Proceedings of the 2000 Winter Simulation Conference
J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.

SIMULATING GI/GI/1 QUEUES AND INSURANCE RISK PROCESSES WITH SUBEXPONENTIAL DISTRIBUTIONS

Nam Kyoo Boots

Department of Econometrics
and Operations Research
Vrije Universiteit
1081 HV Amsterdam, THE NETHERLANDS

Perwez Shahabuddin

Department of Industrial Engineering
and Operations Research
Columbia University
New York, NY 10027, U.S.A.

ABSTRACT

This paper deals with estimating small tail probabilities of the steady-state waiting time in a GI/GI/1 queue with heavy-tailed (subexponential) service times. The problem of estimating infinite horizon ruin probabilities in insurance risk processes with heavy-tailed claims can be transformed into the same framework. It is well-known that naive simulation is ineffective for estimating small probabilities and special fast simulation techniques like importance sampling, multilevel splitting, etc., have to be used. Previous fast simulation techniques for queues with subexponential service times have been confined to M/GI/1 queueing systems. The general approach is to use the Pollaczek-Khintchine transformation to transform the problem into that of estimating the tail distribution of a geometric sum of independent subexponential random variables. However, no such useful transformation exists when one goes from Poisson arrivals to general interarrival-time distributions. We describe an approach that is based on directly simulating the random walk associated with the waiting-time process of the GI/GI/1 queue, using a change of measure called delayed subexponential twisting – an importance sampling idea recently developed and found useful in the context of M/GI/1 heavy-tailed simulations.

1 INTRODUCTION

This paper deals with estimating tail probabilities of the steady-state waiting-time random variable in a GI/GI/1 queue with heavy-tailed service times. In particular we consider service times that are sub-exponentially distributed. If W is the steady-state waiting-time random variable, then the problem is to estimate $P(W > u)$ where u is large. Problems like these arise, for example, while estimating probabilities of extreme delays of packets in communication networks or the packet loss probabilities in such networks. While the queueing systems used to realistically model communication networks are much more complex than the GI/GI/1 queue, this work may be viewed as one of the first steps in that direction in the context of heavy-tailed service times (and related quantities that may be heavy-tailed, e.g., duration of packet transmission times).

The random walk associated with the waiting time in the GI/GI/1 queue has the same probabilistic structure as an insurance risk process where the claims arrive according to an ordinary renewal process (see, e.g., Embrechts and Klüppelberg (1993)). In particular, $P(W > u)$ in the context of a GI/GI/1 queue may be interpreted as the probability of ultimate ruin with initial capital u in the insurance risk process. In this setting, a subexponential claim-size distribution corresponds to a subexponential service-time distribution. Subexponential claim-size distributions are used to model the possibility of large claims.

A large body of work already exists for the rare event simulation of queues and networks of queues for the case where service times and related quantities are light-tailed (see, e.g., Cottrell, Fort and Malgouyres (1983), Parekh and Walrand (1989), Frater, Lenon and Anderson (1991), Sadowsky (1991), Chang, Heildelberger, Juneja and Shahabuddin (1994) and Falkner, Devetsikiotis and Lambadaris (1999); for a survey see Heidelberger (1995)). In this paper we call a distribution light-tailed if its moment generating function is finite in some neighborhood of zero. Importance sampling is a widely used technique in the setting of light-tailed random variables. It involves simulating the system with a new probability dynamics (i.e., a change of probability measure) that makes the rare event happen more frequently and then adjusting the final estimate. The change of probability measure frequently used in the light-tailed case is called "exponential change of measure" or "exponential twisting" (see, e.g., Siegmund (1976), Bucklew (1990), Asmussen (1985) and Lehtohnen and Nyrhinen (1992)). Let $f(\cdot)$ be the density function of a non-negative random variable X and let $M_X(\cdot)$ be its moment generating function. In a queue, the X may correspond to a service

time random variable or an interarrival-time random variable. Then the density obtained by exponentially twisting $f(x)$ by an amount θ is

$$f_\theta(x) \equiv \frac{e^{\theta x} f(x)}{M_X(\theta)}.$$

If the rare event of interest is facilitated by the X being large (cf., small) then one uses a θ that is positive (cf., negative) so that more large (cf., small) samples of X occur under the new measure. However, just arbitrarily choosing θ may result in highly unstable estimates, and large deviations theory has to be used to determine the optimal θ to be used in each case.

Recent data in the telecommunications area shows that very frequently quantities like service times (and related quantities) exhibit heavy-tailed behavior (see, e.g., Leland, Taqqu, Willinger and Wilson (1994)). Note that exponential twisting relies on the existence of the moment generating functions in a neigborhood of zero. When $f(x)$ is heavy-tailed then the moment generating function is infinite for all $\theta > 0$. Consequently most of the techniques and theory developed for rare event simulation in the light-tailed setting are not valid here.

One of the first works in the area of rare event simulation for systems with heavy-tailed random variables is Asmussen and Binswanger (1997). They considered the problem of estimating the probability of ruin for insurance-claim processes with Poisson claim arrivals and subexponentially distributed claim size. As mentioned before, it can easily be shown that this is equivalent to the problem of estimating the tail probability of the steady-state waiting time in a M/GI/1 queue with subexponential service times. They came up with an innovative algorithm based on conditioning and they proved that it works for subexponential service times with a regularly-varying tail. Later, Asmussen, Binswanger and Hojgaard (1998) gave an importance sampling change of measure for the same problem that works for other subexponential distributions, but only if the traffic intensity is below a certain level. A different framework for importance sampling for systems with subexponential distributions was presented in Juneja and Shahabuddin (1999). The idea was "subexponential twisting", i.e., twist at a "subexponential rate" rather than at an exponential rate as is done in exponential twisting. One way of doing subexponential twisting is "hazard rate twisting". Let $\lambda(x) \equiv f(x)/\bar{F}(x)$ be the hazard rate corresponding to $f(x)$ and let $\Lambda(x) = \int_{s=0}^{x} \lambda(s)ds$ be the cumulant function. Note that the tail of any distribution, $\bar{F}(x)$, may be represented as $e^{-\Lambda(x)}$. In hazard rate twisting, the new distribution function is given by

