RARE-EVENT SIMULATION FOR INFINITE SERVER QUEUES

Roberto Szechtman
Peter W. Glynn

Department of Management Science and Engineering
Stanford University
Stanford, CA 94305 U.S.A.

ABSTRACT

We discuss rare-event simulation methodology for computing tail probabilities for infinite-server queues. Our theoretical discussion also offers some new simulation insights into the change-of-measure associated with the Gärtner-Ellis theorem of large deviations.

1 INTRODUCTION

This paper is concerned with rare-event simulation in the setting of infinite-server queues. Infinite-server queues play an important role in queueing theory, as they form a mathematical idealization of systems in which many servers are present. In particular, if a queue possesses a large number of servers, the structure of the infinite-server queue is largely inherited by the many-server system, provided that the fraction of time that the many-server system has all servers busy is small.

Many-server queues have played a fundamental role in the telecommunications modeling environment over the years. In this setting, circuits can be identified with servers. In view of the large number of circuits that are typically available to carry traffic, a many-server queueing model is often appropriate. Furthermore, quality-of-service considerations guarantee that the system will be engineered in such a way that the probability of finding all the servers busy is small.

In the telecommunications setting, such many-server queues typically exhibit "loss" whenever all the servers are busy. In other words, connections are refused whenever all the circuits are busy. The corresponding "loss probability" is a fundamental performance measure for such systems. The tail probability for the number-in-system process for the associated infinite-server queue often is a good approximation to the loss probability for the many-server system. As a consequence, efficient computation of such tail probabilities for the infinite-server queue is of clear applied relevance.

In addition, infinite-server queues form an important class of models in their own right. In addition to their mathematical importance within the queueing context, they arise naturally in the study of electric power consumption. The number of electric power users consuming electricity can be viewed as the number-in-system process for an infinite-server queue. Thus, a tail probability for the infinite-server queue provides important information on peak load demand characteristics for an electric power grid.

This paper is specifically concerned with the use of rare-event simulation as a means of computing tail probabilities for the infinite-server queue. In particular, we develop efficient algorithms for computing tail probabilities for infinite-server queues with a high average arrival rate. In view of the telecommunications and electric power examples described above, this asymptotic setting seems especially natural.

This paper is organized as follows. Section 2 offers a problem formulation and describes the basic estimation approach we shall utilize. In Section 3, we survey related large deviations theory, while Section 4 provides additional discussion of our proposed algorithm. Computational results are given in Section 5.

2 PROBLEM FORMULATION AND BASIC RESULTS

We start by giving a precise description of the $GI/GI/\infty$ queue. Suppose that $(A_k : k \geq 1)$ is a nondecreasing sequence in which A_k corresponds to the arrival time of the k'th customer. If the system starts empty at $t = 0$, and if V_j denotes the "time-in-system" (or "processing time") of the j'th customer, then the number of customers $Q(t)$ in the system at time t is given by

$$Q(t) = \sum_{k=1}^{\infty} I(A_k \leq t \leq A_k + V_k).$$

Let $N(t) = \max\{n \geq 0 : A_n \leq t\}$ be the counting process corresponding to the number of arrivals in $[0, t]$.

(By convention, we set $A_0 = 0$.) Then, $Q(t)$ can be re-expressed in terms of $N(\cdot)$ as

$$Q(t) = \sum_{k=1}^{N(t)} I(A_k + V_k > t),$$

where $I(B)$ is the indicator random-variable (rv) associated with the event B.

Our goal here is to efficiently compute $P(Q(t) > x)$, where x is so large that $\{Q(t) > x\}$ is a "rare-event". Throughout this paper, we will assume that:

A1. $V = (V_n : n \geq 1)$ is a sequence of independent and identically distributed (iid) random variables, independent of $N = (N(t) : t \geq 0)$.

With this assumption in force, we find that

$$EQ(t) = \int_0^t P(V > t - s) EN(ds).$$

Thus, if $N = (N(t) : t \geq 0)$ is a point process with stationary increments and arrival intensity $\lambda = EN(1)$, it follows that

$$EQ(t) = \lambda \int_0^t \bar{F}(s) ds,$$

where $\bar{F}(t) = P(V > t)$ and $F(t) = P(V \leq t)$. The event $\{Q(t) > x\}$ will therefore tend to be "rare" when $x \gg \lambda \int_0^t \bar{F}(s) ds$.

