems has an explicit probabilistic representation (that yields a perfect simulation algorithm in the presence of bounded service times). Unfortunately, the finite server loss system under consideration here enjoys no such probabilistic representation, and the algorithms developed here do not explicitly leverage off these infinite server ideas. Finally, Srikant and Whitt (1996, 1999) study variance reduction techniques for many-server loss systems based on control variate ideas.

In contrast to the earlier efforts described above, the many-server algorithm introduced here does not exploit either memorylessness in the service time distribution (as in the $M/M/s$ queue) or any special probabilistic representation characteristics (as in the infinite-server system). Of course, this extension comes at the cost of needing to enrich the state-space so as to include information on the remaining service time requirements of the customers that are present. To deal with this in full generality would require working with a continuous-time measure-valued process. To avoid technical complications, we choose to work with a slotted time formulation in which the service times have finite support, thereby leading to a finite-dimensional (vector-valued) state descriptor. (It can also be argued that such a formulation is not particularly limiting from a practical standpoint.) The slotted time $M/G/s$ queue on which we focus in Sections 2 through 4 is a well known example of an "insensitive queue" (see, for example, Kelly (1979)), so that there is a closed form expression for the aggregate loss probability in such queues. Because our algorithm does not exploit insensitivity, it can be extended to rare-event computations to which known insensitivity results do not apply (e.g. the distribution of the remaining service requirements conditional on buffer overflow) and to model extensions not covered by insensitivity. As a simple illustration of this latter point, we extended our $M/G/s$ ideas in Sections 5 and 6 to a model that exhibits a correlated arrival stream, namely that of a Markov-modulated variant of the $M/G/s$ queue. While not described in this paper, the same ideas easily extend to slotted $M/G/s$ loss models in which the arrival process is Poisson with a periodic (non-stationary) arrival intensity. Thus, we take the viewpoint that the ideas developed here are an important step towards the development of efficient computational tools for more complex loss systems.

This paper is organized as follows. Section 2 introduces the basic model that is considered in this paper. The formulation is a natural discrete-time analog of the continuous time $M/G/s$ loss system. We work in discrete time in large part to avoid complications that arise in dealing with continuous time measure-valued processes, both at a mathematical level and in creating the data structures necessary to efficiently implement the continuous-time importance sampler. Section 3 describes our rare-event algorithm, while Section 4 proves that the algorithm is efficient. Section 5 describes the extension of the algorithm to a Markov-modulated version of the $M/G/s$ loss system, and Section 6 discusses efficiency and implementation issues. Some extensions involving time inhomogeneous processes and non-Poisson input are given in Section 7. Finally, Section 8 provides a discussion of numerical results.

2 The Basic Model

Our basic model is a loss system with s servers. As will soon become evident, we will extensively exploit the Markovian "measured-valued" description of the system in describing both our algorithm and developing the associated theory. This measure-valued process becomes finite-dimensional (i.e. vector-valued) when the service time requirements are bounded and integer-valued, with slotted time arrivals. In view of this simplification, we will therefore work with a discrete-time formulation of the classical $M/G/s$ loss model. Throughout the rest of the paper we will use the notation $\mathrm{Po}(\lambda)$ to denote a generic Poisson random variable with mean $\lambda > 0$. Similarly we use $\mathrm{Bin}(n, p)$ to denote

a generic binomial r.v. (random variable) with mean np and variance $np(1-p)$ with $n \in \{1,2,...\}$ and $p \in (0,1)$.

Suppose that V is a r.v. corresponding to a generic service time requirement. We assume that V takes values in the integers $\{1,2,...,m\}$. Let $W_n(i)$ be the number of customers in the system at time n with a remaining service time requirement of i time units (so that $W_n \triangleq (W_n(1),...,W_n(m))$ is the measure-valued process mentioned above). Then,

$$Q_n \triangleq \|W_n\| = \sum_{i=1}^{m} W_n(i)$$

is the total number of customers in the system at time n. The state W_{n+1} at time $n+1$ is obtained from W_n according to the following algorithm.

Algorithm A

1. Advance time to just prior to $n+1$, and temporarily set $W_{n+1}(i) = W_n(i+1)$ for $1 \leq i \leq m-1$.

2. Generate a Poisson r.v. χ_{n+1}, independently of W_n, having mean λ, corresponding to the total number of customers arriving at time $n+1$.

3. Each of the arriving χ_{n+1} customers is independently assigned a service time requirement from the distribution of V.

4. If $Q_n - W_n(m) + \chi_{n+1} \leq s$, all the arriving customers are accepted into the system, and each customer is assigned its own server. If $Q_n - W_n(m) + \chi_{n+1} > s$, then $Q_n - W_n(m) + \chi_{n+1} - s$ customers are chosen uniformly and at random (without replacement) from the χ_{n+1} customers that have just arrived and immediately deleted from the system (i.e. "lost"). The remaining $s - Q_n + W_n(m)$ customers are assigned servers.

5. The state vector W_{n+1} is now updated to include the service time requirements of the $\min(\chi_{n+1}, s - Q_n + W_n(m))$ customers that have just been accepted into the system.

6. The $[Q_n - W_n(m) + \chi_{n+1} - s]^+$ customers that have just been accepted into the system start their service times at time $n+1$.

Our interest is focused on computing the equilibrium fraction of customers that are lost in the above system, in the "many server asymptotic regime" in which $s \nearrow \infty$ and $\lambda \nearrow \infty$ so that $\lambda EV/s \triangleq \rho \in (0,1)$. We shall apply an approach for computing such equilibrium quantities that was introduced by Goyal et al. (1992). In particular, note that $W = (W_n : n \geq 0)$ is a finite state irreducible discrete time Markov chain for each $s \geq 1$. Let W_∞ be a r.v. with the stationary distribution $\pi(\cdot)$ of $(W_n : n \geq 0)$. Select a non-empty subset $A \subseteq S \triangleq \{(w_1,...,w_m) \in \mathbb{Z}_+^m : w_1 + ... + w_m \leq s\}$ and let $\pi_A(\cdot) = P(W_\infty \in \cdot \,|\, W_\infty \in A)$. The fraction of loss customers (in steady state) satisfies

$$\beta(s) = \lim_{n \longrightarrow \infty} \frac{\sum_{j=0}^{n-1} [Q_j - W_j(m) + \chi_{j+1} - s]^+}{\sum_{j=0}^{n-1} \chi_j}.$$

The previous limit is well defined almost surely because of the law of large numbers. Moreover, a generalization of Kac's formula (see, for example, p. 123 of Breiman (1968)) yields

$$\beta(s) = \frac{E_{\pi_A}\left(\sum_{j=0}^{T_A-1}[Q_j - W_j(m) + \chi_{j+1} - s]^+\right)}{\lambda}\pi(A), \tag{1}$$

where $E_{\pi_A}(\cdot)$ is the expectation operator under which $P(W_0 \in \cdot) = \pi_A(\cdot)$ and $T_A \triangleq \inf\{n \geq 1 : W_n \in A\}$ is the first return time to A. If we choose A so that $\pi(A)$ is bounded away from zero (as a function of s), it is evident that the computationally challenging term in calculating the steady-state loss probability $\beta(s)$ is the quantity

$$\kappa(s) \triangleq E_{\pi_A}\left(\sum_{j=0}^{T_A-1}[Q_j - W_j(m) + \chi_{j+1} - s]^+\right). \tag{2}$$

Note that $[Q_j - W_j(m) + \chi_{j+1} - s]^+$ is non-zero for $j < T_A$ only when $T \triangleq \inf\{n \geq 1 : Q_{n-1} - W_{n-1}(m) + \chi_n > s\} < T_A$. Since the event $\{T < T_A\}$ will typically be a rare event in our asymptotic regime, computing (2) involves rare-event simulation ideas. As we shall see in Section 3, the set A can be chosen so that the quantity $\pi(A)$ can be efficiently calculated via standard Monte Carlo ideas.

One strategy for computing $\kappa(s)$ is to run a crude Monte Carlo simulation of W up to time n, with n large. We then estimate (2) via

$$\frac{\sum_{j=0}^{n-1}\Gamma_{j+1}(W_j)\, I(W_j \in A)}{\sum_{j=0}^{n-1} I(W_j \in A)},$$

where the collection of r.v.'s $(\Gamma_j(w) : j \geq 1, w \in A)$ is independent of $(W_n : n \geq 0)$ and chosen so that

$$E\Gamma_j(w) = E_w\left(\sum_{j=0}^{T_A-1}[Q_j - W_j(m) + \chi_{j+1} - s]^+\right)$$

for $w \in A$. Let $Y_j = (Y_j(1), ..., Y_j(m))$, where $Y_j(i)$ is the number of customers having service requirement i that arrive at time j. Because of the Poisson "thinning property", the $Y_j(i)$'s are independent and Poisson with $EY_j(i) = \lambda P(V = i)$. In constructing suitable $\Gamma_j(w)$'s, we take advantage of the fact that

$$\begin{aligned}
&E_w\left(\sum_{j=0}^{T_A-1}[Q_j - W_j(m) + \chi_{j+1} - s]^+\right) \\
&= E_w\left(\sum_{j=0}^{T_A-1}[Q_j - W_j(m) + \chi_{j+1} - s]^+ I(T < T_A)\right) \\
&= E_w\left(\phi(W_{T-1}, Y_T)\, I(T < T_A)\right),
\end{aligned}$$

where $\phi(w, y)$ is defined for $\|w + y\| > s + w_m$ via

$$\phi(w, y) = E_w\left(\sum_{j=0}^{T_A-1}[Q_j - W_j(m) + \chi_{j+1} - s]^+ \,\middle|\, W_0 = w, Y_1 = y\right).$$

Since $T = 1$ conditional on $\|w + y\| > s + w_m$, $\phi(\cdot)$ can be estimated via simulating W under its nominal dynamics. As a result, the rare event computation can be reduced to an efficient sampler for expectations defined in terms of (W_{T-1}, Y_T), conditional on $T < T_A$. The next section describes such a rare-event simulation algorithm (Algorithm C below).