$$\bar{F}_\theta(x) = e^{-\Lambda(x)(1-\theta)} \tag{1}$$

where $0 \leq \theta < 1$. As was the case for exponential twisting, an appropriate θ has to be chosen for the given application. In Juneja and Shahabuddin (1999) it was formally shown that a "delayed" version of hazard rate twisting is efficient for the case of estimating $P(W > u)$ in M/GI/1 queues for all traffic intensities (provided the queue is stable) and for almost all subexponential distributions. Independently of Juneja and Shahabuddin (1999), Asmussen, Binswanger and Hojgaard (2000) gave a refinement of the importance sampling algorithm in Asmussen, Binswanger and Hojgaard (1998) that is also provably efficient for all traffic intensities.

All the above techniques relied on the Pollaczek-Khintchine transformation to simulate the M/GI/1 queue. Using this transformation one can express $P(W > u)$ as $P(\sum_{i=1}^{N} Y_i > u)$ where the Y_i's are independent and have the integrated-tail distribution of the service times (explained later), and N is a geometric random variable with parameter ρ, where ρ is the traffic intensity (i.e., the ratio of the expected service time to the expected interarrival time), and N is independent of the Y_i's. In the importance sampling techniques in Asmussen, Binswanger and Hojgaard (2000) and Juneja and Shahabuddin (1999), the "new" distribution is chosen for the Y_i's; the distribution of the N is left unchanged. However, once we go from Poisson arrivals to general interarrival times the distributions of the N and the Y_i's are no longer known in explicit form.

In this paper we attempt to go beyond the restriction imposed by the Pollaczek-Khintchine transformation, and simulate the random walk associated with the GI/GI/1 queue directly using delayed subexponential twisting. Results are mixed; the method works well for some classes of subexponential distributions and not for others. Before we discuss the formal effciency of these methods, we will review the standard criterion used in the simulation literature to evaluate the efficiency of rare event simulation techniques, and a slightly weaker one which we developed. Techniques satisfying the weaker criterion are as good for most practical purposes as the techniques satisfying the usual one.

The standard criterion used to evaluate the efficiency of rare event simulation techniques is "asymptotic optimality" (see, e.g., Heidelberger (1995); sometimes also called "asymptotic efficiency"). Many of the light-tailed simulation techniques and the three heavy-tailed simulation techniques mentioned have been shown to be "asymptotically optimal" under certain assumptions. However, in our experience and in the experience of others (see, e.g., Asmussen, Binswanger and Hojgaard (2000), p. 315) it is difficult to come up with techniques that satisfy this criterion in the heavy-tailed setting beyond the M/GI/1 queue. So instead we settle for something weaker that we call "large set asymptotic optimality." The new criterion is based on the observation that many times the reason why importance sampling does not work well is that the likelihood-ratio on some "small" set (i.e., note that "small" here is in comparison with the rare

set, the probability of which we are trying to estimate) is highly variable; if we exclude this set when we conduct importance sampling, then one gets very good estimates for the remaining "large" part. Now in most simulation experiments in practice, one tries for a fixed relative error (the confidence interval half-width upon the probability one is trying to estimate) of say δ' (usually somewhere between 0.01 and 0.1). And the δ' is usually independent of the rarity of the overall event (i.e., whether one is estimating a probablity of 10^{-2} or 10^{-9} one attempts to achieve the same relative error). If the relative bias, i.e., the ratio of the "small" set probability to the probability of interest is of the same order as δ' (and remains so as the event of interest becomes rarer), then we are not losing much from the practical point of view when we exclude the small set.

Roughly speaking, the class of subexponential distributions most commonly used in practice can be categorized into the following three classes: "Weibull type tails", "lognormal type tails" and "Pareto type tails"; a more formal categorization will be given later on. These are tails with different degrees of "heaviness" ranging from least heavy to most heavy. We show that for the class of subexponential distributions with Weibull type tails we obtain large set asymptotic optimality. For the class of distributions with lognormal type tails, we conjecture large set asymptotic optimality but it is very difficult to formally prove it. For the Pareto type tails we feel that this technique is not large set asymptotically optimal and hence is not recommended for use in this setting. Fortunately, being the class with the heaviest tails, the asympotic approximations for $P(W > u)$ given by heavy-tailed theory are the most accurate here and fairly close to $P(W > u)$.

Section 2 reviews the random walk formulation for estimating $P(W > u)$ in the GI/GI/1 queue and discusses the basic concepts in the theory of subexponential distributions. Section 3 reviews rare event simulation and importance sampling. We also introduce the concept of large set asymptotic optimality in this section. Section 4 presents the simulation algorithm and conditions on the parameters of the service time distribution and the simulation algorithm that guarantee large set asymptotic optimality. Experimental results are presented in Section 5.

2 PRELIMINARIES AND RELATED RESULTS

We start with some commonly used notation. For any functions $z_1(x)$ and $z_2(x)$, we use the notation $z_1(x) \sim z_2(x)$ to mean that $z_1(x)/z_2(x)$ converges to 1 as x goes to infinity. Order statistics of $X_1, \ldots, X_n$ are denoted by $X_{(1)} \leq \cdots \leq X_{(n)}$. The maximum of zero and x is denoted by $\{x\}^+$. Finally, the indicator function is denoted by $I(\cdot)$.

