TI 2001-012/4
Tinbergen Institute Discussion Paper

Simulating Tail Probabilities in GI/GI.1 Queues and Insurance Risk Processes with Sub Exponential Distributions

Nam Kyoo Boots
Perwez Shahabuddin

Simulating Tail Probabilities in GI/GI/1 Queues and Insurance Risk Processes with Subexponential Distributions

Nam Kyoo Boots[*] and Perwez Shahabuddin[†]

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 interarrival times can have any distribution with a finite mean. 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. Though there exists a vast amount of literature on the rare event simulation of queuing systems and networks with light-tailed distributions, previous fast simulation techniques for queues with subexponential service times have been confined to the M/GI/1 queue. The general approach is to use the Pollaczek-Khintchine transformation to convert 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 and evaluate 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. Some quantities other than those mentioned above can also be estimated via this approach.

Keywords: Simulation analysis methodology, variance reduction, importance sampling, rare event simulation, heavy tailed distributions, subexponential distributions, insurance risk, fluid queues, GI/GI/1 queues.

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, if W is the steady-state waiting-time random variable, then the problem is to estimate $P(W > u)$ where u is large.

[*] Dept. of Econometrics and Operations Research, Vrije University, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands. Email: nboots@econ.vu.nl

[†] Dept. of Industrial Engg. & Operations Research, Columbia University, New York, NY 10027, USA. Email: perwez@ieor.columbia.edu

The GI/GI/1 queue is strongly related to the single-source fluid queue (see, e.g., [12]). This is a buffer with a constant out-flow rate and fed by a fluid source which alternates between the on-state and the off-state. In the on-state, the source sends fluid into the queue at a certain fixed rate. Assume the times in the on-state to be subexponentially distributed and the times in the off-state to be generally distributed. Then with proper re-interpretation, the techniques in this paper can be used to estimate the steady-state probability that the buffer content in the fluid queue exceeds u at the beginning of an on-period. Problems like estimating $P(W > u)$ for the GI/GI/1 queue for large u and the above measure for fluid queues, arise, for example, while estimating probabilities of extreme delays and congestion of packets in communication networks or the packet loss probabilities in such networks. While the queuing systems used to realistically model communication networks are usually much more complex than the GI/GI/1 queue and the single source fluid queue, this work may be viewed as one of the first steps in the rare event simulation of those models in the heavy-tailed setting.

However, this is not the only reason why the GI/GI/1 queue has been widely studied in both the light-tailed (see, e.g., [15, 31, 36, 34]) and heavy-tailed setting (see, e.g., [1, 10, 11, 35]); the other reason is its connection with a canonical random walk that arises in many engineering and scientific applications. In particular, it is well-known that $P(W > u)$ for the stable GI/GI/1 queue corresponds to the probability that the maximum of a random walk, whose increment (in this case, the difference of the service time and the interarrival time random variable) has a negative expectation (thus the random walk has a negative drift), exceeds a high level u. Many other problems in engineering and finance can be transformed into the same framework as above. For example, an "inverted" version of this random walk occurs when modeling the capital process of an insurance company with initial capital u, a fixed premium collection rate, random claim sizes (that may be subexponentially distributed), and generally distributed interarrival times of claims (see, e.g. [19]). In this case, the probability mentioned above corresponds to the probability of eventual ruin, i.e., the company eventually goes bankrupt. This will be described in more detail later. Another example is the "dam problem" that is studied by civil engineers and deals with the overflow probabilities of dams and reservoirs (see, e.g., [34]).

The technique we develop in this paper is for the basic random walk problem mentioned above. Hence it can be used for any estimation problem that can be transformed into this random walk problem. However, for simplicity, we will only describe in detail the problem of estimating $P(W > u)$ in the GI/GI/1 queue and the probability of eventual ruin in the insurance model. The same basic simulation technique can also easily be extended to simulate for tail probabilities of the (busy) cycle maxima in the GI/GI/1 queue with heavy-tailed service times. This is stated as an open problem in [5, 8] for even the M/GI/1 case and is the ingredient needed to extend the results in [8] obtained for light-tailed service times to heavy-tailed service times.

A large body of work already exists for the rare event simulation of queues and networks of queues (cf. insurance risk processes) for the case where service times (cf. claim sizes) and related quantities are light-tailed (e.g., [2, 16, 24, 13, 33, 21, 36, 14, 20]; for a partial survey see [28]). 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., [38, 13, 2, 29]). 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

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

is said to be the density obtained by exponentially twisting $f(x)$ by an amount θ. 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 best 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., [30]). Note that exponential twisting relies on the existence of the moment generating functions in a neighborhood 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 [4]. They considered the problem of estimating the probability of ruin for insurance claim processes with Poisson claim arrivals and subexponentially distributed claim size. This is equivalent to the problem of estimating the steady-state waiting-time tail probability in an M/GI/1 queue with subexponential service times. They came up with an innovative algorithm based on conditioning and proved that it works for subexponential service times with a regularly-varying tail. Later, [6] gave an importance sampling change of measure for the same problem that also works for other subexponential distributions, but only if the traffic intensity is below a certain level. A framework for importance sampling for systems with subexponential distributions was presented in [27]. 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 function corresponding to $f(x)$ and let $\Lambda(x) = \int_{s=0}^{x} \lambda(s)ds$ be the hazard function. Note that the tail of any distribution, $\bar{F}(x)$, may be represented as $e^{-\Lambda(x)}$. In hazard rate twisting, the tail of 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 [27] 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 [27], [6] gave a refinement of the importance sampling algorithm in [5] that also worked 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), 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 [6] and [27], 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 generally distributed 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. In the light-tailed case large deviations theory is used to come up with efficient changes of measure. However as mentioned in [3], Pg. 287, and as illustrated by counter examples in [5], it seems that large deviations ideas do not yield good changes of measure in the heavy-tailed case. Hence it is difficult to come up with techniques that satisfy the standard criterion called "asymptotic optimality" (see, e.g., [28]; sometimes also called "asymptotic efficiency") that is used to classify a rare event simulation technique as efficient (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). We show that if we are willing to tolerate a small amount of bias in our estimator, then we can make use of some sample path large deviations ideas in the heavy-tailed setting for at least one class of subexponential distributions. Hence we develop a slightly weaker criterion, which is intended to tolerate a small amount of bias. Techniques satisfying the weaker criterion are as good for most practical purposes as the techniques satisfying the usual one, since this criterion requires that the bias be at most of the same order as the statistical variability in an asymptotical optimal estimator.