3 Rare-event Simulation for the Discrete-time $M/G/s$ Queue

For $0 \leq n < T$, the system experiences no loss, so the dynamics of the many-server queue are identical to those of the infinite server queue for such n. In particular,

$$W_{n+1} = BW_n + Y_{n+1}$$

for $n \geq 0$ and $n < T$, where B is the $m \times m$ matrix in which $B(i, i+1) = 1$ if $1 \leq i \leq m-1$ and $B(i, j) = 0$ otherwise. If we set $R_0 = w = W_0$ and let $(R_n : n \geq 0)$ be the Markov chain driven by the recursion

$$R_{n+1} = BR_n + Y_{n+1}$$

for $n \geq 1$, then $W_n = R_n$ for $n < T$; the sequence $R = (R_n : n \geq 0)$ is the measure-valued description of the infinite-server system associated with $(W_n : n \geq 0)$. Note that

$$R_n = B^n R_0 + \sum_{j=0}^{n-1} B^j Y_{n-j}$$

for $n \geq 0$. Since $B^l = 0$ for $l \geq m$, it follows that

$$R_n = \sum_{j=0}^{m-1} B^j Y_{n-j}$$

for $n \geq m$ and R_n is independent of R_{n-m}. In particular, this implies that $(R_n : n \geq 0)$ converges to stationarity in m steps.

As pointed out in Section 2, the key to a successful rare-event simulation algorithm for computing loss probabilities for our model is the development of an efficient sampler for (W_{T-1}, Y_T), conditional on $T < T_A$. But, given $R_0 = w = W_0$,

$$T = \inf\{n \geq 1 : \|R_n\| > s\}.$$

It follows that the required sampler is equivalent to finding an efficient algorithm for computing expectations defined in terms of $(R_{T-1}, R_T - BR_{T-1})$, conditional on $T < \widetilde{T}_A$, where

$$\widetilde{T}_A = \inf\{n \geq 1 : R_n \in A\}.$$

Let τ be a random variable independent of $(R_n : n \geq 0)$ and taking values on the positive integers. Moreover, assume that $P(\tau = n) > 0$ for $n \geq 1$. Let

$$\widetilde{P}(\cdot) = P_w(\,\cdot \,|\, \|R_\tau\| > s). \tag{3}$$

Note that

$$\begin{aligned}
&\widetilde{P}\left((R_1,...,R_T)\in\cdot,T<\widetilde{T}_A\right)\\
&=\sum_{n=1}^{\infty}\frac{P_w\left((R_1,...,R_T)\in\cdot,T<\widetilde{T}_A,\|R_n\|>s,\tau=n\right)}{P_w\left(\|R_\tau\|>s\right)}\\
&=\sum_{n=1}^{\infty}\frac{P_w\left((R_1,...,R_T)\in\cdot,T<\widetilde{T}_A,T\leq n,\|R_n\|>s\right)P\left(\tau=n\right)}{P_w\left(\|R_\tau\|>s\right)}\\
&=E_w\left(I\left((R_1,...,R_T)\in\cdot,T<\widetilde{T}_A\right)\frac{\sum_{n=T}^{\infty}h_n\left(T,R_T\right)}{P_w\left(\|R_\tau\|>s\right)}\right) \qquad (4)
\end{aligned}$$

where

$$h_n\left(k,r\right)=P\left(\tau=n\right)P_r\left(\|R_{n-k}\|>s\right).$$

Because R converges to stationarity in m steps,

$$\sum_{n=T}^{\infty}h_n\left(T,R_T\right)=\sum_{n=T}^{T+m-1}h_n\left(T,R_T\right)+\sum_{n\geq T+m}P\left(\tau=n\right)P\left(\|R_m\|>s\right)$$

and

$$P_w\left(\|R_\tau\|>s\right)=\sum_{j=1}^{m-1}P_w\left(\|R_j\|>s\right)P\left(\tau=j\right)+P\left(\tau\geq m\right)P_w\left(\|R_\infty\|>s\right).$$

For each j, $\|R_j\|$ is a Poisson r.v., so that $P_w\left(\|R_j\|>s\right)$ and $P_w\left(\|R_\infty\|>s\right)$ can be calculated in closed form. This, of course, simplifies the computation of

$$L^{-1}\left(T,R_T\right)=\frac{\sum_{n=T}^{\infty}h_n\left(T,R_T\right)}{P_w\left(\|R_\tau\|>s\right)}.$$

Since $\sum_{n=k}^{\infty}h_n\left(k,r\right)>0$ for each $r\in\mathbb{Z}_+^m$ and $k\in\mathbb{Z}_+$, it follows that (4) implies the change-of-measure identity

$$\begin{aligned}
&P_w\left((R_1,...,R_T)\in\cdot,T<\widetilde{T}_A\right)\\
&=\widetilde{E}\left(I\left((R_1,...,R_T)\in\cdot,T<\widetilde{T}_A\right)L\left(T,R_T\right)\right). \qquad (5)
\end{aligned}$$

We now turn to the question of generating paths of R under $\widetilde{P}$. Note that

$$\begin{aligned}
\widetilde{P}\left(\cdot\right)&=\sum_{k=1}^{\infty}\frac{P_w\left(\cdot\,,\|R_k\|>s\right)P\left(\tau=k\right)}{P_w\left(\|R_\tau\|>s\right)}\\
&=\sum_{k=1}^{\infty}\frac{P_w\left(\cdot\,|\,\|R_k\|>s\right)P_w\left(\|R_k\|>s\right)P\left(\tau=k\right)}{P_w\left(\|R_\tau\|>s\right)}\\
&=\sum_{k=1}^{\infty}P_w\left(\cdot\,|\,\|R_k\|>s\right)P_w\left(K=k\right), \qquad (6)
\end{aligned}$$

where K is a r.v., independent of $(Y_n : n \geq 1)$ and with probability mass function

$$P_w(K=k) = \frac{P_w(\|R_k\| > s)\, P(\tau = k)}{P_w(\|R_\tau\| > s)}. \tag{7}$$

We are now ready to state our algorithm for generating the r.v. $\Gamma_j(w)$.

Algorithm B

1. Generate the r.v. K
2. If $K \geq m$, use Algorithm C to generate $(Y_{K-m+1}, ..., Y_K)$, conditional on $\|R_K\|_1 > s$. Then, independently generate $(Y_1, ..., Y_{K-m})$ from the nominal distribution for the Y_j's. If $K < m$, use Algorithm C to generate $(Y_1, ..., Y_K)$, conditional on $\|R_K\|_1 > s$ and $R_0 = w$.
3. Compute T from $R_0, ..., R_K$.
4. If $\widetilde{T}_A \leq T$, then output 0 and STOP. Otherwise, put $W_j = R_j$ for $j < T$ and generate W_T from W_{T-1} and Y_T (by independently and uniformly removing exactly $[Q_{T-1} - W_{T-1}(m) + \chi_T - s]^+$ customers from Y_T).
5. Using W_T as the initial condition, generate $W_{T+1}, ..., W_{T_A}$ using the nominal dynamics of the W_j's.
6. Output
$$\sum_{j=T}^{T_A - 1} [\|BW_{j-1} + Y_j\| - s]^+ L(T, R_T).$$

Algorithm C is intended to first generate $\|R_l\|$ with $l \leq m$, conditional on $\|R_l\| > s$ and $R_0 = w$ followed by generating $(Y_1, ..., Y_l)$ conditional on $\|R_l\|$ and $R_0 = w$. Because $\left\|R_l - B^l R_0\right\|$ is a Poisson random variable with mean $\lambda E(\min(V, l))$, the generation of $\|R_l\|$ conditional on $\|R_l\| > s$ and $R_0 = w$, is equivalent to generating r.v.'s from the conditional distribution

$$P(||B^l w|| + \mathrm{Po}(sv_l) \in \cdot \mid \mathrm{Po}(sv_l) > s - ||B^l w||), \tag{8}$$

where

$$v_l = \rho(E\min(V, l))/EV.$$

To generate r.v.'s from the conditional distribution (8), our approach is to use acceptance-rejection based on utilizing a Poisson r.v. with mean $q \triangleq s - \left\|B^l w\right\| + 1$ as the dominating mass function. The acceptance ratios that arise in this setting are of the form $(sv_l/q)^j$, and the average acceptance probability at Step 4 below is

$$\exp(-(q - sv_l)) \left(\frac{q}{sv_l}\right)^q P(\mathrm{Po}(sv_l) \geq q).$$

Since

$$P(\mathrm{Po}(sv_l) \geq q) \sim \frac{1}{(2\pi q)^{1/2}} \exp(-(q - sv_l)) \left(\frac{sv_l}{q}\right)^q$$

as $s \nearrow \infty$, it follows that the expected number of times that Steps 1 through 4 below are executed is of order $s^{1/2}$ as $s \nearrow \infty$.