Our approach to computing $\alpha = P(Q(t) > x)$ will be to apply importance sampling, with the selection of the importance sampling distribution guided by the principles of large deviations theory (Bucklew 1990). The study of large deviations suggests first computing the moment generating function of the rv $Q(t)$. Under A1, we note that

$$\begin{aligned}
&E \exp(\theta Q(t)) \\
&\quad = EE[\exp(\theta Q(t))|N] \\
&\quad = EE[\exp(\theta \sum_{i=1}^{N(t)} I(A_k + V_k > t))|N] \\
&\quad = E \prod_{k=1}^{N(t)} E[\exp(\theta I(V_k > A_k - t))|N] \\
&\quad = E \prod_{k=1}^{N(t)} \left(e^{\theta} \bar{F}(t - A_k) + F(t - A_k) \right) \\
&\quad = E \exp \left(\sum_{k=1}^{N(t)} \log \left(e^{\theta} \bar{F}(t - A_k) + F(t - A_k) \right) \right) \\
&\quad = E \exp \left(\int_{[0,t]} \log(e^{\theta} \bar{F}(t - s) + F(t - s)) N(ds) \right).
\end{aligned}$$

Set $\psi_Q(\theta) = \log E \exp(\theta Q(t))$. Suppose that there exists a positive root θ^* of the equation

$$\psi'_Q(\theta^*) = x.$$

The idea is to then generate variates from the "exponentially twisted" distribution given by

$$P^*(d\omega) = \exp(\theta^* Q(t, \omega) - \psi_Q(\theta^*)) P(d\omega.)$$

If $E^*(\cdot)$ is the expectation operator corresponding to P^*, then $\alpha = P(Q(t) > x)$ can be expressed in terms of $E^*(\cdot)$ via the relation

$$\alpha = E^* \exp(-\theta^* Q(t) + \psi_Q(\theta^*)) I(Q(t) > x).$$

The importance sampling algorithm for computing α now involves first simulating iid replicates of the rv

$$W = \exp(-\theta^* Q(t) + \psi_Q(\theta^*)) I(Q(t) > x)$$

under the probability P^*. The estimator for α is then obtained as the sample mean of the replicates generated.

Example 1. Suppose that $N = (N(t) : t \geq 0)$ is a Poisson process with rate $\alpha > 0$. It is known that the distribution of $Q(t)$ in this $M/G/\infty$ setting is Poisson distributed with parameter $\lambda \int_0^t \bar{F}(s) ds$ (see, for example, page 39 in Ross (1983)). Consequently,

$$\psi_Q(\theta) = \lambda \int_0^t \bar{F}(s) ds (e^{\theta} - 1).$$

Note that $\theta^* = \log(x/(\lambda \int_0^t \bar{F}(s) ds))$ and

$$\begin{aligned}
P^*(d\omega) &= \exp(\theta^* Q(t, \omega) - \lambda(e^{\theta^*} - 1) \int_0^t \bar{F}(s) ds) P(d\omega) \\
&= \exp(\theta^* Q(t, \omega) - (x - \lambda \int_0^t \bar{F}(s) ds)) P(d\omega)
\end{aligned}$$

In particular,

$$\begin{aligned}
&P^*(Q(t) = k) \\
&\quad = \exp(\theta^* k - (x - \lambda \int_0^t \bar{F}(s) ds)) P(Q(t) = k) \\
&\quad = \exp(\theta^* k - (x - \lambda \int_0^t \bar{F}(s) ds)) \\
&\qquad \cdot \left(\frac{(\lambda \int_0^t \bar{F}(s) ds)^k}{k!} \exp(-\lambda \int_0^t \bar{F}(s) ds) \right) \\
&\quad = \exp(-x) \frac{x^k}{k!}.
\end{aligned}$$

Thus, for a Poisson arrival stream, our algorithm replicates

$$\exp(-\theta^* Q(t) - (x - \lambda \int_0^t \bar{F}(s) ds)) I(Q(t) > x),$$

where $Q(t)$ is generated under P^* so that it has a Poisson distribution with mean x. So, our algorithm can be easily implemented in the Poisson setting.

In the next section, we offer some motivation for our choice of P^* as an importance distribution.

3 IMPORTANCE SAMPLING AND THE GÄRTNER-ELLIS THEOREM

As was mentioned in the Introduction, the typical real-world modeling environment that leads to infinite-server queues is one in which the arrival rate is large. Thus, we will consider here the so-called "heavy-traffic" asymptotic regime for infinite-server queues, in which we examine the behavior of a sequence of infinite server queues having an arrival rate tending to infinity.