2.1 The Model

Let F be the cumulative distribution function of the service-time random variable X. We assume that F has a density f. Let $\lambda(x) \equiv f(x)/\bar{F}(x)$ be the hazard-rate function and $\Lambda(x) = \int_{s=0}^{x} \lambda(s)ds$ the cumulant function. It is well-known that $\Lambda(x) = -\log \bar{F}(x)$. We assume that the 0th customer arrives at epoch 0 to an empty *system* and hence has a waiting time in the *queue* $W_0 = 0$. Let $(\xi_n)_{n\geq 0}$ be the sequence of i.i.d. interarrival times and $(X_n)_{n\geq 0}$ be the sequence of i.i.d. service times, i.e., X_n is the service time of the n-th customer and ξ_n the time between the arrival of customer n and $n+1$. We assume the traffic intensity $\rho = E[X]/E[\xi]$ to be smaller than 1 and the sequence of interarrival times to be independent of the sequence of service times. An insightful recursion for the waiting time can be derived; if W_n denotes the waiting time of the n-th customer, then it is well-known that W_n satisfies the so-called Lindley's recursion $W_{n+1} = \{W_n + X_n - \xi_n\}^+$, $n \geq 0$ (see, e.g., Feller (1966)). Expanding this relation recursively gives

$$W_{n+1} = \max\left\{\sum_{i=0}^{n}(X_i - \xi_i), \ldots, \sum_{i=n-1}^{n}(X_i - \xi_i), X_n - \xi_n, 0\right\}. \quad (2)$$

Define the random walk $(M_n)_{n\geq 1}$ by

$$M_n = \sum_{i=0}^{n-1}(X_i - \xi_i), \quad (3)$$

with i.i.d. increments $X_i - \xi_i$ and let $M_0 \equiv 0$. It is easy to see from (2) and (3) that W_n has the same distribution as $\max_{0\leq i\leq n} M_i$. Thus the steady-state waiting time W has the same distribution as $\sup_{n\geq 0} M_n$. In this paper, we assume the interarrival-time distribution to be light-tailed with a finite mean and we simulate for $P(W > u)$, for large u, via the random variable $I(\sup_{n\geq 0} M_n \geq u)$. Let

$$\tau(u) = \inf\{n : n \in \mathbb{N}, M_n \geq u\},$$

be the *hitting time* of level u. Note that $\tau(u)$ is an $\{\infty\}\cup\mathbb{N}$-valued random variable and $P(W > u) = P(\tau(u) < \infty)$.

2.2 Subexponential Distributions and GI/GI/1 Queue Asymptotics

For details about subexponential distributions we refer the reader to the textbook Embrechts, Klüppelberg and Mikosch (1997). Below we give a short summary.

The definition of subexponentiality is due to Chistyakov (1964):

Definition 2.1 *The distribution F is subexponential (denoted by $F \in \mathcal{S}$) if and only if*

$$\frac{P(X_1 + \cdots + X_n > u)}{nP(X_1 > u)} \to 1 \ (u \to \infty), \qquad (4)$$

for all n.

The integrated tail of F is defined by $F_I(x) = \int_0^x \overline{F}(y)dy/E[X]$ when $E[X] < \infty$. Define $\Lambda_I(u)$ and $\lambda_I(u)$ similar to $\Lambda(u)$ and $\lambda(u)$. In this paper F_I is assumed to be subexponential, rather than F. Since the most interesting distributions which are subexponential have integrated tails that are also subexponential and vice versa (this is certainly the case for the ones we use in this paper; see also Embrechts, Klüppelberg and Mikosch (1997)), we continue using the phrase "subexponential service times".

For the GI/GI/1 queue with subexponential service times, the asymptotic waiting-time distribution is given by Pakes (1975):

$$P(W > u) \sim \frac{\rho}{1-\rho}\overline{F}_I(u). \qquad (5)$$

Note that in the asymptotics of the waiting-time distribution, the interarrival-time distribution plays a role only via its first moment. In this paper we use the following assumption for the service times:

Assumption 1 *$F_I \in \mathcal{S}$ and F is in the maximum domain of attraction of the Gumbel distribution.*

This implies that $\max_n X_n$ converges, when properly normalized, to the Gumbel distribution. This is a result from extreme value theory. A function that plays an important role in extreme value theory is the so-called *auxiliary function* $a(u)$. The function $a(u)$ is defined to be any function such that

$$a(u) \sim \frac{\int_u^\infty \overline{F}(x)dx}{\overline{F}(u)} = E[X]\frac{\overline{F}_I(u)}{\overline{F}(u)}.$$

For details we refer the reader to Goldie and Resnick (1988), Asmussen and Klüppelberg (1996), and Embrechts, Klüppelberg and Mikosch (1997). Examples of subexponential distributions that satisfy Assumption 1 are:

- The heavy-tailed Weibull(σ, α) distribution with

  $$F(x) = 1 - e^{-\sigma x^\alpha}, \ \ f(x) = \sigma\alpha x^{\alpha-1}e^{-\sigma x^\alpha},$$

  with $\sigma > 0$ and $0 < \alpha < 1$. In this case we may take

  $$a(u) = \frac{1}{\alpha}u^{1-\alpha}.$$

- The lognormal(α, σ^2) distribution with

  $$F(x) = \Phi\left(\frac{\log x - \alpha}{\sigma}\right)$$

  and

  $$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left[\frac{\log x-\alpha}{\sigma}\right]^2}$$

  with $\alpha \in \mathbb{R}$ and $\sigma > 0$ and where Φ denotes the standard normal distribution function. The mean of the lognormal distribution is given by $e^{\alpha+\frac{1}{2}\sigma^2}$. As auxiliary function we may take

  $$a(u) = \frac{\sigma^2 u}{\log u - \alpha}.$$

The technique in this paper relies heavily on a result in Asmussen and Klüppelberg (1996). Define a conditional distribution $P^{(u)}$ of (M_n) by