Algorithm C

1. Generate a Poisson r.v. Z with mean q.
2. If $Z < q$, return to STEP 1. (This occurs with approximately probability $1/2$ when s is large).
3. Generate an independent uniform r.v. U on $[0,1]$.
4. If $U \leq (sv_l/q)^{Z-q}$, go to STEP 5. Else, return to STEP 1.
5. Distribute the Z customers across the r.v.'s $\left(B^j Y_{l-j}\right)(i)$ for $i \in \{1,...,m\}$ and $j \in \{0,1,...,l-1\}$ as a multinomial r.v. with corresponding number of trials equal to Z and associated multinomial probability $E\left(B^j Y_1\right)(i)/\lambda E\left(\min\left(V,l\right)\right)$ for the r.v. $\left(B^j Y_{l-j}\right)(i)$. (Observe that $\left(B^j Y_{l-j}\right)(i)$ is the i-th component of the vector $B^j Y_{l-j}$, which in turn is $Y_{l-j}(i-j)$ if $j < i$ and 0 otherwise.)

To fully specify the algorithm for computing the equilibrium loss probability for our $M/G/s$ queue, we need to describe the distribution of τ and the choice of the subset A. Note that a heavy-tailed choice for τ is inherently conservative from an algorithmic viewpoint, since it assigns relatively greater mass to paths taking a long time to reach the overflow level s. This, in turn, puts less stress on the sampler in attempting to induce the rare event. This can also be seen by studying the likelihood ratio $L(T,R_T)$. Recall that the likelihood ratio's denominator contains the tail sum $\sum_{n=T}^{\infty} h_n(T,R_T)$, where $h_n(k,r) = P(\tau = n) P_r(\|R_{n-k}\|_1 > s)$, thereby suggesting that $L(T,R_T)$ will be easier to control proof-wise when $P(\tau = n)$ decays slowly. As a consequence, our choice for the distribution of τ is $P(\tau \geq n) = n^{-\gamma}$ for $\gamma > 0$. Furthermore, note that Algorithm B requires "conditioning in" the history of the Y_n's up to time K. In order to guarantee that the number of Y_n's generated by the algorithm is well-behaved (e.g. finite mean), we choose $\gamma > 2$ (so that $Var(\tau) < \infty$).

Regarding the choice of A, we have observed previously that A should be chosen so that $\pi(A)$ is bounded away from zero (as a function of s). Since the dynamics of W and R are typically identical, the stationary distribution of W is close to that of R. Furthermore, in the many-server asymptotic regime it follows easily that R_m obeys a central limit theorem (see, for example, Glynn and Whitt (1991)). This suggests the choice

$$A = \left\{ w \in \mathbb{R}^m : \left(1 - as^{-1/2}\right) \sum_{j=0}^{m-1} E\left(B^j Y_1\right) \leq w \leq \sum_{j=0}^{m-1} E\left(B^j Y_1\right) \left(1 + as^{-1/2}\right) \right\},$$

for $a > 0$.

In the next section we discuss the efficiency of the above algorithm. However, before we move on, it is worth to point out that the sampling strategy allows to estimate conditional expectations involving the measure valued descriptor at the time of a loss. For instance, one compute the distribution of the number of customers in the system with given remaining processing time at the time of a loss. The efficiency analysis given in the next section applies with minor changes to the estimation of such distribution.

4 Algorithmic Efficiency

Given an algorithm for computing an expectation α (depending on a parameter $s \nearrow \infty$), any unbiased estimator $\Lambda \geq 0$ for α must satisfy the inequality

$$E\Lambda^2 \geq \alpha^2.$$

Assuming that $\alpha \in (0,1)$ the previous inequality is equivalent to $\log E\Lambda^2 / \log(\alpha) \leq 2$. In our case, s large corresponds to the expectation $\alpha = \alpha(s)$ being small; it is natural to assert that Λ is a *logarithmically efficient* estimator if

$$\underline{\lim}_{s \longrightarrow \infty} \frac{\log E\Lambda^2}{\log \alpha} = 2;$$

see, for example, Asmussen and Glynn (2007) or Juneja and Shahabuddin (2006) for a discussion of this efficiency criterion in the rare-event simulation setting. We will seek to verify a similar criterion for our estimator.

Recall that we have chosen the set A so that $P(R_\infty \in A)$ is bounded away from zero. Our first result confirms our intuition that the stationary distributions for $(R_n : n \geq 0)$ and $(W_n : n \geq 0)$ are close to one another when s is large, so that $\pi(A) = P(W_\infty \in A)$ is also bounded away from zero.

Proposition 1. *For each $s \geq 1$,*

$$\sup_C |P(R_\infty \in C) - P(W_\infty \in C)| \leq mP(R_\infty > s).$$

Proof. Set $R_0 = W_0$ and note that for each subset C,

$$\frac{1}{n}\sum_{j=0}^{n-1} I(R_j \in C) - \frac{1}{n}\sum_{j=0}^{n-1} I(W_j \in C) \leq \frac{1}{n}\sum_{j=0}^{n-1} I(R_j \neq W_j). \tag{9}$$

Since W_j is formed by dropping customers that would otherwise join the infinite-server process R_j, and since accepted customers initiate their service times at the same instants in both the finite-server and infinite-server systems, it follows that $W_j \leq R_j$ for $j \geq 0$ when $W_0 = R_0$. If $R_{j-i} \leq s$ (and $W_{j-i} \leq s$) for $i \leq m-1$, then there have been no customers lost in the last m time units, so that all the customers present in the infinite-server system at time j were also accepted into the finite-server system, yielding the conclusion that $W_j = R_j$. Thus, if $W_j \neq R_j$, there must exist $i \in \{0, 1, ..., m-1\}$ for which $\|R_{j-i}\| > s$. Consequently,

$$\begin{aligned}\frac{1}{n}\sum_{j=0}^{n-1} I(R_j \neq W_j) &\leq \frac{1}{n}\sum_{j=0}^{n-1}\sum_{i=0}^{(m-1)\wedge j} I(\|R_{j-i}\| > s) \\ &\leq \frac{m}{n}\sum_{i=0}^{n-1} I(\|R_i\| > s). \end{aligned} \tag{10}$$

Since $(R_n : n \geq 0)$ and $(W_n : n \geq 0)$ are irreducible positive recurrent discrete-time Markov chains, the strong law of large numbers for such chains implies, on the basis of (9) and (10), that

$$P(R_\infty \in C) - P(W_\infty \in C) \leq mP(\|R_\infty\| > s).$$

Similarly, we obtain that

$$P(W_\infty \in C) - P(R_\infty \in C) \leq mP(\|R_\infty\| > s),$$

yielding the result. □

Recall that $\|R_\infty\| \stackrel{D}{=} \mathrm{Po}(\rho s)$. Hence, in the many server asymptotic regime

$$\frac{1}{s}\log P\left(\|R_\infty\| > s\right) \longrightarrow -\left(\rho - 1 - \log\rho\right) \tag{11}$$

as $s \nearrow \infty$, so that $P\left(R_\infty > s\right) \longrightarrow 0$ exponentially rapidly in s. Consequently, Proposition 1 implies that $\pi\left(A\right) = P\left(W_\infty \in A\right)$ is bounded away from zero as $s \nearrow \infty$.

We turn next to the analysis of $\beta\left(s\right)$.