Let $N = (N(t) : t \geq 0)$, $V = (V_j : j \geq 1)$ and $(A_n : n \geq 0)$ be defined as in Section 2. To send the infinite-server queue into heavy-traffic, we speed up the arrival process by a factor of n, leaving the processing times unchanged. More specifically, let

$$N_n(t) = N(nt)$$

be the arrival process feeding the n'th system; the arrival time of customer j in the n'th system is then A_j/n. The number-in-system process for the n'th system is then given by

$$Q_n(t) = \sum_{j=1}^{N_n(t)} I\left(\frac{A_j}{n} + V_j > t\right)$$

The mean of $Q_n(t)$ is easily seen to be

$$EQ_n(t) = n\lambda \int_0^t \bar{F}(s)ds.$$

Thus, a "rare-event" for the n'th system is a deviation in which $Q_n(t) > xn$, where $x > \lambda \int_0^t \bar{F}(s)ds$. We are interested in efficient computation of

$$\alpha_n = P(Q_n(t) > xn)$$

when n is large.

The Gärtner-Ellis large deviations theorem describes the asymptotic behavior of α_n for n large. It is generally stated in an abstract form, and concerns a sequence of real-valued rv's $(\beta_n : n \geq 1)$. The main hypothesis underlying the Gärtner-Ellis theorem is the following:

A2. There exists a real-valued function $\psi_\beta(\cdot)$ such that

$$\frac{1}{n} \log E \exp(\theta\beta_n) \to \psi_\beta(\theta) \text{ as } n \to \infty.$$

Assuming that we wish to approximate the probability $P(\beta_n > nx)$, we also require:

A3. There exist positive constants θ^*_β and ϵ such that $\psi_\beta(\cdot)$ is continuously differentiable and strictly increasing on $[-\epsilon, \theta^*_\beta + \epsilon]$, with $\psi'_\beta(0) < \psi'_\beta(\theta^*_\beta) = x$. The following result is due to Gärtner and Ellis (see page 15 of Bucklew 1990).

Theorem 1. Under hypotheses A2 and A3,

$$\frac{1}{n} \log P(\beta_n > nx) \to -\theta^*_\beta x + \psi_\beta(\theta^*_\beta)$$

as $n \to \infty$.

To apply this result to the analysis of $\alpha_n = P(Q_n(t) > xn)$, we set $\beta_n = Q_n(t)$. The validation of hypothesis A2 requires the following condition on the counting process N:

A4. There exists a finite-valued function ψ_N such that for $0 = t_0 < t_1 < \ldots < t_m = t$ and $(\theta_1, \theta_2, \ldots, \theta_m) \in \mathbb{R}^m$,

$$\frac{1}{n} \log E \exp(\sum_{i=1}^{m} \theta_i [N(nt_i) - N(nt_{i-1})]) \to \sum_{i=1}^{m} \psi_N(\theta_i)(t_i - t_{i-1})$$

as $n \to \infty$.

This assumption is satisfied by many different arrival processes; see Dembo and Zajic (1993). The function ψ_N will now be described in a couple of different modeling contexts.

Example 2. Suppose that the arrival process is renewal, so that A_k can be represented as $A_k = U_1 + \cdots + U_k$, where $(U_k : k \geq 1)$ is iid. Under suitable regularity conditions on the U_k's, Glynn and Whitt (1994) show that

$$\psi_N(\theta) = -\kappa^{-1}(-\theta),$$

where $\kappa(\theta) = \log(E \exp(\theta U_1))$, and $\kappa^{-1}(\cdot)$ is the inverse function to κ (ie. $\kappa(\kappa^{-1}(\theta)) = \kappa^{-1}(\kappa(\theta)) = \theta$).

Example 3. Here, we consider a Markov-modulated Poisson process. In other words, there exists an S-valued continuous-time Markov chain $X = (X(t) : t \geq 0)$ with generator B and function $f : S \to (0, \infty)$ such that the intensity of the Poisson (arrival) process at time t is $f(X(t))$. Suppose B is finite and irreducible. Then, $\psi_N(\theta)$ is the eigenvalue of $B + D(\theta)$ having maximal real part, where

$D(\theta) = \mathrm{diag}((e^\theta - 1)f(x) : x \in S)$.

Under Assumption A4, Glynn (1995) proves the following theorem.

Theorem 2. If Assumption A1 and A4 hold, then

$$\frac{1}{n}\log E\exp(\theta Q_n(t)) \to \int_0^t \psi_N(\log(e^\theta \bar{F}(x) + F(x)))dx,$$

as $n \to \infty$.