$$P^{(u)}(\cdot) = P(\cdot \mid \tau(u) < \infty).$$

In case Assumption 1 holds, the asymptotic distribution of the normalized hitting time τ under the $P^{(u)}$-measure is derived in Asmussen and Klüppelberg (1996): $\tau(u)/a(u)$ asymptotically has an exponential distribution. In particular, if $\overset{P^{(u)}}{\to}$ denotes convergence in the conditional distribution, then

$$\frac{\tau(u)}{a(u)} \overset{P^{(u)}}{\to} \frac{\psi}{\mu}, \qquad (6)$$

where ψ is a standard exponential random variable (i.e., with mean 1) and

$$\mu = E[X]\frac{1-\rho}{\rho}.$$

In this paper we will also need the following condition that is satisfied by most of the common subexponential distributions; distributions not satisfying it are mainly pathological cases (see Juneja and Shahabuddin (1999) for a discussion):

Assumption 2 *The hazard-rate function $\lambda(x)$ is eventually decreasing.*

3 RARE EVENT SIMULATION AND IMPORTANCE SAMPLING

3.1 A New Criterion for Rare Event Simulation Efficiency

Let $A(u)$ denote some event parameterized by u with the property that $P(A(u)) \to 0$ as $u \to \infty$. The u is called the rarity parameter. Define $\alpha(u) := P(A(u))$ and let $\hat{\alpha}(u)$

denote an unbiased estimator for $\alpha(u)$, which is obtained by averaging realizations from n i.i.d. simulation replications. If we let $\widehat{\text{Var}}(\hat{\alpha}(u))$ be the sample estimator of the variance of $\hat{\alpha}(u)$, then an $100(1-\eta)\%$ confidence interval based on the central limit theorem is given by

$$\left(\hat{\alpha}_u - \sqrt{\widehat{\text{Var}}(\hat{\alpha}(u))}z_{1-\eta/2}, \hat{\alpha}_u + \sqrt{\widehat{\text{Var}}(\hat{\alpha}(u))}z_{1-\eta/2}\right),$$

where z_a denotes the a-th quantile of the standard normal distribution. A quantity that is a measure of the precision of an estimator is the relative error, which is defined to be the confidence interval half-width upon the quantity one is trying to estimate, i.e.,

$$RE\left[\hat{\alpha}(u)\right] := z_{1-\eta/2}\frac{\sqrt{\text{Var}(\hat{\alpha}(u))}}{\alpha(u)}.$$

The estimator $\hat{\alpha}(u)$ is said to have a bounded relative error, if for fixed "n" the relative error remains bounded as u tends to infinity (Shahabuddin, 1994). Alternatively, the number of samples required to obtain a given relative error remains bounded as u goes to infinity. Since rare event simulation techniques with bounded relative errors are usually very hard to find, in the literature one works with the somewhat weaker notion of *asymptotic optimality* (a.o.).

Definition 3.1 "Asymptotically optimal" *$\hat{\alpha}(u)$ is an asymptotically optimal estimator of $\alpha(u)$ if*

$$\limsup_{u\to\infty}\frac{\log\left(\text{Var}\left[\hat{\alpha}(u)\right]\right)}{\log(\alpha^2(u))} \geq 1. \tag{7}$$

Note that a.o. allows the relative error to grow to infinity for growing u, but that this growth is at a slower rate (compared to the decay rate of $\alpha(u)$).

In many cases the simulation effort per replication is either independent of the rarity parameter u or grows very weakly with it (see, e.g., Shahabuddin (1994) and Juneja and Shahabuddin (1999)). However in cases where the growth of effort is substantial with increasing u (see, e.g., Glasserman, Heidelberger, Shahabuddin and Zajic (1999) and this paper) it is more fair to consider the criterion

$$\limsup_{u\to\infty}\frac{\log\left(\text{work}(u)\times\text{Var}\left[\hat{\alpha}(u)\right]\right)}{\log(\alpha^2(u))} \geq 1 \tag{8}$$

(see, e.g., Glynn and Whitt (1992)). Here work(u) denotes the computation effort per simulation replication as a function of u. If $\hat{\alpha}(u)$ satisfies (8), then it is called *work-normalized a.o.*. As mentioned in the Introduction, we have not been able to find a work-normalized a.o. simulation algorithm for the GI/GI/1 case and hence we introduce the weaker criterion *work-normalized large set a.o.*, and prove that it is satisfied under certain conditions.

In the following definition, think of δ as the maximum *asymptotic relative bias* that one is willing to tolerate in the simulation.

Definition 3.2 "Large set asymptotically optimal" *Let $\delta \in (0,1)$ be a fixed constant. If*

1. *there exists a decomposition of $\alpha(u)$ into two positive quantities $\alpha(u) = \gamma(u) + \epsilon(u)$ s.t.*

$$\limsup_{u\to\infty}\frac{\epsilon(u)}{\alpha(u)} \leq \delta,$$

2. *there exists an unbiased estimator $\hat{\gamma}(u)$ of $\gamma(u)$ s.t.*

$$\limsup_{u\to\infty}\frac{\log\left(\text{Var}\left[\hat{\gamma}(u)\right]\right)}{\log(\gamma^2(u))} \geq 1, \tag{9}$$

then $\hat{\gamma}(u)$ is said to be a large set a.o. estimator of $\alpha(u)$.

Similar to work-normalized a.o., we can define work-normalized large set a.o.

Let $\alpha_a(u)$ be an *asymptotic approximation* to $\alpha(u)$, i.e., $\alpha_a(u) \sim \alpha(u)$. Since $\alpha_a(u)$ may be regarded as an estimator with zero variance, it can be checked that it is also large set a.o. Unlike the approximations in the light-tailed setting which are asymptotic in the log (i.e., $\log\alpha_a(u) \sim \log\alpha(u)$), in the heavy-tailed setting approximations that satisfy $\alpha_a(u) \sim \alpha(u)$ do exist and hence are competitive with large set a.o. rare event simulation methods. We now briefly discuss the advantage and disadvantage of each.