Theorem 1. *The following limits hold as* $s \nearrow \infty$

i.) $\left(\log P\left(\|R_\infty\| > s\right)\right)^{-1}\log\beta\left(s\right) \longrightarrow 1$
ii.) $\left(\log P\left(\|R_\infty\| > s\right)\right)^{-1}\log\beta\left(s\right) \longrightarrow 1$
iii.) $\left(\log E\left(\left[\|R_\infty\| - s\right]^+\right)\right)^{-1}\log\beta\left(s\right) \longrightarrow 1$
iv.) $\left(\log P\left(Q_\infty = s\right)\right)^{-1}\log\beta\left(s\right) \longrightarrow 1$
v.) $s^{-1}\log\beta\left(s\right) \longrightarrow -[\rho - 1 - \log\left(\rho\right)]$

Proof. Since $\|R_\infty\| \stackrel{D}{=} \mathrm{Po}(\rho s)$ with $\rho \in (0,1)$,

$$\begin{aligned} P\left(\|R_\infty\| = s\right) \leq P\left(\|R_\infty\| \geq s\right) &= \sum_{k=0}^{\infty}\exp\left(-\rho s\right)\frac{\left(\rho s\right)^{s+k}}{(s+k)!} \\ &= \exp\left(-\rho s\right)\frac{\left(\rho s\right)^{s}}{s!}\left(1 + \sum_{k=1}^{\infty}\frac{\rho^k s^k}{(s+1)\cdot\ldots\cdot(s+k)}\right) \\ &\leq P\left(\|R_\infty\| = s\right)\left(1 + \sum_{k=1}^{\infty}\rho^k\right). \end{aligned}$$

So, $\log P\left(\|R_\infty\| = s\right)/\log P\left(\|R_\infty\| \geq s\right) \longrightarrow 1$ as $s \nearrow \infty$. Also,

$$\begin{aligned} P\left(\|R_\infty\| = s+1\right) &\leq E\left(\left[\|R_\infty\| - s\right]^+\right) \\ &= \sum_{k=1}^{\infty} k\exp\left(-\rho s\right)\frac{\left(\rho s\right)^{s+k}}{(s+k)!} \\ &\leq P\left(\|R_\infty\| = s\right)\sum_{k=1}^{\infty} k\rho^k, \end{aligned}$$

and hence $\log P\left(\|R_\infty\| = s\right)/\log\left(E[\|R_\infty\| - s]^+\right) \longrightarrow 1$ as $s \nearrow \infty$. Kac's formula implies that

$$\begin{aligned} P\left(Q_\infty = s\right) &= E_{\pi_A}\left(\sum_{j=0}^{T_A - 1} I\left(Q_j = s\right)\right)/E_{\pi_A}\left(T_A\right) \\ &= E_{\pi_A}\left(\sum_{j=0}^{T_A - 1} I\left(Q_j = s, \widetilde{T} < T_A\right)\right)\pi\left(A\right) \\ &\geq P_{\pi_A}\left(\widetilde{T} < T_A\right)\pi\left(A\right), \end{aligned} \tag{12}$$

where $\widetilde{T} = \inf\{n \geq 1 : Q_n \geq s\} = \inf\{n \geq 1 : \|R_n\| \geq s\}$. Also, because the W_j's are dominated by the R_j's,

$$P\left(Q_\infty = s\right) \leq P\left(\|R_\infty\| \geq s\right). \tag{13}$$

Starting from (1), a similar argument to that for (12) shows that

$$\beta\left(s\right) \geq P_{\pi_A}\left(T < T_A\right) \frac{\pi\left(A\right)}{\lambda}. \tag{14}$$

In addition,

$$\begin{aligned}
&\frac{1}{n}\sum_{j=0}^{n-1}\left[Q_j - W_j\left(m\right) + \chi_{j+1} - s\right]^+ \\
&= \frac{1}{n}\sum_{j=0}^{n-1}\left[\sum_{l=1}^{m-1} W_j\left(l\right) - W_j\left(m\right) + \chi_{j+1} - s\right]^+ \\
&\leq \frac{1}{n}\sum_{j=0}^{n-1}\left[\sum_{l=1}^{m-1} R_j\left(l\right) - W_j\left(m\right) + \chi_{j+1} - s\right]^+ \\
&= \frac{1}{n}\sum_{j=0}^{n-1}\left[\|BR_j + Y_{j+1}\| - s\right]^+ \\
&= \frac{1}{n}\sum_{j=0}^{n-1}\left[\|R_{j+1}\| - s\right]^+.
\end{aligned}$$

Sending $n \nearrow \infty$, we conclude that

$$\beta\left(s\right) \leq E\left(\left[\|R_\infty\| - s\right]^+\right)/\lambda. \tag{15}$$

Combining (11) to (15) and the earlier results deduced in this proof, we see that the argument is complete once we show that $s^{-1}\log P_{\pi_A}\left(T < T_A\right)$ and $s^{-1}\log P_{\pi_A}\left(\widetilde{T} < T_A\right)$ converge to $-(\rho - 1 - \log\rho)$ as $s \nearrow \infty$. Without loss of generality we assume that $P\left(V = m\right) > 0$ (for otherwise we can re-define the endpoint m of the support of V). In fact, given our development above, we just require a lower bound. Note that

$$\begin{aligned}
&P_{\pi_A}\left(\widetilde{T} < T_A\right) \\
&\geq \int_A \pi_A\left(dw\right) P_w\left(\widetilde{T} = m < T_A\right) \\
&\geq \int_A \pi_A\left(dw\right) P\left(\|R_\infty\| = s\right) \\
&\times P_w\left(\|R_i\| < s, R_i \notin A, 1 \leq i \leq m-1 \middle| \|R_m\| = s\right).
\end{aligned}$$

But, conditional on $\|R_m\| = s$, for $1 \leq i \leq m-1$, $\left\|R_i - B^i w\right\|$ is distributed $\mathrm{Bin}(s, E\min\left(V,i\right)/EV)$. Because $w \in A$ is within $O\left(s^{1/2}\right)$ of $\sum_{j=0}^{m-1} B^j EY_1$, and $\sum_{j=0}^{m-1} B^j E[Y_1(i)] = \lambda E[V-i]^+/EV$ it follows that for each $\varepsilon > 0$

$$\sup_{w\in A} P\left(\left|s^{-1}\|R_i\| - \rho\frac{E[V-i]^+}{EV} - \frac{E\min\left(V,i\right)}{EV}\right| > \varepsilon \middle| \|R_m\| = s\right) \longrightarrow 0$$

as $s \nearrow \infty$. Since $\rho EV < \rho E[V-i]^+ + E(\min(V,i)) < EV$ for $1 \leq i \leq m-1$, it follows that $\inf_{w \in A} P_w(\|R_i\| < s, R_i \notin A | \|R_m\| = s) \longrightarrow 1$ as $s \nearrow \infty$. But $s^{-1} \log P(\|R_\infty\| = s) \longrightarrow -(\rho - 1 - \log \rho)$ as $s \nearrow \infty$, proving that

$$s^{-1} \log P_{\pi_A}\left(\widetilde{T} < T_A\right) \longrightarrow -(\rho - 1 - \log(\rho))$$

as $s \nearrow \infty$. An identical argument works for $s^{-1} \log P_{\pi_A}(T < T_A)$, finishing the proof. □

In view of the discussion at the beginning of this section and Theorem 1, we will establish the effectiveness of our algorithm by proving that for each $w \in A$,

$$\underline{\lim}_{s \longrightarrow \infty} \frac{\widetilde{E}\left(\Gamma_1^2(w) L(T, R_T)^2\right)}{\beta(s)} = 2 \tag{16}$$

Theorem 2. *For each $w \in A$, $\Gamma_1(w)$ satisfies (16).*

Proof. Observe that

$$\begin{aligned}
&\widetilde{E}\left(\Gamma_1^2(w) L(T, R_T)^2\right) \\
&= \widetilde{E}\left(\left(\sum_{j=0}^{T_A-1} [Q_j - W_j(m) + \chi_{j+1} - s]^+ L(T, R_T)\right)^2\right) \\
&= E_w\left(\left(\sum_{j=0}^{T_A-1} [Q_j - W_j(m) + \chi_{j+1} - s]\right)^2 I(T < T_A) L(T, R_T)\right).
\end{aligned}$$

But for $\|r\| > s$,

$$\frac{L(k,r)}{P_w(\|R_\tau\| > s)} \leq \frac{1}{h_k(k,r)} = \frac{1}{P(\tau = k)} \leq \frac{2^\gamma k^\gamma}{\gamma},$$

and thus

$$\begin{aligned}
\widetilde{E}\left(\Gamma_1^2(w) L(T, R_T)^2\right) \leq \frac{2^\gamma}{\gamma} &P_w(\|R_\tau\| > s) \\
&\times E_w\left(\left(\sum_{j=1}^{T_A} \chi_j\right)^2 T_A^\gamma I(T < T_A)\right).
\end{aligned} \tag{17}$$

Clearly,

$$\begin{aligned}
&P(\|R_m\| > s) P(\tau = m) \\
&\leq P_w(\|R_\tau\| > s) \\
&\leq \sum_{j=1}^{m-1} P_w(\|R_j\| > s) + P(\tau \geq m) P(\|R_\infty\| > s).
\end{aligned}$$