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 divided by the probability one is trying to estimate) of say δ' (usually somewhere between 0.01 and 0.1). And the δ' is usually chosen independent of the rarity of the overall event (i.e., whether one is estimating a probability 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 to be estimated 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. We call a technique large set asymptotically optimal if it is able to estimate the probability of such a large set in an asymptotical optimal fashion; we make this more precise later.

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 asymptotic 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 the probability of eventual ruin in insurance risk theory, and discusses the basic concepts in theory of subexponential distributions. Section 3 reviews rare event simulation and importance sampling. We also formalize 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 guarantees large set asymptotic optimality. In this section we also present bounds on the variance and prove the large set asymptotically optimal property. Practical insights into the simulation algorithm as well as conjectures for distributions that do not satisfy the assumptions of Section 3 and Section 4 are presented in Section 5 and Section 6. Experimental results are presented in Section 7. Section 8 summarizes some further research we are doing in this area.

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 the ratio of $z_1(x)$ to $z_2(x)$ 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 $\{x\}^+$. We define $F^{\leftarrow}(y) = \inf\{x : F(x) = y\}$. If the inverse function of F is well defined, then $F^{\leftarrow} \equiv F^{-1}$. Finally, the indicator function is denoted by $I(\cdot)$ and $\bar{F}(x) := 1 - F(x)$.

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 corresponding to $f(x)$ and let $\Lambda(x) = \int_{s=0}^{x} \lambda(s) ds$ be the hazard function (e.g., [9]). It is well-known that $\Lambda(x) = -\log \bar{F}(x)$. We assume that the first customer arrives at epoch 0 to an empty *system* and hence has a waiting time in the *queue* $W_1 = 0$. Let $(\xi_n)_{n \geq 1}$ be the sequence of i.i.d. interarrival times and $(X_n)_{n \geq 1}$ 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 both the interarrival-time distribution and the service-time distribution to have finite means, 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 $W_n = \{W_{n-1} + X_{n-1} - \xi_{n-1}\}^+$, $n \geq 2$, see, e.g., Feller [22]. Expanding this relation recursively gives

$$W_n = \max\left\{\sum_{i=1}^{n-1}(X_i - \xi_i), \ldots, \sum_{i=n-2}^{n-1}(X_i - \xi_i), X_{n-1} - \xi_{n-1}, 0\right\}. \tag{2}$$

Define the random walk $(M_n)_{n\geq 2}$ by

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

with i.i.d. increments $X_i - \xi_i$ and let $M_1 \equiv 0$. Define $\mu = -E[X - \xi] = E[X](1-\rho)/\rho$, i.e., the negative of the expected value of the increments of the random walk $(M_n)_{n\geq 1}$. Since $E[X] < E[\xi]$, $\mu > 0$. Hence the random walk has a negative drift and $P(\sup_{n\geq 1} M_n > u) \to 0$ as $u \to \infty$. It is easy to see from (2) and (3) that W_n has the same distribution as $\max_{1\leq i\leq n} M_i$. Thus the steady-state waiting time W has the same distribution as $\sup_{n\geq 1} M_n$. Thus $P(W > u) = P(\sup_{n\geq 1} M_n > u) = E[I(\sup_{n\geq 1} M_n > u)]$ and we simulate for $P(W > u)$, for large u, via the random variable $I(\sup_{n\geq 1} M_n > u)$. Let

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

be the *hitting time* of level u. Note that $\tau(u)$ is an $\{\infty\}\cup\mathbb{N}$-valued random variable and $P(\sup_{n\geq 1} M_n > u) = P(\tau(u) < \infty)$. There is a significant amount of literature for efficiently estimating quantities like $P(\sup_{n\geq 1} M_n > u)$ for large u when both the X_i's and ξ_i's are light-tailed (e.g., [2, 29, 38]). The basic contribution of this paper is to develop an efficient technique for the case where the X_i's are subexponentially distributed; the interarrival-time distribution can either be light-tailed or heavy-tailed.

2.2 Ruin probability in a renewal insurance risk process with subexponentially distributed claims

An important quantity that is studied in insurance mathematics is the ruin probability. We show that the results derived in this paper can also be applied to the renewal risk model where the possibility of large claims are modeled by using subexponentially distributed claim sizes. For more information about the use of subexponentially distributed claim sizes in risk processes and about risk processes in general, we refer the reader to [18] and the review paper [19].

Consider a insurance risk model where the period between the arrival of claim $n-1$ and n is denoted by ξ'_n and the size of claim n is denoted by X'_n. We assume both the sequence of interarrival times and the sequence of claim sizes to be i.i.d. and the two sequences to be independent of each other. We also assume that both interarrival times and the claim sizes have finite means. Premium comes in at a constant rate c. Let u denote the initial capital and let $N(t)$ denote the number of arrivals in the interval $[0,t]$. Then the capital at time t, i.e. $\{U(t)\}_{t\geq 0}$, is given by

$$U(t) = u + ct - \sum_{i=0}^{N(t)} X_i', \ t \geq 0.$$

A quantity of interest is the probability of ruin before time T with initial capital u:

$$\psi(u,T) = P(U(t) < 0 \text{ for some } t < T).$$

The probability of ultimate ruin $\psi(u)$ is given by $\psi(u,\infty)$. Since ruin can only occur at the claim arrival times,

$$\begin{aligned}\psi(u) &= P(U(t) < 0 \text{ for some } t \geq 0) = P\left(u + ct - \sum_{i=1}^{N(t)} X_i' < 0 \text{ for some } t \geq 0\right) \\ &= P\left(u + c\sum_{i=1}^{n} \xi_i' - \sum_{i=1}^{n} X_i' < 0 \text{ for some } n \geq 1\right) = P\left(\sup_{n\geq 1} \sum_{i=1}^{n} \left(X_i' - c\xi_i'\right) > u\right).\end{aligned}$$

If we take $\xi_n = c\xi_n'$ and $X_n = X_n'$, and as in (3), define the random walk $M_1 = 0$ and $M_n := \sum_{i=1}^{n-1}(X_i - \xi_i)$ for $n \geq 2$, then $\psi(u) = P(\sup_{n\geq 1} M_n > u) = P(\tau(u) < \infty)$ and we get the same random walk estimation problem as before.

2.3 Subexponential distributions and GI/GI/1 queue asymptotics

For details about subexponential distributions we refer the reader to [18]. Below we give a short summary.

The definition of subexponentiality is due to [17]:

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

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

for all n.