Even if we come up with a.o. simulation methods (in contrast to large set a.o. simulation methods) for the heavy-tailed case, asymptotic approximations have relative errors going to zero, whereas a.o. is weaker than bounded relative error in the simulation. Also approximations take negligible computation time as compared to simulation. So the only advantage of simulation methods is for u fixed (say at u_0) and in the "practical range" (in contrast to $u\to\infty$). Then the relative bias in the asymptotic approximations, i.e., $(\alpha_a(u_0)-\alpha(u_0))/\alpha(u_0)$ is also fixed and beyond our control. However, in simulation one has the choice of decreasing the relative error by running more simulations (i.e., putting in more effort). In this practical range where asymptotic approximations are not accurate, it is still worthwhile to come up with a.o. simulation techniques if they improve considerably over naive ones. As mentioned before, this has been done for certain cases in Asmussen and Binswanger (1997), Asmussen, Binswanger and Hojgaard (2000) and Juneja and Shahabuddin (1999).

One would prefer this to be the case for large set a.o. techniques also. However, in the definition of large set a.o., one can also think of $\epsilon(u_0)$ as a bias term over which one has no control. So on top of Definition 3.2, we place another stringent requirement of having an additional parameter β

in the decomposition that gives control over such bias terms for fixed u.

Condition 3.3 Additional condition for definition of large set asymptotic optimality: *For any fixed u, there exists a family of decompositions parameterized by β (i.e., $\alpha(u) = \gamma_\beta(u) + \epsilon_\beta(u)$) such that:*

$$\limsup_{\beta\to\infty} \frac{\epsilon_\beta(u)}{\alpha(u)} = 0.$$

With this new additional condition, asymptotic approximations are no longer work-normalized large set a.o. To simplify notation, we will use $\gamma(u) \equiv \gamma_\beta(u)$ and $\epsilon(u) \equiv \epsilon_\beta(u)$.

3.2 Importance Sampling

The simulation method we use in this paper is importance sampling. Suppose the stochastic process that we wish to simulate is defined on some probability space with measure P. Let Q be some other measure on the same probability space such that P is absolutely continuous relative to Q. One can then express

$$\alpha(u) = E_Q\left(I(A(u))\frac{dP}{dQ}\right),$$

where dP/dQ is called the likelihood-ratio and subscript Q indicates that the expectation is with respect to the new measure Q. In importance sampling one generates the sample paths under the Q measure, computes the likelihood-ratio in each case and estimates $\alpha(u)$ by the sample mean of the $I(A(u))(dP/dQ)$'s. The underlying idea is that the event that is rare under P is not rare under Q and in order to get an unbiased estimator we have to multiply the estimator by some correction factor, which turns out to be the likelihood-ratio.

As mentioned in Section 1, for subexponential distributions one may use hazard rate twisting (HR) where the new distribution is given by (1). The density corresponding to $\overline{F}_\theta$ is given by

$$f_\theta(x) = (1-\theta)\lambda(x)e^{-(1-\theta)\Lambda(x)}. \tag{10}$$

For X_1 with density f, HR leads to a likelihood-ratio of $f(X_1)/f_\theta(X_1)$ and an unbiased estimator for $P(X_1+\cdots+X_n>)$ is given by

$$\begin{aligned}&\prod_{i=1}^{n}\frac{f(X_i)}{f_\theta(X_i)}I(X_1+\cdots+X_n>u)\\ =\;&(1-\theta)^{-n}e^{-\theta\sum_{i=1}^{n}\Lambda(X_i)}I(X_1+\cdots+X_n>u).\end{aligned}$$

Under some mild regularity conditions, for the choice of

$$\theta \equiv \theta_u = 1 - \frac{c}{\Lambda(u)}, \tag{11}$$

where c is any positive constant, HR is proved to be a.o. for estimating $P(X_1+\cdots+X_n>u)$ in Juneja and Shahabuddin (1999).

Weighted delayed hazard rate twisting (WDHR) extends HR by introducing a weighting parameter w and a delaying parameter $x^\star$. The WDHR density is defined by

$$f_{\theta_u,x^\star}(x) = \begin{cases} \frac{f(x)}{1+w} \text{ for } x \le x^\star, \\ \left(1-\frac{F(x^\star)}{1+w}\right)\frac{f_{\theta_u}(x)}{\overline{F}_{\theta_u}(x^\star)} \text{ for } x > x^\star. \end{cases} \tag{12}$$

If we let N be a geometrically distributed random variable with $P(N=n) = \rho^n(1-\rho)$ for $n \ge 0$, then it is well-known for the M/GI/1 queue that (see, e.g., Feller (1966))

$$P(W>u) = P(Y_1+\cdots+Y_N \ge u), \tag{13}$$

where the sequence of i.i.d. random variables (Y_i) are distributed as the integrated tail of the service-time distribution. In Juneja and Shahabuddin (1999) it is proved that for θ_u given by (11) and for certain choices of $x^\star \equiv x_u^\star$ and w (and under some mild regularity conditions), WDHR is a.o. for estimating $P(Y_1+\cdots+Y_N \ge u)$. Unfortunately, these results cannot be applied to the GI/GI/1 queue, since for non-Poisson arrivals, the Y_i's no longer have the integrated-tail distribution of the service times, but another distribution for which no explicit form is known in general. Besides, $P(N=n) = \hat{\rho}^n(1-\hat{\rho})$ for $n \ge 0$ and some $\hat{\rho}$ for which again no explicit expression is known. The techniques in Asmussen and Binswanger (1997) and Asmussen, Binswanger and Hojgaard (2000) also rely on (13) and hence are only applicable for M/GI/1 queues.