Set

$$\nu(\theta) = \int_0^t \psi_N(\log(e^\theta \bar{F}(s) + F(s)))ds.$$

Suppose that the function $\nu(\cdot)$ has the property that there exist positive constants θ^*_∞ and ϵ such that $\nu(\cdot)$ is continuously differentiable and strictly increasing on $[-\epsilon, \theta^*_\infty + \epsilon]$, with $\nu'(0) < \nu'(\theta^*_\infty) = x$. Then, Theorem 1 ensures that

$$\frac{1}{n}\log P(Q_n(t) > xn) \to -\theta^*_\infty x + \nu(\theta^*_\infty) \qquad (1)$$

as $n \to \infty$. The above limit suggests the approximation

$$P(Q_n(t) > xn) \approx \exp(n(-\theta^*_\infty x + \nu(\theta^*_\infty))),$$

when n is large.

Simulation offers a means of computing $P(Q_n(t) > xn)$ to a much higher level of precision than that associated with the approximation. The rare-event simulation algorithm proposed in Section 2, when applied to the computation of $\alpha_n = P(Q_n(t) > xn)$, suggests the use of the importance distribution

$$P^*_n(d\omega) = \exp(\theta^*_n Q_n(t,\omega) - \log E\exp(\theta^*_n Q_n(t)))P(d\omega),$$

where θ^*_n is the root of $d/d\theta \log E\exp(\theta^*_n Q_n(t)) = xn$ and $P(\cdot)$ is the original probability associated with the probability space supporting $Q_n(t)$. An estimator for α_n is then obtained via the sample mean of replications of the rv

$$W_n = \exp(-\theta^*_n Q_n(t) + \log E\exp(\theta^*_n Q_n(t)))I(Q_n(t) > xn)$$

simulated under the distribution P^*_n. Let $E^*_n(\cdot)$ be the expectation operator corresponding to P^*_n. The Cauchy-Schwarz inequality implies that for any unbiased estimator W_n of α_n,

$$EW_n^2 \geq (EW_n)^2 = \alpha_n^2.$$

Thus, under the conditions leading to (1), we have

$$\underline{\lim}_{n\to\infty} \frac{1}{n}\log EW_n^2 \geq -2\theta^*_\infty x + 2\psi_N(\theta^*_\infty). \qquad (2)$$

We will show momentarily that the lower bound on the right-hand side of (2) is achieved asymptotically by simulating W_n under P^*_n. In other words, the rv W_n, when simulated under P^*_n, achieves (in logarithmic scale) the highest possible asymptotic efficiency (in the sense of minimizing the second moment of the estimator). We view this as an asymptotic justification for our use of the algorithm suggested in Section 2.

In fact, this result holds in great generality. To make this point clear, we shall show that the result holds in the general Gärtner-Ellis setting.

Theorem 3. Assume hypotheses A2 and A3 hold.
i.) Let $(W_n : n \geq 1)$ be a sequence of estimators of $P(\beta_n > nx)$ that is unbiased, in the sense that $EW_n = P(\beta_n > nx)$. Then,

$$\underline{\lim}_{n\to\infty} \frac{1}{n}\log EW_n^2 \geq -2\theta^*_\beta x + 2\psi_\beta(\theta^*_\beta).$$

ii.) Suppose that $\tilde{W}_n = \exp(-\tilde{\theta}_n\beta_n + \log E\exp(\tilde{\theta}_n\beta_n))I(\beta_n > nx)$ is simulated under $\tilde{P}_n(d\omega) = \exp(\tilde{\theta}_n\beta_n(\omega)) - \log E\exp(\tilde{\theta}_n\beta_n))P(d\omega)$, where $\tilde{\theta}_n$ is the root of $d/d\theta \log E\exp(\tilde{\theta}_n\beta_n) = nx$ and $P(\cdot)$ is the original probability associated with β_n. Then

$$\tilde{E}_n\tilde{W}_n = P(\beta_n > nx)$$

and

$$\lim_{n\to\infty} \frac{1}{n}\log \tilde{E}_n\tilde{W}_n^2 = -2\theta^*_\beta x + 2\psi_\beta(\theta^*_\beta),$$

where $\tilde{E}_n(\cdot)$ is the expectation operator associated with $\tilde{P}_n(\cdot)$.