The integrated tail of F is defined by $F_I(x) = \int_0^x \bar{F}(y)dy/E[X]$ when $E[X] < \infty$. Let $\lambda_I(x)$ be the hazard-rate function and $\Lambda_I(x)$ be the hazard function corresponding to F_I. In this paper F_I rather than F is assumed to be subexponential. 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 [18]), 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 [32]:

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

Note that in the asymptotics of the waiting-time distribution, the interarrival-time distribution plays a role only via its first moment. Our technique works under the following assumption on the service times:

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

$F \in$MDA(Gumbel) means 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

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

For details we refer the reader to [26, 7, 18]. Examples of subexponential distributions that satisfy Assumption 1 are:

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

  $$F(x) = 1 - e^{-\sigma x^\alpha}, \ f(x) = \sigma\alpha x^{\alpha-1}e^{-\sigma x^\alpha} \ (\sigma > 0, \ 0 < \alpha < 1).$$

  In this case we may take

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

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

  $$F(x) = \Phi\left(\frac{\log x - \alpha}{\sigma}\right) \text{ and } f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left[\frac{\log x-\alpha}{\sigma}\right]^2} \ (\alpha \in \mathbb{R}, \ \sigma > 0),$$

  where Φ denotes the standard normal cumulative distribution function (cdf). 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 [7]. Define a conditional distribution $P^{(u)}$ of the random walk (M_n) by

$$P^{(u)}(\cdot) = P(\cdot \mid \tau(u) < \infty). \tag{6}$$

In case Assumption 1 holds, the asymptotic distribution of the normalized hitting time τ under the $P^{(u)}$-measure is derived in [7]: $\tau(u)/a(u)$ asymptotically has an exponential distribution. In particular, if $\overset{P^{(u)}}{\to}$ denotes convergence in the conditional distribution, then

$$\frac{\tau(u)}{a(u)} \overset{P^{(u)}}{\to} \frac{\psi}{\mu}, \tag{7}$$

where ψ is a standard exponential random variable, i.e., it has mean 1 (recall that $-\mu$ is the mean increment of the random walk (M_n)).

An important subclass of the subexponential distributions is the class of regularly-varying distributions.

Definition 2.2 *The distribution F is regularly varying of index $\alpha > 0$ (denoted by $F \in \mathcal{R}_\alpha$) if and only if*

$$\lim_{t\to\infty} \frac{\bar{F}(tx)}{\bar{F}(t)} = L(x)x^{-\alpha-1} \ (x > 0),$$

for some slowly varying function L, i.e., L is Lebesgue measurable and

$$\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1 \ (t > 0).$$

Note that these distributions may be said to have a "heavier tail" than the ones satisfying Assumption 1. An example of a regularly-varying distribution is the Pareto(σ, α) distribution with

$$F(x) = 1 - \left(1 + \frac{x}{\sigma}\right)^{-\alpha-1} \ (\alpha > 0, \ \sigma > 0). \tag{8}$$

In this case we may take

$$a(u) = \frac{\sigma + u}{\alpha}.$$

For regularly-varying service-time distributions F of index $\alpha > 0$ the convergence given by (7) still goes through, but with

$$P(\psi > x) = \left(1 + \frac{x}{\alpha}\right)^{-\alpha}. \tag{9}$$

This is in contrast to $F(\cdot)$ satisfying Assumption 1 where $\tau(u)$ has approximately an exponential tail (for large u). It is also one of the key reasons why the techniques which we discuss in this paper are not useful for distributions that are regularly-varying.

In this paper we pay special attention to the Weibull, lognormal and the Pareto distributions, since they are not only among the best known subexponential distributions, but they also illustrate the merits of the different assumptions we use for the service-time distribution. In our subsequent analysis we will also need the following assumption that is satisfied by most of the common subexponential distributions; distributions not satisfying it are mainly pathological cases (see [27] 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 $u \to \infty$. For example, $A(u) = \{\tau(u) < \infty\}$. 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. naive simulation replications. If we let $\widehat{\mathrm{Var}}\left[\hat{\alpha}(u)\right]$ be the sample estimator of $\mathrm{Var}\left[\hat{\alpha}(u)\right] = \mathrm{Var}\left[I(A(u))\right]/n$, then a $100(1-\eta)\%$ confidence interval based on the central limit theorem is given by

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

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.,

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

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 (e.g., [37]). 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 rare event simulation 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 $\alpha(u)$ iff

$$\liminf_{u\to\infty} \frac{\log\left(\mathrm{Var}\left[\hat{\alpha}(u)\right]\right)}{\log(\alpha^2(u))} \geq 1. \tag{10}$$

One typically tries to achieve the same prefixed relative error for each value of u. Informally, asymptotic optimality means that the number of replications N required to achieve a prefixed relative error is bounded, or grows very slowly as u becomes large; N is smaller than some constant times $-\log \alpha(u)$. This is in contrast to naive simulation where N is proportional to $1/\alpha(u)$.

In many cases the simulation effort per replication is either independent of the rarity parameter u or grows very weakly with it (e.g., [37, 27]). However, in cases where the growth of effort is substantial with increasing u (e.g., [23] and this paper) it is more fair to use $\mathrm{work}(u) \times \mathrm{Var}\left[\hat{\alpha}(u)\right]$ instead of $\mathrm{Var}\left[\hat{\alpha}(u)\right]$ in (10) (see, e.g., [23]). Here $\mathrm{work}(u)$ denotes the expected computational effort per simulation replication as a function of u. In that case (10) becomes

$$\liminf_{u\to\infty} \frac{\log\left(\mathrm{work}(u) \times \mathrm{Var}\left[\hat{\alpha}(u)\right]\right)}{\log(\alpha^2(u))} \geq 1. \tag{11}$$

If $\hat{\alpha}(u)$ satisfies (11), then it is called *work-normalized asymptotically optimal.* As mentioned in the Introduction, we have not been able to find a work-normalized asymptotically optimal simulation algorithm for the GI/GI/1 case and hence we introduce the weaker criterion *work-normalized large set asymptotic optimality*, 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.*

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

2. *there exists an unbiased estimator $\hat{\gamma}(u)$ of $\gamma(u)$ that is a.o., i.e.,*

$$\liminf_{u\to\infty} \frac{\log\left(\mathrm{Var}\left[\hat{\gamma}(u)\right]\right)}{\log(\gamma^2(u))} \ge 1, \tag{12}$$

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

In defining *work-normalized large set asymptotic optimality* we simply replace $\mathrm{Var}\left[\hat{\gamma}(u)\right]$ by $\mathrm{work}(u)\times \mathrm{Var}\left[\hat{\gamma}(u)\right]$ in (12).

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 asymptotically unbiased estimator with zero variance, it can be checked in (12) 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 $\alpha_a(u) \sim \alpha(u)$ (e.g., (5)) 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 biases going to zero, whereas asymptotic optimality 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 [4, 6, 27].

One would prefer to have this control over the bias for large set a.o. techniques also. However, in Definition 3.2, 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 in 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 to make the event $A(u)$ (that is rare under P) 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.