Since $\|R_j\|$ is distributed Poisson with mean $\rho s E(\min(V,i))/EV + \left\|B^i w\right\|$ with $w \in A$,

$$P_w(\|R_j\| > s) = o(P(\|R_\infty\| > s)).$$

We then conclude that

$$\frac{\log P_w\left(\|R_\tau\| > s\right)}{\log P\left(\|R_\infty\| > s\right)} \longrightarrow 1 \tag{18}$$

as $s \nearrow \infty$. Finally, Holder's inequality and Wald's identity imply that

$$\begin{aligned}
&E_w\left(\left(\sum_{j=1}^{T_A}\chi_j\right)^2 T_A^\gamma I\left(T < T_A\right)\right) \\
&\leq E_w\left(T_A^{2+\gamma}\left(\max_{1\leq j\leq T_A}\chi_j^2\right) I\left(T < T_A\right)\right) \\
&\leq E_w^{1/p} T_A^{(2+\gamma)p} \times E_w^{1/q}\sum_{j=1}^{T_A}\chi_j^{2q} \times P_w^{1/r}\left(T < T_A\right) \\
&\leq E_w^{1/p} T_A^{(2+\gamma)p} \times E_w^{1/q} T_A \times E^{1/q}\chi_1^{2q} \times P_w^{1/r}\left(T < T_A\right)
\end{aligned} \tag{19}$$

for $1/p + 1/q + 1/r = 1$ and $p, q, r \geq 1$. Set $r = 1 + \varepsilon$. If we can show that $E_w T_A^n$ is uniformly bounded in s for each $n \geq 1$, then it follows from (19) that

$$\begin{aligned}
&\overline{\lim}_{s\longrightarrow\infty}\frac{1}{s}\log E_w\left(\left(\sum_{j=1}^{T_A}\chi_j\right)^2 T_A^\gamma I\left(T < T_A\right)\right) \\
&\leq -\frac{1}{1+\varepsilon}\left(\rho - 1 - \log\left(\rho\right)\right);
\end{aligned} \tag{20}$$

since $\varepsilon > 0$ was arbitrary, (18) and (20) then prove the theorem. To deal with $E_w T_A^n$ for arbitrary $n \geq 1$, note that the independence of R_m, $R_{m+1}, ...$ implies that

$$\begin{aligned}
&P_w\left(T_A > km\right) \\
&\leq P_w\left(R_m \notin A, R_{2m} \notin A, ..., R_{km} \notin A\right) \\
&= P\left(R_\infty \notin A\right)^k = \left(1 - P\left(R_\infty \in A\right)\right)^k,
\end{aligned}$$

proving that $E_w T_A^n$ is uniformly bounded as $s \nearrow \infty$ for each $n \geq 1$ (by the choice of A and the central limit theorem for R_∞). As a result, we have proved that our estimator is efficient. □

5 The Markov Modulated $M/G/s$ Queue

Now we consider a variation of the basic model in which the arrival process is autocorrelated. The process $W' = (W'_n : n \geq 0)$ denotes the measure valued description of the loss system. Similarly, to emphasize the distinction between the model treated here and the basic model developed in earlier sections we use the super-index " $'$ " for quantities that correspond to their natural counterparts. So, for instance, the number of customers that arrive at time n is denoted by χ'_n.

We suppose that there exists an irreducible finite state $\mathcal{S}$-valued Markov chain $X = (X_n : n \geq 0)$ that modulates the sequence of arriving customers $(\chi'_n : n \geq 0)$. Specifically, conditional on X, we assume that the χ'_n's are independent Poisson r.v.'s for which $E\left(\chi'_n \middle| X\right) = \lambda\left(X_n\right)$ for some function $\lambda : \mathcal{S} \longrightarrow \mathbb{R}_+$. To simplify the arguments that follow, we shall assume that the transition matrix $(p\left(x,y\right) : x, y \in \mathcal{S})$ satisfy $p\left(x,x\right) > 0$ for $x \in \mathcal{S}$. Let μ be the stationary distribution of X. The rest of the model is identical to that described in Algorithm A. The reader should note that the service

times are not assumed to depend on X. We introduce this assumption to simplify the exposition but we shall describe at the end of Section 6 how to deal with the case in which the service times depend on X.

It is immediate from the description in the previous paragraph that for each $s \geq 1$, the process $((W'_n, X_n) : n \geq 0)$ is an irreducible finite-state Markov chain on $S \times \mathcal{S}$, having a stationary distribution denoted by π'.

We will consider a many-server asymptotic regime in which $\lambda(x)$ and s go to infinity in such a way that for each $x \in \mathcal{S}$, $\rho(x) \triangleq \lambda(x) EV/s$ is fixed. A key difference in this setting is the role that the Markov modulation plays in the large deviations behavior of the system. We shall assume that

$$\rho' \triangleq \max_{x \in \mathcal{S}} \rho(x) < 1. \tag{21}$$

In order to motivate this assumption, observe that if there exists $z \in \mathcal{S}$ such that $\rho(z) > 1$, then the associated loss probability is bounded away from zero and therefore it could be estimated via crude Monte Carlo using standard techniques. Indeed, to see that (21) is required to induce a rare-event environment note that since $p(x, x) > 0$ for $x \in \mathcal{S}$, then the number of busy servers will typically reach s if X exhibits a path segment $(X_{j+1}, ..., X_{j+m})$ for which $\rho(X_{j+1}) = ... = \rho(X_{j+m}) > 1$.

Our strategy in computing loss probabilities in this setting will be to exploit the ideas already developed in Sections 2 through 4. Note that the Poisson structure of the χ'_j's can be exploited just as in the earlier $M/G/s$ context, once one conditions on X. The main complication that is introduced is that the χ'_j's are then a sequence of non-identically distributed (but still conditionally independent) Poisson r.v.'s. However, since the key properties of the Poisson r.v.'s are preserved in the presence of non-stationarity, the algorithms and analysis of Sections 2 through 4 remain largely intact.

Kac's formula again applies in this context, yielding the relationship

$$\begin{aligned} &\frac{\sum_{j=0}^{n-1} \left[Q'_j - W'_j(m) + \chi'_{j+1} - s \right]^+}{\sum_{j=0}^{n-1} \chi'_j} \\ &\longrightarrow \beta'(s) \triangleq \frac{E_{\pi'_{A'}} \left(\sum_{j=0}^{T_{A'}-1} \left[Q'_j - W'_j(m) + \chi'_{j+1} - s \right]^+ \right)}{E_\mu \lambda(X_0)} \pi'(A') \end{aligned} \tag{22}$$

a.s. as $n \nearrow \infty$, where $E_{\pi'_{A'}}(\cdot)$ is the expectation operator under which

$$P\left(W'_0 \in \cdot, X_0 \in \cdot\right) = \pi'_{A'}(\cdot) \triangleq P\left(W'_\infty \in \cdot, X_\infty \in \cdot \,\middle|\, (W'_\infty, X_\infty) \in A'\right)$$

and $T_{A'} \triangleq \inf\{n \geq 1 : (W'_n, X_n) \in A'\}$ is the first return time to A'. Let x_* be selected so that $\rho' = \rho(x_*)$, put

$$\Delta = \left\{ w \in S : \left| \sum_{j=1}^{m} E\left(B^j Y_j \middle| X_j = x_*\right) - w \right| \leq a s^{-1/2} \sum_{j=1}^{m} E\left(B^j Y_j \middle| X_j = x_*\right) \right\}$$

(where the absolute value above is interpreted component by component) and set

$$A' \triangleq \Delta \times \{x_*\}.$$

In order to see that $\pi'(A')$ is bounded away from zero, note that

$$R'_\infty \stackrel{D}{=} R'_m = Y'_m + BY'_{m-1} + ... + B^{m-1}Y'_1, \tag{23}$$

assuming that X_0 is drawn from μ. Therefore,

$$\begin{aligned}\pi\left(A'\right) &= E_\mu\left(P\left(R_m \in \Delta, X_m = x_* \middle| X_0\right)\right) \\ &\geq \mu(x_*)\, p(x_*, x_*)^m\, P\left(R_m \in \Delta \middle| X_0 = ... = X_m = x_*\right).\end{aligned}$$

By the central limit theorem (see, for instance, Glynn and Whitt (1991) we have that $P(R_m \in \Delta | X_0 = ... = X_m = x_*)$ is bounded away from zero as $s \nearrow \infty$. On the other hand, exactly the same argument as that given in the proof of Proposition 1 yields

$$\sup_C \left|P\left(R'_\infty \in C\right) - P\left(W'_\infty \in C\right)\right| \leq mP\left(\left\|R'_\infty\right\| > s\right).$$

But, representation (23) implies

$$\begin{aligned}&P(\mathrm{Po}\left(s\rho'\right) > s)\mu(x_*)\, p(x_*, x_*)^m \\ &\leq P\left(\left\|R'_\infty\right\| > s\right) \leq P(\mathrm{Po}\left(s\rho'\right) > s)\end{aligned} \tag{24}$$

for all $s \geq 1$. Hence,

$$\frac{1}{s}\log P\left(\left\|R'_\infty\right\| > s\right) \longrightarrow -\left(\rho' - 1 - \log\rho'\right)$$

and we conclude that $\pi(A')$ is bounded away from zero as $s \nearrow \infty$. On this basis, we concentrate on the numerator of the expression for $\beta'(s)$, namely,

$$\kappa'(s) \triangleq E_{\pi'_{A'}}\left(\sum_{j=0}^{T_{A'}-1}\left[Q'_j - W'_j(m) + \chi'_{j+1} - s\right]^+\right), \tag{25}$$

which involves the occurrence of the rare event $\{T' < T_{A'}\}$ where

$$T' \triangleq \inf\{n \geq 1 : Q'_{n-1} - W'_{n-1}(m) + \chi'_n > s\}.$$

Computing (25) can be done following the same strategy described in Section 2 based on the construction of r.v.'s $\Gamma'_j(w,x)$, $j \geq 1$, with the property that

$$E\Gamma_j(w,x) = E_{(w,x)}\left(\sum_{j=0}^{T_{A'}-1}\left[Q'_j - W'_j(m) + \chi'_{j+1} - s\right]^+\right).$$