Proof. Part i.) follows in the same way as does (2) above. For part ii.), it is easily verified that $\log E\exp(\theta\beta_n)$ is convex in θ; see Dembo and Zeitouni (1998). It follows from convexity that

$$\frac{1}{n}\frac{d}{d\theta}\log E\exp(\theta\beta_n) \to \frac{d}{d\theta}\psi_\beta(\theta)$$

as $n \to \infty$; see Dembo and Zeitouni (1998). Furthermore, $\psi'_\beta(\cdot)$ is continuous and strictly increasing on $(-\epsilon, \theta^*_\beta + \epsilon)$. It therefore follows easily that $\tilde{\theta}_n \to \theta^*_\beta$ as $n \to \infty$ and $n^{-1}\log E\exp(\tilde{\theta}_n\beta_n) \to \psi_\beta(\theta^*_\beta)$ as $n \to \infty$. Since $\tilde{\theta}_n$ is positive for n sufficiently large,

$$\begin{aligned}\tilde{W}_n &= \exp(-\tilde{\theta}_n\beta_n + \log E\exp(\tilde{\theta}_n\beta_n))I(\beta_n > nx)\\ &\leq \exp(-\tilde{\theta}_n nx + \log E\exp(\tilde{\theta}_n\beta_n))I(\beta_n > nx)\\ &\leq \exp(-\tilde{\theta}_n nx + \log E\exp(\tilde{\theta}_n\beta_n)).\end{aligned}$$

Hence,

$$\frac{1}{n}\log \tilde{E}_n W_n^2 \leq -2\tilde{\theta}_n x + \frac{2}{n}\log E\exp(\tilde{\theta}_n\beta_n).$$

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

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log \tilde{E}_n \tilde{W}_n^2 \leq -2\theta_\beta^* x + 2\psi_\beta(\theta_\beta^*).$$

Because $\tilde{W}_n$ is clearly unbiased for estimation of $P(\beta_n > nx)$, part i.) immediately yields

$$\lim_{n\to\infty} \frac{1}{n} \log \tilde{E}_n W_n^2 = -2\theta_\beta^* x + 2\psi_\beta(\theta_\beta^*),$$

as desired.

Thus, the importance distribution $\tilde{P}_n$ is always guaranteed to yield a "change-of-measure" that is asymptotically optimal (in logarithmic scale). So, the Gärtner-Ellis theory establishes that the importance sampling algorithm introduced in Section 2 is asymptotically optimal in "heavy traffic".

Unfortunately, the importance distribution P^* suggested in Section 2 is, in general, impossible to implement from a practical standpoint. While implementation is clearly possible when the distribution of $Q(t)$ is known, such knowledge will never be available in situations of practical interest (for in such cases, simulation would be unnecessary). Any realistic implementation of importance sampling must involve describing the change-of-measure at the level of the "building-blocks" of the process. In the setting of the infinite-server queue, the change-of-measure must be described at the level of the inter-arrival times and processing times. Since P^* does not lend itself to such a description (because it "twists" $Q(t)$ and not the inter-arrival times and processing times), we must instead search for an alternative change-of-measure that (hopefully) coincides asymptotically with P^*. The same general remarks unfortunately also apply to the Gärtner-Ellis change-of-measure described in Theorem 3.

In Section 4, we explore an alternative change-of-measure to P^* that has the appropriate asymptotic structure.

4 AN IMPLEMENTABLE RARE-EVENT SIMULATION ALGORITHM

We wish to find a "change-of-measure" that coincides asymptotically (for large n) with the probability P_n^* discussed earlier. Recall that P_n^* is defined through the "twisting-parameter" θ_n^*, where $\theta_n^* \to \theta_\infty^*$ and θ_∞^* is defined as the root of

$$\nu'(\theta_\infty^*) = \int_0^t \psi_N'(e^{\theta_\infty^*}\bar{F}(t-s) + F(t-s)) \cdot \frac{e^{\theta_\infty^*}\bar{F}(t-s)}{e^{\theta_\infty^*}\bar{F}(t-s) + F(t-s)} ds = x.$$

Recall that $e^{\theta_\infty^*}\bar{F}(t-s)/(e^{\theta_\infty^*}\bar{F}(t-s) + F(t-s))$ is the parameter of a Bernoulli random variable having mean $e^{\theta_\infty^*}\bar{F}(t-s)/(e^{\theta_\infty^*}\bar{F}(t-s) + F(t-s))$. Note that, $\psi_N'(e^{\theta_\infty^*}\bar{F}(t-s) + F(t-s))$ is the (asymptotic) mean of the arrival process associated with exponential twist $e^{\theta_\infty^*}\bar{F}(t-s) + F(t-s)$. This suggests an importance sampling algorithm in which the arrival process (or, equivalently, the inter-arrival times) is twisted at time s to have instantaneous arrival rate $\psi_N'(e^{\theta_\infty^*}\bar{F}(t-s)+F(t-s))$, and the Bernoulli rv indicating that a customer arriving at time s stays until time t (i.e. has a processing time greater than $t-s$) is twisted to have mean $e^{\theta_\infty^*}\bar{F}(t-s)/(e^{\theta_\infty^*}\bar{F}(t-s) + F(t-s))$.