In the literature, importance sampling for queues is almost exclusively limited to exponential twisting. We illustrate the application of exponential twisting by means of two examples. Let $Z_1, \ldots, Z_k$ be light-tailed, non-negative valued, i.i.d. random variables with moment generating function $M_Z(\nu)$ and density h. Suppose we are interested in the probability $P(Z_1 + \cdots + Z_k > u)$ for large u. Under the importance sampling measure, the density h is replaced by a version that is exponentially twisted by an amount of ν, i.e., $h_\nu(x) = h(x)e^{\nu x}/M_Z(\nu)$ for some $\nu > 0$. In that case, the likelihood-ratio is given by

$$\prod_{i=1}^{k} \frac{h(Z_i)}{h_\nu(Z_i)} = M_Z(\nu)^k e^{-\nu \sum_{i=1}^k Z_i}.$$

Hence, an unbiased estimator for $P(Z_1 + \cdots + Z_k > u)$ is given by

$$I(Z_1 + \cdots + Z_k > u) M_Z(\nu)^k e^{-\sum_{i=1}^k Z_i}.$$

and its second moment is bounded by

$$E\left[I(Z_1 + \cdots + Z_k > u)[M_Z(\nu)]^{2k} e^{-2\nu \sum_{i=1}^k Z_i}\right] \leq [M_Z(\nu)]^{2k} e^{-2\nu u}.$$

Then for a given u and k one can choose a ν so as to minimize this second moment. Similar methods hold for the GI/GI/1 queue with light-tailed service times and light-tailed interarrival times where one simulates the random walk (M_n) mentioned in Section 2.1 directly, but using exponentially twisted versions of X_i and ξ_i. In this case the optimal ν is one that satisfies $M_X(-\nu)M_\xi(\nu) = 1$ and $\nu > 0$; it can be shown that an unique solution exists under fairly general conditions. The ξ is then exponentially twisted by ν and the X by $-\nu$.

However as pointed out earlier, exponential twisting is limited to random variables which have a tail that decays at an exponential or faster rate, as then one can come up with a normalizing constant that turns out to be the moment generating function. Subexponential random variables fail to have such a finite normalizing constant. In such cases, as mentioned in the Introduction, one may use hazard rate twisting (HRT) where the new distribution F_θ is given by (1). The density corresponding to F_θ is given by

$$f_\theta(x) = (1-\theta)\lambda(x)e^{-(1-\theta)\Lambda(x)}. \tag{13}$$

For Z_1 subexponential with density f, HRT leads to a likelihood-ratio of $f(Z_1)/f_\theta(Z_1)$ and thus an unbiased estimator for $P(Z_1 + \cdots + Z_k > u)$ is given by

$$\prod_{i=1}^{k} \frac{f(Z_i)}{f_\theta(Z_i)} I(Z_1 + \cdots + Z_k > u) = (1-\theta)^{-k} e^{-\theta \sum_{i=1}^{k} \Lambda(Z_i)} I(Z_1 + \cdots + Z_k > u).$$

Under some mild regularity conditions, for the choice of

$$\theta \equiv \theta_u = 1 - \frac{c}{\Lambda(u)}, \tag{14}$$

where c is any positive constant, HRT is proved to be a.o. for estimating $P(Z_1 + \cdots + Z_k > u)$ in [27].

Weighted delayed hazard rate twisting (WDHRT) extends HRT by introducing a weighting parameter w and a delaying parameter $x_u^\star$ chosen as a function of u. The WDHRT density is defined by

$$f_{\theta_u, x_u^\star}(x) = \begin{cases} \frac{f(x)}{1+w} \text{ for } x \le x_u^\star, \\ \left(1 - \frac{F(x_u^\star)}{1+w}\right) \frac{f_{\theta_u}(x)}{\bar{F}_{\theta_u}(x_u^\star)} \text{ for } x > x_u^\star. \end{cases} \tag{15}$$

In [27], $x_u^\star$ satisfies

$$\Lambda(x_u^\star) = 2\log\left(\frac{\Lambda(u)}{d}\right),$$

where d is some constant, the basic intention being that $\Lambda(x_u^\star)$ should grow at the rate of $\log(\Lambda(u))$. Note that in this case

$$P(Z \le x_u^\star) = \frac{F(x_u^\star)}{1+w} \to \frac{1}{1+w} \ (u \to \infty) \text{ and } P(Z > x_u^\star) \to \frac{w}{1+w} \ (u \to \infty).$$

If we let N be a geometrically distributed random variable with $P(N = k) = \rho^k(1-\rho)$ for $k \ge 0$, then it is well-known for the M/GI/1 queue that (e.g., [22])

$$P(W > u) = P(Y_1 + \cdots + Y_N > u), \tag{16}$$

where the sequence of i.i.d. random variables (Y_i) are distributed as the integrated tail of the service-time distribution. In [27] it is proved that for θ_u given by (14) and for certain choices of $x_u^\star$ and w (independent of u), WDHRT is a.o. for estimating $P(Y_1 + \cdots + Y_N > u)$ under some mild regularity conditions. 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, $P(N = k) = \hat{\rho}^k(1 - \hat{\rho})$ for $k \geq 0$ and some $\hat{\rho}$ for which again no explicit expression is known. The techniques in [4, 5, 6] also rely on (16) and hence are only applicable to M/GI/1 queues.

4 The Simulation Algorithm and Variance Bounds

For the GI/GI/1 case, as mentioned in Section 2.1, we estimate $P(W > u)$ by directly simulating the random walk $(M_n)_{n\geq 1}$ and estimate $P(\sup_{n\geq 1} M_n > u) = P(\tau(u) < \infty)$ (instead of using expressions like (16)). We use WDHRT for the service times, i.e., we use the density $f_{\theta_u, x_u^\star}(x)$ given in (15) ($f(x)$ is now the service-time distribution) for some specified w, θ_u and $x_u^\star$, to simulate the service times. This requires some stringent conditions on the choice of $x_u^\star$ and unlike the case in [27], requires $w \equiv w_u$ to depend on u. We argue later in this section why we do not apply any change of measure to the interarrival-time distribution.