In turn, to construct such $\Gamma'_j(w,x)$'s we take advantage of the fact that

$$\begin{aligned}&E_{(w,x)}\left(\sum_{j=0}^{T_{A'}-1}\left[Q'_j - W'_j(m) + \chi_{j+1} - s\right]^+\right) \\ &= E_{(w,x)}\left(\phi'\left(W'_{T'-1}, X_{T'-1}, Y'_{T'}\right) I\left(T' < T_{A'}\right)\right),\end{aligned} \tag{26}$$

where

$$\phi'(w,x,y) = E\left(\sum_{j=0}^{T_{A'}-1}[Q'_j - W'_j(m) + \chi'_{j+1} - s]^+ \middle| W_0 = w, X_0 = x, Y'_1 = y\right),$$

$Y'_j = \left(Y'_j(1),...,Y'_j(m)\right)$ with the $Y'_j(i)$'s being conditionally independent given X and $E\left(Y'_j(i)\middle| X_j\right) = \lambda(X_j)P(V=i)$. The Monte Carlo evaluation of expectations such as (26) can be easily done if one takes advantage of an efficient importance sampling estimator for $T' < T_{A'}$, which we shall develop next.

The description of our estimator takes advantage of the associated infinite server system $(R'_n : n \geq 0)$ satisfying

$$R'_{n+1} = BR'_n + Y'_{n+1}$$

for $n \geq 0$. As in Section 3, the fact that $W'_n = R'_n$ for $n < T'$ (assuming $W'_0 = w = R'_0$) yields that

$$T' = \inf\{n \geq 1 : \left\|R'_n\right\| > s\},$$

and that $\left(W'_{T'-1}, X_{T'-1}, Y'_{T'}\right)$ given $T' < T_{A'}$ has the same distribution as $\left(R'_{T'-1}, X_{T'-1}, R'_{T'} - BR'_{T'-1}\right)$, conditional on $T' < \widetilde{T}_{A'}$, where

$$\widetilde{T}_{A'} = \inf\{n \geq 1 : R'_n \in A'\}.$$

Consequently, we concentrate our efforts on describing a sampler that is efficient for estimating the probability of the event $T' < \widetilde{T}_{A'}$.

We run a straightforward adaptation of the algorithm described in Section 3 conditional on the chain X. In order to provide an explicit representation of our importance sampling estimator it is convenient to introduce the notation $P_w^X(\cdot) = P_{(w,x)}(\cdot \,|\, X_0, X_1, ...)$; the associated expectation operator is denoted by $E_w^X(\cdot)$. We let

$$\widetilde{P}^X(\cdot) = P_w^X\left(\cdot \,|\, \left\|R'_\tau\right\| > s\right)$$

(with τ as defined in Section 3). Following the development leading to equation (4) we conclude that

$$\begin{aligned}&\widetilde{P}^X\left(\left(R'_1,...,R'_{T'}\right) \in \cdot\,, T' < \widetilde{T}_{A'}\right)\\ &= E_w^X\left(I\left(\left(R'_1,...,R'_{T'}\right) \in \cdot\,, T' < \widetilde{T}_{A'}\right)\frac{\sum_{n=T'}^{\infty} h_n^X\left(T', R'_{T'}\right)}{P_w^X\left(\|R'_\tau\| > s\right)}\right)\end{aligned} \tag{27}$$

where

$$h_n^X(k,r) = P(\tau = n)\, P_r^X\left(\left\|R'_n\right\| > s \middle| R'_k = r\right).$$

Our proposed importance sampling distribution is given by $\widetilde{P}(\cdot)$ defined via

$$\widetilde{P}(\cdot) = E_w\left(E_w^X\left(I\left(\left(R'_1,...,R'_{T'}\right) \in \cdot\,, T' < \widetilde{T}_{A'}\right)\frac{\sum_{n=T'}^{\infty} h_n^X\left(T', R'_{T'}\right)}{P_w^X\left(\|R'_\tau\| > s\right)}\right)\right),$$

(i.e. no importance sampling is applied to X).

In order to evaluate $h_n^X(k,r)$, note that $P_r^X(\|R'_n\| > s | R'_k = r)$ can be computed in closed form using the Poisson distribution. Now, in contrast to our analysis in Section 3, the pair of infinite series $\sum_{n=T'}^{\infty} h_n^X\left(T', R'_{T'}\right)$ and $\sum_{n=1}^{\infty} h_n^X(1,w) = P_w^X(\|R'_\tau\| > s)$ cannot be evaluated in closed form; instead, they have to be truncated at some level thereby introducing a bias. We will deal with this issue later, but for the moment it suffices to say that one can explicitly control the relative bias of the estimator by truncating at level $t(s) = s^{1+\delta}$ for any $\delta > 0$.

Our development in (27) yields the likelihood ratio identity

$$\frac{d\widetilde{P}}{dP_w} = L_X^{-1}\left(T', R'_{T'}\right) = \frac{\sum_{n=T'}^{\infty} h_n^X\left(T', R'_{T'}\right)}{P_w^X(\|R'_\tau\| > s)}.$$

One can generate paths under $\widetilde{P}(\cdot)$ by adapting Algorithm B. The adaptation of Algorithm B to our current setting proceeds as follows. First, step 1 requires simulating X. In practice, because of the truncation issue mentioned before, it suffices to simply generate $X_0, ..., X_{t(s)}$. Then one samples K' with probability mass function proportional to $P_w^X(\|R'_k\| > s) P(\tau = k) I(k \leq t(s))$. Steps 2 to 5 just involve obvious notational changes (i.e. replacing the Y_j's by Y'_j's and χ_j's by χ'_j's). Finally, in step 6, the output involves computing $L_X^{-1}\left(T', R'_{T'}\right)$ truncating the series up to $t(s)$.

In our current setting, there is also a direct adaptation of Algorithm C in order to generate $\|R'_l\|$ with $l \leq m$, conditional on X, $\|R'_l\| > s$ and $R'_0 = w$, and then $(Y'_1, ..., Y'_l)$ given $\|R'_l\|$, X and $R'_0 = w$. The idea is to fix X and use the fact that $\left\|R'_l - B^l R'_0\right\|$ is a Poisson random variable with mean sv_l^X, where

$$v_l^X \triangleq \frac{\rho(X_l) P(V \geq 1) + ... + \rho(X_1) P(V \geq l)}{EV}.$$

The generation of $\|R'_l\|$ conditional on $\|R'_l\| > s$, X and $R'_0 = w$, is equivalent to generating r.v.'s from the conditional distribution

$$P(||B^l w|| + \text{Po}\left(sv_l^X\right) \in \cdot \mid \text{Po}\left(sv_l^X\right) > s - ||B^l w||).$$

Sampling from the previous distribution can be done by means of acceptance rejection as suggested in Section 3. The expected number proposals required to obtain an acceptance is also of order $O\left(s^{1/2}\right)$ (uniformly over $X_1, ..., X_m$). In conclusion, a small variation of Algorithm C can be executed conditional on X. The only changes arise in step 4, where v_l is replaced by v_l^X, and in step 5 where the Z number of customers are distributed across the $\left(B^j Y'_{l-j}\right)(i)$'s (for $i \in \{1, ..., m\}$ and $0 \leq j < l$) according to a multinomial distribution with associated multinomial probability

$$\frac{E_w^X\left(B^j Y'_{l-j}\right)(i)}{sv_l^X} = \frac{E_w^X Y'_{l-j}(i-j)}{sv_l^X} = \frac{\lambda(X_{l-j}) P(V = i-j)}{sv_l^X}$$

for $\left(B^j Y'_{l-j}\right)(i)$.

6 Efficiency Analysis and Bias Control

The technical development behind the asymptotic behavior of $\beta'(s)$ as $s \nearrow \infty$ is similar to that of Section 4. In particular, we have the following result which parallels the statement of Theorem 1.