To precisely state our rare-event simulation algorithm, we need to specify the arrival process more exactly. Set $U_k = A_k - A_{k-1}$ for $k \geq 1$.

A5. $(U_k : k \geq 1)$ is iid, with $\kappa(\theta) := \log E \exp(\theta U_1)$ for $\theta \in \mathbb{R}$.

Our goal is to compute $\alpha = P(Q(t) > x)$ where $x \gg EQ(t)$.

Algorithm.

1. Compute the root θ^* to the equation

$$\frac{d}{d\theta} \int_0^t \kappa^{-1}(-\log(e^\theta \bar{F}(t-s) + F(t-s)))ds|_{\theta=\theta^*} = -x$$

and select m, the total number of replications.

2. Set $A \leftarrow 0$, $L \leftarrow 1$, $Q \leftarrow 0$, $W \leftarrow 0$.

3. Generate U from the distribution

$$\exp(\kappa^{-1}(-\log(e^{\theta^*}\bar{F}(t-A) + F(t-A)))x - \kappa(\kappa^{-1}(-\log(e^{\theta^*}\bar{F}(t-A) + F(t-A)))))P(U \in dx) = (e^{\theta^*}\bar{F}(t-A) + F(t-A)) \cdot \exp(\kappa^{-1}(-\log(e^{\theta^*}\bar{F}(t-A)+F(t-A)))x)P(U \in dx).$$

4. $L \leftarrow L \cdot (e^{\theta^*}\bar{F}(t-A) + F(t-A))^{-1} \cdot \exp(-\kappa^{-1}(-\log(e^{\theta^*}\bar{F}(t-A) + F(t-A)))U)$.

5. $A \leftarrow A + U$.

6. If $A > t$, go to 11.

7. Else, generate a Bernoulli rv I with parameter

$$\frac{e^{\theta^*}\bar{F}(t-A)}{e^{\theta^*}\bar{F}(t-A) + F(t-A)}.$$

8. $Q \leftarrow Q + I$.

9. $L \leftarrow L \cdot e^{-\theta^* I}(e^{\theta^*}\bar{F}(t-A)+F(t-A))$.

10. Go to 3.

11. $W \leftarrow I(Q>x)L$.

12. Replicate steps 2 through 11 n independent times, thereby computing $W_1, W_2, \ldots, W_n$.

13. The estimator for α is $n^{-1}\sum_{i=1}^{n} W_i$.

A natural question that arises here is the efficiency of the algorithm just described. As in Sections 2 and 3, we offer an asymptotic "heavy-traffic" analysis of the estimator above.

Suppose that we wish to compute $\alpha_n = P(Q_n(t) > nx)$, where $Q_n(t)$ is as described earlier. The arrival process for system n is accelerated by a factor of n, so that the j'th inter-arrival time in system n is just U_j/n. Thus, the logarithmic moment generating function $\kappa_n(\cdot)$ for the inter-arrival times in system n is given by $\kappa_n(\theta) = \log E\exp(\theta U_1/n) = \kappa(\theta/n)$. It is then easily verified that $\kappa_n^{-1}(\theta) = n\kappa^{-1}(\theta)$. Hence, the root θ^* of the equation

$$\begin{aligned}
&\frac{d}{d\theta}\int_0^t \kappa_n^{-1}(-\log(e^{\theta^*}\bar{F}(t-s)+F(t-s)))ds|_{\theta=\theta^*} \\
&\quad = n\frac{d}{d\theta}\int_0^t \kappa^{-1}(-\log(e^{\theta^*}\bar{F}(t-s)+F(t-s)))ds|_{\theta=\theta^*} \\
&\quad = -nx
\end{aligned}$$

appearing in step (1) of the algorithm is independent of n. Furthermore, the likelihood ratio of step (11) then is equal to