Let Q be the new probability measure corresponding to applying WDHRT to the service times on the sample paths of (M_n). Let Z denote the resulting likelihood-ratio. In order to prove variance reduction, we have to upper bound $E_Q[Z^2 I(\tau(u) < \infty)]$ in an appropriate manner. It is useful to rewrite $E_Q[Z^2 I(\tau(u) < \infty)]$ as $E[ZI(\tau(u) < \infty)]$, since we know the asymptotic (and conditional) hitting-time distribution under the old measure, but we do not know it under the importance sampling measure. Note that

$$E[ZI(\tau(u) < \infty)] = P(\tau(u) < \infty) \sum_{k=1}^{\infty} E[Z \mid \tau(u) = k] P^{(u)}(\tau(u) = k) \tag{17}$$

(see (6) for the definition of $P^{(u)}(\cdot)$). Hence in order to obtain variance reduction it is sufficient to prove

$$\sum_{k=1}^{\infty} E[Z \mid \tau(u) = k] P^{(u)}(\tau(u) = k) < 1,$$

since naive simulation gives a second moment of $P(\tau(u) < \infty)$. Instead we prove the stronger result of (work-normalized) large set a.o. For any preselected asymptotic relative bias δ, we will use the decomposition

$$P(\tau(u) < \infty) \equiv \alpha(u) = \gamma(u) + \epsilon(u),$$

where

$$\gamma(u) = P(\tau(u) \leq k_0(u)),\ \epsilon(u) = P(k_0(u) < \tau(u) < \infty)$$

and

$$k_0(u) = -\frac{a(u)\log\delta}{\mu} = -\frac{\rho a(u)\log\delta}{(1-\rho)E[X]}. \tag{18}$$

Using (7), we have

$$\begin{aligned}\frac{\epsilon(u)}{\alpha(u)} = \frac{P(k_0(u) < \tau(u) < \infty)}{P(\tau(u) < \infty)} &= P(\tau(u) > k_0(u) \mid \tau(u) < \infty) = P^{(u)}\left(\frac{\tau(u)}{a(u)} > \frac{k_0(u)}{a(u)}\right) \\ &= P^{(u)}\left(\frac{\tau(u)}{a(u)} > \frac{-\log\delta}{\mu}\right) \to \delta\end{aligned}$$

as $u \to \infty$, thus satisfying Part 1 of Definition 3.2. We will show that $\gamma(u) = P(\tau(u) \leq k_0(u))$ may be estimated (work-normalized) a.o. using WDHRT, thus giving a (work-normalized) large set a.o. estimator for $P(\tau(u) < \infty)$. Also note that selecting

$$k_0(u) = -\frac{\beta a(u)\log\delta}{\mu} = -\frac{\rho a(u)\log\delta^{\beta}}{(1-\rho)E[X]} \tag{19}$$

gives us the flexibility required to fulfill Condition 3.3 for any fixed u. For simplicity we will use $\beta = 1$, but all the results and proofs go through with δ replaced by δ^{β}.

An important question in using WDHRT is the choice of the importance sampling parameters θ_u, w_u and $x_u^{\star}$. It is standard intuition in importance sampling for rare event simulation, that the new measure we select should induce sample paths to mimic as closely as possible the sample paths under the original measure conditioned on the rare event happening. The parameters θ_u, w_u and $x_u^{\star}$ are selected keeping this in mind. Using results from [7], one can heuristically argue that the probability law of the interarrival times under the $P^{(u)}$-measure is "rather close" to the probability law of the interarrival times under the original unconditioned measure, thus we do not apply any importance sampling to the interarrival times.

For reasons similar to those in [27], we use θ_u given by the equation

$$\theta_u = 1 - \frac{1}{\Lambda(u)}. \tag{20}$$

Furthermore, we argue that w_u should become smaller for growing ρ and u. This can be intuitively seen as follows: Since $a(u)$ tends to infinity as u goes to infinity (see Section 2.3 and [7]) and μ is decreasing as a function of ρ (if we keep $E[X]$ fixed), from (7) it follows that large ρ and/or large u tends to give more mass of the conditioned hitting-time distribution to high values. Therefore the big service time causing the rare event to happen also tends to take place later. Since w_u is controlling the chance of a big service time (i.e., a service time larger than $x_u^{\star}$), it makes sense to write w_u as a function of u and ρ. A smaller value of w_u decreases the chance of a service time larger than $x_u^{\star}$, so the big service time causing the random walk to exceed u tends to happen later. As a consequence, it makes sense to take a smaller w_u for larger ρ and/or u. In fact we will show that to obtain work-normalized large set a.o., it suffices to use $w = w_u$ given by

$$w_u = \frac{c_1\mu}{a(u)}, \tag{21}$$

where c_1 is some positive constant.

We will need the distribution functions F to satisfy the following assumption.

Assumption 3 *The $F(\cdot)$ is such that there exists some constant $b > 1$ 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 $b > 1/\alpha$.) For reasons similar to those in [27], we want $\Lambda(x_u^\star) \propto \log \Lambda(u)$. In particular we use $x_u^\star$ satisfying

$$\Lambda(x_u^\star) = b \log \Lambda(u), \tag{22}$$

where b is the constant in Assumption 3. Since $\Lambda(x) = -\log(1 - F(x))$,

$$x_u^\star = F^{\leftarrow}\left(1 - e^{-b\log\Lambda(u)}\right) = F^{\leftarrow}\left(1 - \Lambda(u)^{-b}\right),$$

which is an useful representation of $x_u^\star$ from the computational point of view. Note that $x_u^\star$ goes to infinity as u goes to infinity, because $\Lambda(u) \to \infty$.

Finally, we will also need the following assumption for reasons that will become clear later.

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

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

Assumption 4 is satisfied by the commonly used subexponential distributions in MDA(Gumbel), like the Weibull and the lognormal distribution.

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

Algorithm 1 "Weighted delayed hazard rate twisting of the service times"

1. *Draw i.i.d. samples $\xi_1, \ldots, \xi_k$ from the interarrival-time distribution and i.i.d. samples $X_1, \ldots, X_k$ using the density $f_{\theta_u, x_u^\star}(x)$, where k is the minimum of $k_0(u)$ and $\min\left\{i : \sum_{j=1}^{i}(X_j - \xi_j) > u\right\}$.*
2. *Compute the likelihood-ratio Z given by*

$$Z = \frac{f(X_1)}{f_{\theta_u, x_u^\star}(X_1)} \cdots \frac{f(X_k)}{f_{\theta_u, x_u^\star}(X_k)}.$$

3. *An average of many independent samples of $ZI\left(\sum_{j=1}^{k}(X_j - \xi_j) > u\right)$ is an unbiased estimator for $P(\tau(u) \le k_0(u))$ which is used as an estimator for $P(\tau(u) < \infty)$.*

Theorem 4.1 *Algorithm 1 results in a work-normalized large set a.o. estimator for $P(\tau(u) < \infty)$ with $\gamma(u) = P(\tau(u) \le k_0(u))$ and $\epsilon(u) = P(k_0(u) < \tau(u) < \infty)$.*

As mentioned before, the only thing which needs to be shown is that the estimator of $P(\tau(u) \leq k_0(u))$ is work-normalized a.o. The formal proof is given in Appendix A. Below we describe the basic approach.