Theorem 3. *The following limits hold as* $s \nearrow \infty$

i.) $\left(\log P\left(\|R'_\infty\| > s\right)\right)^{-1} \log \beta'(s) \longrightarrow 1$
ii.) $\left(\log P\left(\|R'_\infty\| > s\right)\right)^{-1} \log \beta'(s) \longrightarrow 1$
iii.) $\left(\log E\left(\left[\|R'_\infty\| - s\right]^+\right)\right)^{-1} \log \beta'(s) \longrightarrow 1$
iv.) $\left(\log P\left(Q'_\infty = s\right)\right)^{-1} \log \beta'(s) \longrightarrow 1$
v.) $s^{-1} \log \beta'(s) \longrightarrow -[\rho' - 1 - \log(\rho')]$

Proof. The proof is completely analogous to that of Theorem 1, taking advantage of the bounds in (24) for parts i.), ii.) and iii.). The only significant change arises in the development of a suitable asymptotic lower bound for $s^{-1} \log P_{\pi'_{A'}}\left(\widetilde{T}' < T_{A'}\right)$, where $\widetilde{T}' = \inf\{n \geq 1 : Q'_n \geq s\} = \inf\{n \geq 1 : \|R'_n\| \geq s\}$. However, note that

$$\begin{aligned}
&P_{\pi_{A'}}\left(\widetilde{T}' < T_{A'}\right) \\
&\geq \int_{\Delta \times \{x_*\}} \pi_{A'}(dw, dx)\, P_{(w,x)}\left(\widetilde{T}' = m < T_{A'}\right) \\
&\geq \int_{\Delta \times \{x_*\}} \pi_{A'}(dw, dx)\, P\left(\left\|R'_m\right\| = s, X_1 = ... = X_m = x_*\right) \\
&\times P_w\left(\left\|R'_i\right\| < s, R'_i \notin A, i \leq m-1 \middle| \left\|R'_m\right\| = s, X_1 = ... = X_m = x_*\right).
\end{aligned}$$

The argument can be completed exactly as in the proof of Theorem 1 using the fact that conditional on $\|R'_m\| = s$ and on $X_0 = ... = X_m = x_*$, the random variables, $\left\|R'_i - B^i w\right\|$, $i \leq m-1$ are distributed $\text{Bin}(s, E\min(V, i)/EV)$. □

We now turn to the efficiency analysis of the estimator $\Gamma_1(w, x_*)^2 L_X\left(T', R'_{T'}\right)$. We have the following result which establishes logarithmic efficiency.

Theorem 4. *For each $w \in \Delta$, we have*

$$\underline{\lim}_{s \longrightarrow \infty} \frac{\log \widetilde{E}\left(\Gamma_1(w, x_*)^2 L_X\left(T', R'_{T'}\right)\right)}{\log \beta'(s)} = 2$$

satisfies (16).

Proof. The proof of Theorem 4 is similar to that of Theorem 2. One can use the Poisson r.v.'s with rate ρ' to stochastically dominate the count of loss customers. The main distinction arises in the analysis of the moments of $T_{A'}$, which are required to be uniformly bounded in s in order to complete the proof. Lemma 1 below then provides the necessary elements to analyze the moments of $T_{A'}$ thereby completing the proof of the result. □

The next lemma estimates the tails of $T_{A'}$. The analysis is useful both to complete the proof of Theorem 4 and to estimate the relative bias induced by the truncation at level $t(s)$ required to implement the procedure. To state our result first, let $\sigma + 1$ be the cardinality of the set $\mathcal{S}$. Since $p(\cdot)$ is irreducible and $p(x, x) > 0$ for $x \in \mathcal{S}$ it follows that $p(\cdot)$ is also aperiodic and therefore

$$\eta \triangleq \min_{y \in \mathcal{S}} p^\sigma(x, y) > 0.$$

Define

$$\theta = P\left(\cap_{j=1}^{m}\{\left|Y_1'(j) - EY_1'(j)\right| \leq as^{-1/2}EY_1'(j)\}\Big|\, X_1 = x_*\right)^m$$

and set

$$\phi = \eta p\left(x_*, x_*\right)^m \theta.$$

Observe that ϕ is bounded away from zero in s and that it can be computed explicitly using the Poisson distribution for each $s \geq 1$.

Lemma 1. *Let $l = \sigma + m - 1$, then for each $w \in \Delta$ and every $k \in \{0, 1, ...\}$*

$$P_{(w,x_*)}\left(T_{A'} > (k+1)l\right) \leq (1-\phi)^k.$$

Proof. Put

$$C_{lk+\sigma+i} = \cap_{j=1}^{m}\{\left|Y'_{lk+\sigma+i}\left(j\right) - EY'_{lk+\sigma+i}\left(j\right)\right| \leq as^{-1/2}EY'_{lk+\sigma+i}\left(j\right)\}$$

and set

$$\zeta\left(X_{kl}\right) = P\left(X_{lk+\sigma+i} = x_*, C_{lk+\sigma+i}, 0 \leq i \leq m-1 | \, X_{kl}\right).$$

Then, for each $w \in \Delta$ we have

$$\begin{aligned}
&P_{(w,x_*)}\left(T_{A'} > (k+1)l\right) \\
&\leq E_{(w,x_*)}\left(I\left(T_{A'} > kl\right)\left(1 - \zeta\left(X_{kl}\right)\right)\right) \\
&\leq P_{(w,x_*)}\left(T_{A'} > kl\right)\left(1-\phi\right).
\end{aligned}$$

Iterating the previous inequality we conclude the result. □

We finish this section with an estimate of the bias that arises by truncating the infinite series in the definition of $L_X^{-1}\left(T', R'_{T'}\right)$. Sampling K' according to the probability mass function

$$P\left(K' = k\right) = \frac{P_w^X\left(\|R'_k\| > s\right) P\left(\tau = k\right) I\left(\tau \leq t\left(s\right)\right)}{P\left(\|R'_\tau\| > s | \, \tau \leq t\left(s\right)\right)}$$

as suggested for the practical implementation of our procedure induces an unbiased estimator for the expectation

$$E_{\pi'_{A'}}\left(\phi'\left(W'_{T'-1}, X_{T'-1}, Y'_{T'}\right) I\left(T' < T_{A'}, T' \leq t\left(s\right)\right)\right).$$

The bias obtained when estimating the numerator of $\beta'\left(s\right)$ is then bounded by

$$\begin{aligned}
&E_{\pi'_{A'}}\left(\phi'\left(W'_{T'-1}, X_{T'-1}, Y'_{T'}\right) I\left(t\left(s\right) < T' < T_{A'}\right)\right) \\
&\leq E_{\pi'_{A'}}\left(\phi'\left(W'_{T'-1}, X_{T'-1}, Y'_{T'}\right) I\left(t\left(s\right) < T' < T_{A'}\right)\right) \\
&\leq E_{\pi'_{A'}}\left(\phi'\left(W'_{T'-1}, X_{T'-1}, Y'_{T'}\right)^2 I\left(T' < T_{A'}\right)\right)^{1/2} P_{\pi'_{A'}}\left(T_{A'} > t\left(s\right)\right)^{1/2} \\
&\leq E_{\pi'_{A'}}\left(\phi'\left(W'_{T'-1}, X_{T'-1}, Y'_{T'}\right)^2 I\left(T' < T_{A'}\right)\right)^{1/2} \left(1-\phi\right)^{\lfloor t(s)/(2l)\rfloor}.
\end{aligned}$$

Therefore, the relative bias can be bounded by

$$\frac{E_{\pi'_{A'}}\left(\phi'\left(W'_{T'-1},X_{T'-1},Y'_{T'}\right)^2 I\left(T'<T_{A'}\right)\right)^{1/2}}{E_{\pi'_{A'}}\left(\phi'\left(W'_{T'-1},X_{T'-1},Y'_{T'}\right) I\left(T'<T_{A'}\right)\right)}\left(1-\phi\right)^{\lfloor t(s)/(2l)\rfloor}$$
$$\leq \frac{E_{\pi'_{A'}}\left(\phi'\left(W'_{T'-1},X_{T'-1},Y'_{T'}\right)^2 I\left(T'<T_{A'}\right)\right)^{1/2}}{P_{\pi'_{A'}}\left(T'<T_{A'},T'\leq t\left(s\right)\right)}\left(1-\phi\right)^{\lfloor t(s)/(2l)\rfloor}.$$

The right hand side of the previous expression decreases super-exponentially fast in s if $t\left(s\right)$ is super-linear in s. Moreover, the term $\left(1-\phi\right)^{\lfloor t(s)/(2l)\rfloor}$ can be easily estimated since ϕ can be explicitly evaluated. On the other hand, the probability $P_{\pi'_{A'}}\left(T'<T_{A'},T'\leq t\left(s\right)\right)$ can also be estimated efficiently via Monte Carlo (the algorithms explained above allow to do precisely this); furthermore exponential decay of such probability coincides with that studied in Theorem 3. Consequently, one can select $t\left(s\right)=\kappa s^{3/2}$ for a suitable constant $\kappa>0$ and estimate the ratio

$$\left(1-\phi\right)^{\lfloor t(s)/(2l)\rfloor}/P_{\pi'_{A'}}\left(T'<T_{A'},T'\leq t\left(s\right)\right)$$

with good relative precision and high confidence. Note that an upper bound for

$$E_{\pi'_{A'}}[\phi'\left(W'_{T'-1},X_{T'-1},Y'_{T'}\right)^2 I\left(T'<T_{A'}\right)]^{1/2}$$

can easily be obtained either by explicitly or by simulation; high relative precision is not required for this calculation. Combining these observations we conclude that the relative bias of our estimator can be reduced to a desired accuracy at a relatively low (sub-exponential in s) computational cost.