$$\begin{aligned}
L_n &= (e^{\theta^*}\bar{F}(t)+F(t))^{-1} \\
&\quad \cdot \exp(-n\sum_{j=0}^{N_n(t)} \kappa^{-1}(-\log(e^{\theta^*}\bar{F}(t-\frac{A_j}{n}) \\
&\quad + F(t-\frac{A_j}{n})))\frac{U_j}{n} - \theta^* Q_n(t)) \\
&= (e^{\theta^*}\bar{F}(t)+F(t))^{-1}\cdot\exp(-n\int_0^{n^{-1}A_{N_n(t)+1}} \\
&\kappa^{-1}(-\log(e^{\theta^*}\bar{F}(t-n^{-1}A_{N_n(s)})+F(t-A_{N_n(s)})))ds \\
&- \theta^* Q_n(t))
\end{aligned} \tag{3}$$

Let $E_*(\cdot)$ denote the expectation operator associated with the "change-of-measure" for system n.

Theorem 4. Suppose that U is a bounded rv, and that A1, A4, and A5 hold. Assume that there exists a root θ^* to the equation

$$\frac{d}{d\theta}\int_0^t \kappa^{-1}(-\log(e^{\theta^*}\bar{F}(t-s)+F(t-s)))|_{\theta=\theta^*} = -x$$

and that $r(\cdot)$ is continuously differentiable on $[0, t+1]$, where $r(s) = \kappa^{-1}(-\log(e^{\theta^*}\bar{F}(t-s)+F(t-s)))$. Then,

$$\begin{aligned}
&\frac{1}{n}\log E_* I(Q_n(t) > xn)L_n^2 \\
&\to -2\theta^* x - 2\int_0^t \kappa^{-1}(-\log(e^{\theta^*}\bar{F}(t-s)+F(t-s)))ds
\end{aligned}$$

as $n \to \infty$.

Proof. The key is formula (3) for the likelihood ratio. The exponent appearing in (3) is just

$$\begin{aligned}
&-n\int_0^{n^{-1}A_{N_n(t)+1}} r(n^{-1}A_{N_n(s)})ds - \theta^* Q_n(t) \\
&\quad = -n\int_0^t r(s)ds - \theta^* Q_n(t) \\
&\quad -n\sum_{j=0}^{N_n(t)}\int_{A_j/n}^{A_{j+1}/n}[r(n^{-1}A_{N_n(s)}) - r(s)]ds \\
&\quad \le -n\int_0^t r(s)ds - \theta^* Q_n(t) \\
&\quad +n\sup_{0\le u\le t+1}|r'(u)|\sum_{j=0}^{N_n(t)}\int_{A_j/n}^{A_{j+1}/n}(s-A_j/n)ds \\
&\quad \le -n\int_0^t r(s)ds - \theta^* Q_n(t) \\
&\quad +n\sup_{0\le u\le t+1}|r'(u)|\sum_{j=0}^{N_n(t)+1} U_j^2/n^2 \\
&\quad = -n\int_0^t r(s)ds - \theta^* Q_n(t) + O(1/n)
\end{aligned}$$

where the $O(1/n)$ term above is deterministic. It follows that

$$\begin{aligned}
&I(Q_n(t) > xn)L_n^2 \\
&\quad \le I(Q_n(t) > xn)(e^{\theta^*}\bar{F}(t)+F(t))^{-1} \\
&\qquad \exp(-2n\int_0^t r(s)ds - 2\theta^* Q_n(t) + O(1/n)) \\
&\quad \le \exp(-2n\int_0^t r(s)ds - 2\theta^* xn + O(1/n)).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\overline{\lim_{n\to\infty}}\frac{1}{n}\log E_* I(Q_n(t) > xn)L_n^2 \\
&\le -2\int_0^t \kappa^{-1}(-\log(e^{\theta^*}\bar{F}(t-s)+F(t-s)))ds - 2\theta^* x.
\end{aligned}$$

The "limit-infimum" result necessary to reach our desired limit follows from the same argument as in Section 3 (namely, $I(Q_n(t) > xn)L_n$ is unbiased for $P(Q_n(t) > xn)$, and the latter probability converges in logarithmic scale via the Gärtner-Ellis theorem).

Theorem 4 establishes that our algorithm produces estimates that are asymptotically optimal (in logarithmic scale).

5 A NUMERICAL EXAMPLE

In this section, we provide a numerical example to complement the theoretical developments of the previous sections. More precisely, we find $\alpha = P(Q(t) > x)$ via simulation for two different systems and for several different values of x. We first consider an $M/M/\infty$ system, and secondly a $G/M/\infty$ system with iid inter-arrival times distributed as an hyper-exponential (H_2) rv with density

$$f(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x},\ x \geq 0.$$

The H_2 distribution is the mixture of two exponential distributions, and for this reason it is useful when modeling the arrivals of two different classes of customers.