First we partition the set $\{\tau(u) = k\}$ into several subsets and derive variance bounds on each of these subsets. Let A_n^k be the set of sample paths of M_n where $\{\tau(u) = k\}$ and the number of the first k service times higher than $x_u^\star$ equals n, $n \leq k$. Define $A^k = \cup_{n=1}^k A_n^k$, $A = \cup_{k=1}^\infty A^k$ and $A' = \cup_{k=1}^\infty A_0^k$ (thus $A \cup A' = \{\tau(u) < \infty\}$ and $A \cap A' = \emptyset$). For notational convenience, we assume that $k_0(u)$ always has an integer value. In this way we are able to partition the rare event $\{\tau(u) \leq k_0\}$ into two sets. These two sets are,

$$A \cap \{\tau(u) \leq k_0(u)\} \equiv \cup_{k=1}^{k_0(u)} A^k, \text{ and } A' \cap \{\tau(u) \leq k_0(u)\}.$$

We then use the following steps to upper bound $E[ZI(\tau(u) \leq k_0(u))]$:

- First, we upper bound $E[ZI(A_n^k)]$ for $n = 1, \ldots, k$. We show that the upper bound for $n = 1$ can also be used for upper bounding $E[ZI(A_n^k)]$ for $n = 2, \ldots, k$.
- Subsequently, we derive an upper bound on $E[ZI(A)I(\tau(u) \leq k_0(u))]$ by summing up the bounds on $E[ZI(A^k)]$ for $k \leq k_0(u)$.
- We also show that for u large enough, $A' \cap \{\tau(u) \leq k_0(u)\} = \emptyset$ and hence $P(A' \cap \{\tau(u) \leq k_0(u)\}) = 0$. This result follows directly from Assumption 4.

All this is summarized in the following proposition. The proof is given in Appendix A.

Proposition 4.2 *For u large enough,*

(i) $E[ZI(A)I(\tau(u) \leq k_0(u))] \leq K_1 a(u)\Lambda(u) e^{-\Lambda_I(u)} e^{-\theta_u \Lambda(u - k_0(u)x_u^\star + x_u^\star)}$.

where K_1 is some positive constant (i.e., quantity independent of u) and

(ii) $E[ZI(A')I(\tau(u) \leq k_0(u))] = 0$.

The a.o. property follows from the idea that $\exp(-\theta_u \Lambda(u - k_0(u)x_u^\star + x_u^\star))$, $\exp(-\Lambda(u))$, $\exp(-\Lambda_I(u))$ and $P(\tau(u) < \infty)$ are asymptotically equivalent in the log, and the rate of increase of $a(u)\Lambda(u)$ is much slower than the rate of decrease of $\exp(-\Lambda_I(u))$.

5 Practical issues

In this section we discuss the more practical aspects of the WDHRT simulation algorithm. In order to prove that Algorithm 1 is large set work-normalized a.o., we make use of the fact that for all u large enough (say larger than u_0), $A' \cap \{\tau(u) \leq k_0(u)\} = \emptyset$. However, in all the experimental results we present for Weibull service times, the actual value of u is smaller than u_0. Thus it is possible that for practical values of u, Algorithm 1 induces a lot of variance on the set $A' \cap \{\tau(u) \leq k_0(u)\}$. In Section 5.1

we derive an upper bound on $P(A')$ and an upper bound on the variance of Algorithm 1 on the set $A' \cap \{\tau(u) \le k_0(u)\}$ that holds for *all* u. In the previous section we only gave some restrictions on the values of our parameters. In Section 5.2, 5.3 and 5.4 we give some heuristic arguments to choose them in the best possible way, since the quality of our simulation results can depend heavily on the particular choice of some parameters (*even though choosing them in this way is not necessary for work-normalized large set a.o.*).

5.1 Upper bounding $P(A')$ and the corresponding variance

Note that on the set A' all the service times are bounded by $x_u^\star$. Let $(\tilde{X}_i)$ be a sequence of i.i.d. random variables with distribution $F_{\tilde{X}}$ and density $f_{\tilde{X}}$ with

$$f_{\tilde{X}}(x) = \begin{cases} \frac{f(x)}{F(x_u^\star)} \text{ for } x \le x_u^\star, \\ 0 \text{ for } x > x_u^\star. \end{cases}$$

Consider the alternative GI/GI/1 queue with the service times (X_i) replaced by $(\tilde{X}_i)$. It is easy to see that $E[\tilde{X}] < E[X]$, since

$$\begin{aligned} E[X] &= E[X \mid X \le x_u^\star]P(X \le x_u^\star) + E[X \mid X > x_u^\star]P(X > x_u^\star) \\ &\ge E[X \mid X \le x_u^\star]P(X \le x_u^\star) + x_u^\star P(X > x_u^\star) \\ &\ge E[X \mid X \le x_u^\star]P(X \le x_u^\star) + E[X \mid X \le x_u^\star]P(X > x_u^\star) \\ &= E[X \mid X \le x_u^\star] = E[\tilde{X}]. \end{aligned}$$

Thus $E[\tilde{X}] < E[\xi]$, since $E[X] < E[\xi]$. This implies that the new queue is also stable. Let $\tilde{\tau}(u)$ be the hitting time in the new queuing system. From

$$\begin{aligned} P(A') &= \sum_{k=1}^{\infty} P(\tau(u) = k \mid X_1 \le x_u^\star; \cdots; X_k \le x_u^\star) P(X_1 \le x_u^\star; \cdots; X_k \le x_u^\star) \\ &\le \sum_{k=1}^{\infty} P(\tau(u) = k \mid X_1 \le x_u^\star; \cdots; X_k \le x_u^\star) = \sum_{k=1}^{\infty} P(\tilde{\tau}(u) = k) = P(\tilde{\tau}(u) < \infty), \end{aligned}$$

it follows that $P(A') \le P(\tilde{\tau}(u) < \infty)$. Since $\tilde{X}$ has a finite support, its moment generating function is finite everywhere. We can then use (for fixed u) a variance/expectation bounding method that is also used for light-tailed theory.