We close our discussion on Markov modulated models with the adaptation of our ideas to the case in which the service times are allowed to depend on X. We let $V\left(x\right)$ be a generic service time corresponding to a customer that arrives at a time at which the underlying Markov chain takes value x. Note that the representation given for R'_{∞} given in equation (23) is applicable. In fact, Theorem 3 holds in this case by suitably chainging the definition of ρ', which now takes the form

$$\rho'=\max_{\{x_1,...,x_m:p(x_1,x_2)\times....\times p(x_{m-1},x_m)>0\}}\sum_{l=0}^{m-1}\sum_{j=l+1}^{m}\lambda\left(x_{m-l}\right)P\left(V\left(x_{m-l}\right)=j\right). \tag{28}$$

As before, $\rho'<1$ is a condition required to make the loss probability converge to zero as $s\nearrow\infty$.The difference between our previous development involves finding a suitable definition for the set A'. Note that our previous analysis relies on a definition of A' that is guided by a suitable central limit theorem argument. At the crux of our arguments is the fact that one can isolate the most likely m-step trajectory in the Markov modulated chain and, conditional on $||R_m||=s$, there is a law of large numbers description that moves the vector valued process outside the set A' given a suitable trajectory for the process X (at this step it is useful to have defined A via the central limit theorem).

When the service times depend on the underlying Markov modulated chain one needs to expand the size of the state descriptor in order to appropriately define a suitable equilibrium set A'. In particular, one needs to define $R'_n(i,j)$ as the number of customers that are of type j (i.e. customers that arrived at a time when the underlying Markov chain took value j) and that have i units of remaining service requirement at time n. Then,

$$R_{n+1}(i,j)=R_n(i+1,j)I(i\leq m-1)+Y_{n+1}(i,j)I\left(X_{n+1}=j\right),$$

where $Y_{n+1}(i,j)$ is Poisson with rate $s\lambda(j)\,P(V(j)=i)$ and all the $Y_n(i,j)$'s are independent. The appropriate definition of the set A' now can be done in terms of the solution to the optimization problem (28). In particular, suppose that the solution to this optimization problem is given by a sequence $(x_1^*,...,x_m^*)$. Then, consider the deterministic system

$$\overline{r}_{n+1}(i,j) = \overline{r}_n(i+1,j)I(i \leq m-1) + \overline{y}_{n+1}(i,j)I\left(x_{n+1}^* = j\right),$$

where $\overline{y}_{n+1}(i,j) = s\lambda(j)\,P(V(j)=i)$ for all $0 \leq n \leq m-1$ assuming that $\overline{r}_0(i,j) = 0$. Then we put $A' \subseteq R^{m\times|\mathcal{S}|} \times \mathcal{S}$ where $|\mathcal{S}|$ is the cardinality of the state-space of the underlying Markov chain defined as follows. First put

$$\Delta = \{r(i,j) : |r(i,j) - \overline{r}_m(i,j)| \leq s^{1/2}, 1 \leq i \leq m, 1 \leq j \leq |\mathcal{S}|\}$$

and let

$$A' = \Delta \times \{x_1^*\}.$$

Both the implementation indicated in Section 5 and the analysis given in this section carry over in a completely analogous manner.

7 Numerical Experiments

We applied our procedure to a couple of examples both in the standard $M/G/s$ case and the model with Markov modulated input. We tested the performance of our algorithms for different sizes of estimated loss probabilities, ranging from quantities of order 10^{-1} to 10^{-10}. This range allows us to test the performance of our algorithms relative to crude Monte Carlo in cases in which the events are not very rare.

We use the method of batch means to provide an associated confidence interval for the loss probability of interest. The number of batches is equal to 20 in all the experiments. In order to make the performance analysis comparable in terms of the coefficient of variation of the estimators we fixed a CPU time budget of roughly 120 seconds (plus the time required to complete the calculations for the last batch in order to make all the batches of the same size). The column "Estimator" provides the corresponding point estimate associated with the corresponding method. C.V. is the empirical coefficient of variation (empirical standard deviation divided by the estimator). Finally, we display an approximate 95% confidence interval.

Our first set of results, shown in the next two tables, correspond to the basic $M/G/s$ model explained in Section 2 $m = 10$. We assume that the service times are uniform on $\{1,2,...,10\}$. The traffic intensity, ρ, is set equal to 0.1. We observe that substantial variance reduction is observed for loss probabilities of order 10^{-4}. For probabilities of order 10^{-8} our procedure is vastly superior to crude Monte Carlo. It is worth noting that the behavior of the empirical coefficient of variation for our importance sampler suggests even better performance than simply logarithmic efficiency.

Crude Monte Carlo

	Estimator	C.V.	95% C.I.
$s = 10$	0.10007	0.062	[0.09563, 0.10451]
$s = 25$	0.01092	0.297	[0.00857, 0.01322]
$s = 50$	3.2307e-04	1.165	[5.3748e-05, 5.9240e-04]
$s = 100$	0	N/A	N/A

Importance Sampling

	Estimator	C.V.	95% C.I.
$s = 10$	0.09651	0.462	[0.06455, 0.12847]
$s = 25$	0.00763	0.894	[0.00274, 0.01251]
$s = 50$	2.5603e-05	0.879	[9.4973e-06, 4.1709e-05]
$s = 100$	6.8965e-10	1.289	[5.3544e-11, 1.3257e-09]

For the case of Markov modulated input we assume that the service times are uniformly distributed on the set $\{1, 2, ..., 5\}$. The transition matrix is

$$p = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

and the traffic intensities are given values $\rho(1) = 0.1$ and $\rho(2) = 0.2$. The relative bias was set less than 1% using $t(s) = s^{3/2}$ and applying the bounds obtained in the previous section. The conclusions in terms of algorithmic performance are consistent with our previous observations. Our numerical output is shown in the next tables.

Crude Monte Carlo

	Estimator	C.V.	95% C.I.
$s = 10$	0.05278	0.086	[0.04951, 0.05604]
$s = 25$	0.00473	0.200	[0.00405, 0.00541]
$s = 50$	1.3461e-04	1.943	[-5.2554e-05, 3.2178e-04]
$s = 100$	0	*N/A*	*N/A*

Importance Sampling

	Estimator	C.V.	95% C.I.
$s = 10$	0.04912	0.373	[0.03613, 0.06217]
$s = 25$	0.00476	1.254	[0.00048, 0.0090]
$s = 50$	6.6055e-05	0.742	[3.0948e-05, 1.0116e-04]
$s = 100$	5.6104e-08	0.889	[2.0426e-08, 9.1782e-08]

Acknowledgement. *We are grateful for the suggestions made by the referee and Associate Editor. This research was partially supported by grants NSF-0902075 and NSF-0846816.*

References

[1] Asmussen, S. (1985), Conjugate processes and the simulation of ruin problems. *Stochastic Processes and their Applications*, **20**, 213-229.

[2] Asmussen, S, and Glynn, P. W. (2007), *Stochastic Simulation: Algorithms and Analysis*. Springer, New York.

[3] Breiman, L. (1968), *Probability*. Addison-Wesley, Massachusetts.

[4] Chang, C. S., Heidelberger, P., Juneja, S., and Shahabuddin, P. (1994), Effective bandwidth and fast simulation of ATM intree networks. *Performance Evaluation*, **20**, 45-65.

[5] Cottrell, M., J, C. Fort and G, Matgouvres (1983), Large deviations and rare events in the study of stochastic algorithms. *IEEE Transactions on Automatic Control*, **28**, 907-920.

[6] Glynn, P. W. (1995), Large deviations for the infinite server queue in heavy traffic. *IMA Vol. 71 in Mathematics and its Applications*, Springer-Verlag, 387-395.

[7] Glynn, P. W. and Whitt, W. (1991) A new view of the heavy-traffic limit theorem for many-server queues. *Advances in Applied Probability*, **23**, 188-209.

[8] Goyal, A., Shahabuddin, P., Heidelberger, P., Nicola, V. F., and Glynn, P. W. (1992), A unified framework for simulating Markovian models of highly dependable systems. *IEEE Transactions on Computers*, **41**, 36-51.

[9] Heidelberger, P. (1995), Fast simulation of rare events in queueing and reliability models, *ACM Transactions on Modeling and Computer Simulation (TOMACS)*. **5**1, 43-85.

[10] Juneja, S., and Shahabuddin, P. (2006). Rare event simulation techniques: an introduction and recent advances. *Handbook in Operations Research and Mangement Sciences*, Vol. 13: Simulation, Chapter 11, 291-350, Henderson S., and Nelson B. (Eds.), Elsevier.

[11] Kelly, F. (1979) *Reversibility and Stochastic Networks.* Wiley, Chichester.

[12] Siegmund, D. (1976) Importance sampling in the Monte Carlo study of sequential tests. *Annals of Statistics*, **4**, 673-684.

[13] Srikant, R., and Whitt W. (1996), Simulation run lengths to estimate blocking probabilities. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, **6**, 7 - 52.

[14] Srikant, R., and Whitt W. (1999), Variance reduction in simulations of loss models. *Operations Research*, **47**, 509-523.

[15] Szechtman, R., and Glynn., P. W. (2002), Rare event simulation for infinite server queues. *Proceedings of the 2002 Winter Simulation Conference*, Yucesan, E., Chen, C. -H., Snowdon, J. L., and Charnes, J. M. (Eds.), 416-423.