4 THE SIMULATION ALGORITHM

For the GI/GI/1 case, instead of using (13), we simulate the waiting-time distribution by directly simulating the random walk (M_n) defined in Section 2.1. We use WDHR for the service times, i.e., use the density $f_{\theta_u,x_u^\star}(x)$ with some specified θ_u and $x_u^\star$, to simulate the service times. We do not apply any change of measure to the interarrival-time distribution. This requires some stringent conditions on the choice of $x_u^\star$ and unlike the case in Juneja and Shahabuddin (1999), requires w to depend on u.

For any preselected asymptotic relative bias δ, we will use the decomposition

$$\alpha(u) = P(\tau(u) \le k_0(u)) + P(\tau(u) > k_0(u)),$$

where

$$k_0(u) = -\frac{a(u)\log\delta}{\mu} = -\frac{\rho a(u)\log\delta}{(1-\rho)E[X]}. \qquad (14)$$

Using (6), it is easy to check that

$$\frac{P(\tau(u) \le k_0(u))}{\alpha(u)} \to 1-\delta$$

as $u \to \infty$. We will now show that $P(\tau(u) \le k_0(u))$ may be estimated (work-normalized) a.o. using WDHR, thus giving a (work-normalized) large set a.o. estimator for $\alpha(u)$. Also note that selecting

$$k_0(u) = -\frac{\beta a(u)\log\delta}{\mu} = -\frac{\beta\rho a(u)\log\delta}{(1-\rho)E[X]} \qquad (15)$$

gives us the flexibility required to fulfill Condition 3.3. However, for ease of presentation we will use $\beta = 1$.

An important question in using WDHR is the choice of the importance sampling parameters θ_u, w_u and $x_u^\star$. For reasons similar to those in Juneja and Shahabuddin (1999), we use θ_u given by the equation

$$\theta_u = 1 - \frac{1}{\Lambda(u)}. \qquad (16)$$

Furthermore, we use $w = w_u$ given by

$$w_u = \frac{c_1\mu}{a(u)}, \qquad (17)$$

where $0 < c_1 < 1$ is some constant.

We will need F to satisfy the following assumption.

Assumption 3 *The $F(\cdot)$ is such that there exists some positive constant b satisfying*

$$\lim_{u\to\infty} \frac{\Lambda(u)^{-b+1}}{w_u} = 0.$$

(For instance, for the Weibull service times with $F(x) = 1 - e^{x^\alpha}$, Assumption 3 holds with $1/\alpha < b$.) We then use $x_u^\star$ satisfying

$$\Lambda(x_u^\star) = b\log\Lambda(u), \qquad (18)$$

where b is the constant in Assumption 3. Note that $x_u^\star$ goes to infinity as u goes to infinity.

Finally, we will also need the following assumption, which is satisfied by the commonly used subexponential distributions that are in the maximum domain of attraction of the Gumbel distribution, like the Weibull and the lognormal distribution.

Assumption 4 *The $F(\cdot)$ has an auxiliary function $a(u)$ such that*

$$\frac{a(u)x_u^\star}{u} \to 0 \ (u \to \infty).$$

The algorithm for estimating $P(W > u)$, using the above given values of θ_u, w_u and x_u^* is as follows:

Algorithm 4.1 "Weighted delayed hazard rate twisting of the service times"

1. *Draw i.i.d. samples $\xi_0, \ldots, \xi_k$ from the interarrival time distribution and i.i.d. samples $X_0, \ldots, X_k$ using the density $f_{\theta_u, x_u^\star}(x)$, where k is the minimum of $k_0(u)$ and* $\inf\left\{i : \sum_{j=0}^{i}(X_j - \xi_j) > u\right\}$.
2. *Define Z by*

$$Z = I\left(\sum_{j=0}^{k}(X_j - \xi_j) > u\right)\prod_{i=0}^{k}\frac{f(X_i)}{f_{\theta_u, x_u^\star}(X_i)}.$$

3. *An average of many independent samples of Z is an unbiased estimator for $P(\tau(u) \le k_0(u))$ which is used as an estimator for $P(W > u)$.*

Under the given assumptions one can prove the following theorem.

Theorem 4.2 *Algorithm 4.1 results in a work-normalized large set asymptotically optimal estimator for $P(W > u)$ with $\gamma(u) = P(\tau(u) \le k_0(u))$, $\epsilon(u) = P(\tau(u) > k_0(u))$.*

For the proof of the theorem we refer the reader to Boots and Shahabuddin (2000).

5 EXPERIMENTAL RESULTS

In this section we present some experimental results using Algorithm 4.1 (A4.1). We present results only for the M/GI/1 queue, since for this case we can compare with accurate simulation estimates based on the Pollaczek-Khintchine transformation (P-K) from Juneja and Shahabuddin (1999). We will also compare each case with the best known asymptotic approximation (AA) for $P(W > u)$ given by (5).

We first use service times that are Weibull distributed. It can easily be checked that this class of distributions satisfies the assumptions in Section 4. For the experiments we assume the service times to have the distribution $1-e^{-\sqrt{x}}$. We abort a simulation, if more than $k_0(u) = \max\{-a(u)\log\delta/\mu, 50\}$ customers have arrived for $\delta = 0.001$. We use $b = 2.1$, consistent with Assumption 3.