Let $M_{\tilde{X}}(\nu)$ and $M_\xi(\nu)$ be the moment generating functions of $\tilde{X}$ and ξ respectively. Define $\nu_{x_u^\star} > 0$ as the solution of

$$M_{\tilde{X}}(\nu) M_\xi(-\nu) = 1, \ \nu > 0. \tag{23}$$

From importance sampling theory for light-tailed distributions, it is well-known that such a ν exists and is unique, see, e.g., [13] (the proof relies on the convexity of $M_{\tilde{X}}(\nu)M_\xi(-\nu)$). Define the exponentially twisted density (by amount $\nu_{x_u^\star}$) corresponding to $\tilde{X}$ by

$$f_{\tilde{X}}^{\nu_{x_u^\star}}(x) = \frac{f_{\tilde{X}}(x) e^{\nu_{x_u^\star}}}{M_{\tilde{X}}(\nu_{x_u^\star})},$$

for $0 \leq x \leq x_u^\star$. Similarly, define the exponentially twisted density (by amount $-\nu_{x_u^\star}$) corresponding to ξ by

$$f_\xi^{-\nu_{x_u^\star}}(x) = \frac{f_\xi(x)e^{-\nu_{x_u^\star}}}{M_\xi(-\nu_{x_u^\star})}.$$

If we denote with $\tilde{E}$ the expectation under the importance sampling measure, then

$$\begin{aligned} P(A') &\leq P(\tilde\tau(u) < \infty) = E[I(\tilde\tau(u) < \infty)] \\ &= \tilde{E}\left[I(\tilde\tau(u) < \infty)e^{-[(\tilde{X}_1-\xi_1)+\cdots+(\tilde{X}_{\tilde\tau(u)}-\xi_{\tilde\tau(u)})]\nu_{x_u^\star}}\left[M_{\tilde{X}}(\nu_{x_u^\star})M_\xi(-\nu_{x_u^\star})\right]^{\tau(u)}\right] \\ &\leq \tilde{E}\left[I(\tilde\tau(u) < \infty)e^{-u\nu_{x_u^\star}}\right] \leq e^{-u\nu_{x_u^\star}}, \end{aligned} \tag{24}$$

using (23) and the fact that on the set $\{\tilde\tau(u) < \infty\}$, $(\tilde{X}_1 - \xi_1) + \cdots + (\tilde{X}_{\tilde\tau(u)} - \xi_{\tilde\tau(u)}) > u$.

In fact, we just proved part (i) of the following theorem:

Theorem 5.1 *Let u be fixed and let $\nu_{x_u^\star}$ be the unique solution to (23), then*

(i) $P(A') \leq e^{-u\nu_{x_u^\star}}$

and

(ii) $E[ZI(A' \cap \{\tau(u) \leq k_0(u)\}] \leq e^{-(u\nu_{x_u^\star}+c_1\log\delta)}$.

Proof. For (ii), note that for $x \leq x_u^\star$, $f(x)/f_{\theta_u,x_u^\star}(x) = 1 + w_u$. Hence

$$\begin{aligned} E[ZI(A' \cap \{\tau(u) \leq k_0(u)\})] &\leq (1+w_u)^{k_0(u)}P(A') \leq e^{-u\nu_{x_u^\star}}\left[\left(1+\frac{c_1\mu}{a(u)}\right)^{\frac{a(u)}{c_1\mu}}\right]^{-c_1\log\delta} \\ &\leq e^{-(u\nu_{x_u^\star}+c_1\log\delta)}. \end{aligned}$$

Here we use the fact that $(1 + x^{-1})^x \leq e$ for all $x > 0$. □

Since we have not been able to come up with an useful upper bound on $\nu_{x_u^\star}$, we have to solve (23) numerically.

5.2 The choice of b

It is noteworthy that a smaller choice of b corresponds to a smaller $x_u^\star$ and which suggests a smaller $P(A')$. Using WDHRT for estimating $P(\tau(u) \leq k_0(u))$ implies that in fact no HRT is done on the set of sample paths in A'. The only difference with naive simulation is that the chance of such paths is smaller under the importance sampling measure for positive w_u, and that increases the variance contribution due to importance sampling on that set of sample paths. This suggests that one should keep the set A' as small as possible and hence to choose b rather small. For instance, for the Weibull case we have the restriction $b > 1/\alpha$. Hence, for $\alpha = .5$, a choice of $b = 2.1$ seems reasonable.

5.3 The choice of w_u

Recall that $w_u = c_1\mu/a(u)$ for some $c_1 > 0$. To derive an upper bound on the variance using importance sampling, sum the right-hand side of (31) over $1 \leq k \leq k_0(u)$. Then we need to choose w_u such that

$$\sum_{k=1}^{k_0(u)} \frac{(1+w_u)^k}{w_u + \Lambda(u)^{-b}} \left(P^{(u)}(\tau(u) > k-1) - P^{(u)}(\tau(u) > k)\right) \tag{25}$$

is as small as possible (since the remaining part is not a function of w_u). Suppose we use the approximations $P^{(u)}(\tau(u) > k) \approx e^{-k\mu/a(u)}$ and $P^{(u)}(\tau(u) = k) \approx \mu/a(u)e^{-k\mu/a(u)}$, which are both based on (7). To avoid the geometric growth in k of $\left(P^{(u)}(\tau(u) > k-1) - P^{(u)}(\tau(u) > k)\right)(1+w_u)^k \approx \mu/a(u)e^{-k\mu/a(u)}(1+c_1\mu/a(u))^k$, it is required that

$$e^{-\frac{\mu}{a(u)}}\left(1 + \frac{c_1\mu}{a(u)}\right) < 1. \tag{26}$$

Hence we recommend c_1 to be selected such that (26) is satisfied. A more refined heuristic for the choice of c_1 is to use the one that minimizes (7), where we use the approximation $P^{(u)}(\tau(u) > k) \approx e^{-k\mu/a(u)}$ when conducting the minimization. As expected, in all our experiments this optimal c_1 satisfies (26) (see Table 3).

5.4 The choice of δ

In this paragraph we present some guidelines to choose the relative bias δ. If we usually want to achieve a relative error (confidence interval half-width upon the estimated quantity) of δ' in unbiased simulations, it makes sense to choose δ somewhat smaller than δ', say $\delta = \delta'/10$. We then use $k_0(u) = -\beta a(u)\log\delta/\mu = -a(u)\log\delta^\beta/\mu$, with $\beta \geq 1$ (e.g. $\beta = 2$). Recall that the factor β is used because just using $k_0(u) = -a(u)\log\delta/\mu$ guarantees that the asymptotic relative bias is less than δ; the actual (non-asymptotic) relative bias may be higher than δ.

6 Relaxing the assumptions on the service-time distribution

In this section we discuss the merits of omitting Assumption(s) 1, 3 and 4. We show that, although our previous analysis is not valid anymore, omitting Assumption 3 may lead to an efficient algorithm anyway. In Section 6.1 we illustrate this by considering lognormal service times, since this is the most important distribution that satisfies Assumptions 1, 2 and 4, but not Assumption 3. Of course, similar analysis can be done for other distributions. More problematic is leaving out Assumptions 1, 3 and 4. Unfortunately, the important class of regularly-varying distributions falls in this category. We discuss the particular case of Pareto service times in Section 6.2.