In these simulations, we choose $t = 500$, $\lambda = 1$ in the $M/M/\infty$ system; $\lambda_1 = 1$, $\lambda_2 = 2$, and $p = 0.8$ in the $H_2/M/\infty$ system; and $\mu = 0.01$ for both systems. The values of x considered are $x = 120, 130, 140$ for both systems, so that $\{Q(t) > x\}$ becomes a "rare-event" as x increases.

We obtain two estimators. The first one is $\tilde{\alpha}(n)$, the estimator obtained by conventional Monte Carlo simulation resulting from computing the sample mean formed from $n = 1000$ iid replications of the random variable $I(Q > x)$. Our second estimator is $\alpha(n)$, formed by computing the average of n iid replicates of the rv $I(Q > x)L$, where Q is obtained using the Algorithm described in the previous section.

In order to compare the efficiency of these estimators, we repeat $m = 1000$ times the simulation just described. The sample mean (sample standard deviation) over m of these estimators produces $\tilde{\alpha}(n,m)$ and $\alpha(n,m)$ ($\tilde{s}(m)$ and $s(m)$).

In Tables 1 and 2 we summarize our results. In each case we display $\tilde{\alpha}(n,m)$, $\alpha(n,m)$, $\tilde{s}(m)$, and $s(m)$. In addition, to validate our results we also include the true value of α; a known value in the $M/M/\infty$ setting, and obtained with a very long simulation in the $H_2/M/\infty$ case. To make more explicit the impact of our estimator, the last row in the tables shows the ratio of the estimator standard deviations $s(m)/\tilde{s}(m)$.

The conclusion we draw from these simulations is that our estimator becomes much more efficient than the conventional Monte Carlo estimator as the tail parameter x increases, for both the $M/M/\infty$ and the $H_2/M/\infty$ systems.

Table 1: $M/M/\infty$ Tail Probability Simulation

	Tail parameter x		
Parameter	120	130	140
α	0.0192	0.0014	4.77e-5
$\tilde{\alpha}(n,m)$	0.0184	0.0012	6.5e-5
$\alpha(n,m)$	0.0194	0.0015	4.75 e-5
$\tilde{s}(m)$	4.25e-3	1.16e-3	2.37e-4
$s(m)$	3.41e-3	3.26e-4	1.65e-5
$\tilde{s}(m)/s(m)$	1.25	3.56	14.4

Table 2: $H_2/M/\infty$ Tail Probability Simulation

	Tail parameter x		
Parameter	120	130	140
α	0.173	0.0334	0.0034
$\tilde{\alpha}(n,m)$	0.175	0.0345	0.0030
$\alpha(n,m)$	0.174	0.0337	0.0033
$\tilde{s}(m)$	0.0503	0.0269	6.7e-3
$s(m)$	0.0457	9.83e-3	1.23e-3
$\tilde{s}(m)/s(m)$	1.10	2.74	5.45

REFERENCES

Bucklew, J. A. 1990. *Large Deviations Techniques in Decision, Simulation, and Estimation.* New York: Wiley.

Dembo, A., and T. Zajic. 1995. Large Deviations: From Empirical Mean and Measure to Partial Sum Process. *Stoc. Proc. and Appl.* 57: 191-224.

Dembo, A., and O. Zeitouni. 1998. *Large Deviations Techniques and Applications.* New York: Springer-Verlag.

Glynn, P. W., and W. Whitt. 1994. Large Deviations Behavior of Counting Processes and their Inverses. *Queueing Systems* 17: 107-128.

Glynn, P. W. 1995. Large Deviations for the Infinite Server Queue in Heavy Traffic . *IMA Volume 71 in Mathematics and its Applications.* New York: Springer-Verlag 387-395.

Ross, S. H. 1983. *Stochastic Processes.* New York: Wiley.

AUTHOR BIOGRAPHIES

ROBERTO SZECHTMAN completed a Ph.D.in the Department of Management Science and Engineering at Stanford University in 2001. His research interests include applied probability, simulation theory, and supply chain management.

PETER W. GLYNN is professor of Management Science and Engineering, and was named to the Thomas W. Ford chair in 1996. He develops computational algorithms and simulation techniques for complex stochastic systems. Applications of his work include performance analysis for computer, telecommunications, and manufacturing systems.