The values of the other parameters used by the algorithm are given in Table 1. They were determined using a heuristic approach further detailed in Boots and Shahabuddin (2000). Note that for the general subexponential Weibull distribution

Table 1: Values of the Parameters

u	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
100	$w_u = .1693, c_1 = .56$	$w_u = .0503, c_1 = .50$	$w_u = 0.0135, c_1 = .41$
	$x_u^\star = 23.38$	$x_u^\star = 23.38$	$x_u^\star = 23.38$
200	$w_u = .1185, c_1 = .56$	$w_u = .0364, c_1 = .51$	$w_u = 0.0105, c_1 = .45$
	$x_u^\star = 30.95$	$x_u^\star = 30.95$	$x_u^\star = 30.95$
400	$w_u = .0827, c_1 = .55$	$w_u = .0261, c_1 = .52$	$w_u = 0.0079, c_1 = .47$
	$x_u^\star = 39.58$	$x_u^\star = 39.58$	$x_u^\star = 39.58$
800	$w_u = .058, c_1 = .55$	$w_u = .0186, c_1 = .53$	$w_u = 0.0058, c_1 = .49$
	$x_u^\star = 49.26$	$x_u^\star = 49.26$	$x_u^\star = 49.26$

(i.e., $\alpha \neq 1/2$), it is difficult to compute the integrated-tail distribution. This indicates that even for the M/GI/1 case, Algorithm 4.1 is easier to implement than the one in Juneja and Shahabuddin (1999), for service-time distributions for which the integrated-tail distribution is difficult to compute. However, it is usually far less efficient in terms of simulation time.

The results from Juneja and Shahabuddin (1999) were based on 10,000,000 replications, in order to get accurate estimates for comparison purposes. For A4.1, we use 300,000 replications for each simulation. The percentages after the estimates are the relative half-widths of the 99%-confidence intervals, i.e., the relative error of the estimate. The standard effort of any simulation algorithm is defined as the variance per simulation replication times the CPU time per simulation replication. The numbers in the parenthesis besides the A4.1 estimator denote the *efficiency ratio* which is the ratio of the standard effort of naive simulation and the standard effort of A4.1. For naive simulation the standard effort is estimated by using the estimate of $P(W > u)$ from A4.1 and then using the formula $P(W > u)(1 - P(W > u))$ for the variance per replication; for the CPU time per replication we simulate the random walk up to k_0 (as otherwise there is a positive probability that the simulation may never end) without using importance sampling. The efficiency ratio may be interpreted as the number of times longer naive simulation will need to run to achieve the same relative accuracy as simulation with the new algorithm. We have not given any *performance* comparison with the algorithm in Juneja and Shahabuddin (1999), as that algorithm can only be used for the special case of M/GI/1 systems. The number in the parenthesis besides the AA denote the relative bias of AA, i.e., $100\% \times |\hat{\alpha}(u) - \alpha_a(u)|/\hat{\alpha}(u)$, where $\hat{\alpha}(u)$ is the accurate simulation estimate from Juneja and Shahabuddin (1999). Estimates in Table 2 for high values of ρ are not accurate for low u and the given number of simulation replications. This is also the case for P-K and $u = 400$. However, for large u the asymptotics take effect and the accuracy improves. From Table 2 we also see that for the given choice of run-lengths, AA is outperformed. Also there is no way to change the relative bias of AA; in contrast one can increase k_0 (to decrease the relative bias) and run more simulation replications to improve the simulation estimate.

We also conducted experiments using lognormal service times that are reported in Boots and Shahabuddin (2000). It can be checked that this family of distributions does not satisfy Assumption 3 . However, we still conjecture in Boots and Shahabuddin (2000) that the algorithm is (work-normalized) large set a.o. and give an intuitive argument in its support. The experimental results in Boots and Shahabuddin (2000) also support that claim.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation (U.S.A.) Career Award Grant DMI 96-25297. The work was partially performed while the second author was a guest of the Tinbergen Institute at Vrije University. The authors would also like to thank S. Juneja and A. Ridder for helpful discussion and comments.

REFERENCES

Asmussen, S. 1985. Conjugate processes and the simulation of ruin problems. *Stochastic Processes and their Applications* 20: 213-229.

Asmussen, S., and K. Binswanger. 1997. Simulation of ruin probabilities for subexponential claims. *ASTIN Bulletin* 27 (2): 297-318.

Asmussen, S., K. Binswanger and B. Hojgaard. 1998. Rare event simulation for heavy-tailed distributions. Research Report, Dept. of Mathematical Statistics, Lund University, Box 118, SE-22100 Lund, Sweden.

Asmussen, S., K. Binswanger and B. Hojgaard. 2000. Rare event simulation for heavy-tailed distributions. *Bernoulli* 6 (2): 303-322.

Asmussen, S., and C. Klüppelberg. 1996. Large deviations results for subexponential tails, with applications to insurance risk. *Stochastic Processes and their Applications* 64: 103-125.

Table 2: Estimates of $P(W > u)$ for the M/GI/1 Queue with Weibull$(1, 1/2)$ Service Times

u		$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
100	A4.1	$2.31E-4 \pm 1.7\%$ $(2.4E2)$	$1.38E-3 \pm 3.4\%$ (3.8)	$1.68E-2 \pm 10.8\%$ (0.07)
	P-K	$2.30E-4 \pm 1.3\%$	$1.41E-3 \pm 1.3\%$	$1.89E-2 \pm .67\%$
	AA	$1.17E-4$(49.1%)	$5.00E-4$(64.5%)	$1.50E3$(92.1%)
200	A4.1	$4.71E-6 \pm 2.0\%$ $(5.9E3)$	$2.46E-5 \pm 3.5\%$ $(1.5E2)$	$6.41E-4 \pm 38.8\%$ (.46)
	P-K	$4.61E-6 \pm 1.5\%$	$2.55E-5 \pm 3.15\%$	$7.37E-4 \pm 3.3\%$
	AA	$3.64E-6$(21.0%)	$1.09E-5$(57.3%)	$3.28E-5$(95.5%)
400	A4.1	$1.65E-8 \pm 2.5\%$ $(9.2E5)$	$7.12E-8 \pm 3.1\%$ $(1.3E5)$	$1.53E-6 \pm 69.5\%$ (13.7)
	P-K	$1.66E-8 \pm 1.6\%$	$7.11E-8 \pm 2.75\%$	$1.62E-6 \pm 43.3\%$
	AA	$1.44E-8$(13.3%)	$4.33E-8$(39.1%)	$1.30E-7$(92.0%)
800	A4.1	$5.54E-12 \pm 3.0\%$ $(1.8E9)$	$2.04E-11 \pm 3.1\%$, $(4.5E8)$	$1.27E-10 \pm 9.4\%$ $(1.5E6)$
	P-K	$5.45E-12 \pm 2.0\%$	$2.04E-11 \pm 1.8\%$,	$1.36E-10 \pm 7.9\%$
	AA	$5.08E-12$(6.8%)	$1.52E-12$(25.5%)	$4.57E-11$(66.4%)

Boots, N. K., and P. Shahabuddin. 2000. Simulating tail probabilities in GI/GI/1 queues and insurance risk processes with subexponential distributions. Research Report, Dept. of Industrial Engineering and Operations Research, Columbia University, NY 10027.

Bucklew, J. A. 1990. *Large Deviations techniques in decision, simulation, and estimation*. John Wiley & Sons, Inc.

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

Cottrell, M., J. C. Fort and G. Malgouyres. 1983. Large deviations and rare events in the study of stochastic algorithms. *IEEE Transactions on Automatic Control* AC28: 907-920.

Chistyakov, V. P. 1964. A theorem on sums of independent positive random variables and its applications to branching processes. *Theory of Probability and its Applications* 9: 640-648.

Embrechts, P., C. Klüppelberg and T. Mikosch. 1997. *Modelling extremal events*. Berlin Heidelberg: Springer-Verlag.

Embrechts, P., and C. Klüppelberg. 1993. Some aspects of insurance mathematics. *Theory of Probability and its Applications* 38 (2): 262-295.

Falkner, M., M. Devetsikiotis and I. Lambadaris. 1999. Fast simulation of networks of queues using effective and decoupling bandwidths. *ACM Transactions on Modeling and Computer Simulation* 9: 45-58.

Frater, M. R., T. M. Lenon and B. D. O Anderson. 1991. Optimally efficient estimation of the statistics of rare events in networks. *IEEE Transactions on Automatic Control* 36: 1395-1405.

Feldmann, A., and W. Whitt. 1998. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation* 31: 245-279.

Feller, W. 1966. *An introduction to probability theory and its applications, volume II*. John Wiley & Sons, Inc.

Glasserman, P., P. Heidelberger, P. Shahabuddin and T. Zajic. 1999. Multilevel splitting for estimating rare event probabilities. *Operations Research* 47: 585-600.

Glynn, P.W., and Iglehart, D.L. 1989. Importance sampling for stochastic simulations, *Management Science* 35 (11): 1367-1393.

Glynn, P.W., and Whitt, W. 1992. The asymptotic efficiency of simulation estimators. *Operations Research* 40: 505-520.

Goldie, C.M., and S. Resnick. 1988. Distributions that are both subexponential and in the domain of attraction of an extreme-value distribution. *Advances in Applied Probability* 20: 706-718.

Juneja, S., and P. Shahabuddin. 1999. Simulating heavy-tailed processes using delayed hazard rate twisting. Research Report, Dept. of Industrial Engineering and Operations Research, Columbia University, NY 10027.

Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation* 6: 43-85.

Lehtohnen, T., and H. Nyrhinen. 1992. Simulating level-crossing probabilities by importance sampling. *Advances in Applied Probability* 24: 858-874.

Leland, W., M. Taqqu, W. Willinger and D. Wilson. 1994. On the self-similar nature of Ethernet trafic. *IEEE/ACM Transactions on Networking* 2: 1-15.

Pakes, A.G. 1975. On the tails of waiting time distributions. *Journal of Applied Probability* 12: 555-564.

Parekh, S., and J. Walrand 1989. A quick simulation method for excessive backlogs in network of queues. *IEEE Transactions on Automatic Control*: 54-56.

Sadowsky, J.S. 1991. Large deviations and efficient estimation of excessive backlogs in GI/G/m queue. *IEEE Transactions on Automatic Control* 36 (1991): 1383-1394.

Shahabuddin, P. 1994. Importance sampling for the simulation of highly reliable Markovian systems. *Management Science* 40: 333-352.

Siegmund, D. 1976. Importance sampling in the Monte Carlo study of sequential tests. *The Annals of Statistics* 4: 673-684.

AUTHOR BIOGRAPHIES

NAM KYOO BOOTS is a Ph.D. student at the Department of Econometrics and Operations Research at Vrije University in the Netherlands since 1997. He received a M.S. in Econometrics (1997) and a M.S. in Mathematics (1998), both from Vrije University. His research interests include rare event simulation in stochastic models with heavy-tailed random variables. His e-mail address is <nboots@econ.vu.nl>.

PERWEZ SHAHABUDDIN is an Associate Professor in the Industrial Engineering and Operations Research Department at Columbia University. From 1990 to 1995, he was a Research Staff Member at IBM T.J. Watson Research Center. He received a B.Tech. in Mechanical Eng. from the Indian Institute of Technology, Delhi (1984), and a M.S. in Statistics and a Ph.D. in Operations Research from Stanford University (1990). He is a recipient of a 1996 National Science Foundation (U.S.A) Career Award, a 1998 IBM University Partnership Award and the 1996 Oustanding Simulation Publication Award given by INFORMS College on Simulation. He serves on the editorial boards of *Management Science, Stochastic Models, IEEE Transactions on Reliability, IIE Transactions on Operations Engineering,* and *ACM Transactions on Modeling and Computer Simulation* (Guest Editor). His e-mail address is <perwez@ieor.columbia.edu>.