6.1 Lognormal service times

In this paragraph, we assume the service times to have a lognormal distribution. Recall that Assumptions 1 and 2 are satisfied. In Appendix B we show that Assumption 4 is also satisfied. Unfortunately, it is easy to check that Assumption 3 does not hold, but in Appendix B we argue heuristically that Algorithm 1 may work anyway.

6.2 Regularly-varying service times

For Pareto service times (see (8)),

$$\frac{a(u)x_u^\star}{u} = \frac{(\sigma+u)x_u^\star}{\alpha u} \to \infty \ (u \to \infty),$$

since $x_u^\star$ goes to infinity. Hence Assumption 4 does not hold. Also,

$$\frac{a(u)}{\Lambda(u)} = \frac{(\sigma+u)}{\alpha(\alpha+1)\log(1+u/\sigma)} \to \infty,$$

as u goes to infinity. Hence Assumption 3 does not hold. Using standard theory of regularly-varying distributions, it can easily be checked that these go through for the other regularly-varying distributions. Therefore, a similar analysis as for the lognormal case becomes impossible. Indeed numerical experiments give little hope for this case.

Also, from (9) it follows that solving $k_0(u)$ from the equation

$$P^{(u)}(\tau(u) > k_0(u)) \to \delta \ (u \to \infty),$$

gives

$$k_0(u) = \frac{a(u)\alpha\rho\left[\delta^{-1/\alpha}-1\right]}{E[X](1-\rho)} = \frac{(\sigma+u)\alpha\rho\left[\delta^{-1/\alpha}-1\right]}{E[X](1-\rho)}. \tag{27}$$

Hence, $k_0(u)$ grows linearly with u and for realistic values of the parameters, WDHRT becomes very time consuming, since it gives a lot of mass at high values of the hitting time. This implies that the simulation effort grows very fast for increasing u. The following theorem states that, even if one is able to come up with a method for estimating $\gamma(u)$ that has bounded relative error, it still does not guarantee *work-normalized* large set a.o.:

Theorem 6.1 *Consider Pareto service times with corresponding $k_0(u)$ given by (27). For any unbiased importance sampling estimator $\hat\gamma(u)$ of $\gamma(u) = P(\tau(u) \le k_0(u))$ with the property that work$(u) \propto k_0(u)$ and that the relative error of $\hat\gamma(u)$ is larger then some positive constant independent of u, the $\hat\gamma(u)$ is not a work-normalized large set a.o. estimator for $P(\tau(u) < \infty)$.*

Proof. Suppose that the relative error of $\hat\gamma(u)$ is larger then some positive constant K_1. But then (5) and the fact that $0 < \gamma(u) < 1$, imply that there exist positive constants K_2, K_3 and K_4 such that

$$\liminf_{u\to\infty} \frac{\log\left(\mathrm{Var}\left[\hat\gamma(u)\right] \times \mathrm{work}(u)\right)}{\log(\gamma^2(u))} \quad \le \quad \lim_{u\to\infty}\left[\frac{\log((K_1)^2\,\gamma^2(u))}{\log(\gamma^2(u))} + \frac{\log(K_3(u))}{2\log(K_2 P(\tau(u) < \infty))}\right]$$

$$\leq \lim_{u\to\infty}\left[\frac{\log((K_1)^2\,\gamma^2(u))}{\log(\gamma^2(u))}+\frac{\log(K_3(u))}{2\log(K_2K_4u^{-\alpha})}\right]=1-\frac{1}{2\alpha}<1,$$

if $\mathrm{work}(u) \propto k_0(u)$. □

7 Experimental results

In this section we present some experimental results using Algorithm 1 (A1) for Weibull and lognormal service times. We first present results for the M/GI/1 queue, since for this case we can compare the results with those from [27] and [4]. For the sake of comparison we also present estimates based on the best-known asymptotic approximation (AA) for $P(W > u)$ given by (5). As argued in Section 6.2, A1 fails to be efficient for Pareto service times. However, we claimed in the Introduction that in this case AA is of a better quality than for the case of Weibull or lognormal service times. Experimental results that compare AA with accurate estimates using the algorithm from [4] for M/GI/1 systems that have Pareto service times, support this claim. We also present some examples for deterministic interarrival times and we compare the results with estimates resulting from naively simulating the hitting time of u in the random walk (M_n).

For the sake of clarity, we present in Tables 1 and 2 a classification of the subexponential distributions as used in this paper. The Weibull, lognormal and Pareto distribution in this classification correspond, respectively, to one of the three most important regimes used in the literature.

Table 1: Different assumptions on the service-time distributions

(i)	$F \in \mathcal{S}$, $F_I \in \mathcal{S}$
(ii)	$F \in$MDA(Gumbel)
(iii)	$\lambda(x)$ is eventually decreasing
(iv)	$a(u)x_u^*/u \to 0$ $(u \to \infty)$
(v)	$\Lambda(u)^{-b+1}/w_u \to 0$ $(u \to \infty)$

Table 2: When and when not our technique will work

The Roman numbers correspond to the assumptions in Table 1	
(i) and not (ii) $\Rightarrow$ A1 will probably not work	e.g. Pareto
$(i), (ii), (iii), (iv)$ and not (v) $\Rightarrow$ A1 will probably work	e.g. lognormal
$(i), (ii), (iii), (iv), (v)$ $\Rightarrow$ A1 is work-normalized large set a.o.	e.g. Weibull

Note that a regenerative simulation method for estimating $P(W > u)$ based on estimating the average number of customers in a regenerative cycle (a regenerative cycle is taken to be the period between two consecutive epochs at which a customer arrives at an empty queue) with a waiting time higher than u, produces very unstable estimates. This unstability is a result of the enormous fluctuation in cycle lengths that are caused by the subexponential service times. Indeed, we conducted some experiments that support this claim and hence it is not advisable to use this regenerative method.

7.1 Weibull service times

We use Weibull service times, with the specific distribution function given by $1 - \exp(-\sqrt{x})$, $x > 0$. It can easily be checked that $E[X] = 2$ and this class of distributions satisfies the assumptions in Section 4. We use $k_0(u) = \max\{-a(u) \log \delta^\beta / \mu, 50\}$ to guarantee that the actual value of $k_0(u)$ is not too small. This can also be interpreted as $k_0(u) = -\max\{a(u) \log \delta^\beta / \mu, 50\}$, where we take $\delta^\beta = 0.001$ for u such that $k_0(u) \geq 50$ and δ^β such that $k_0(u) = 50$ for the smaller u. We use $b = 2.1$, consistent with Assumption 3.

Table 3: Values of the Parameters (For b we use 2.1)

u	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
100	$w_u = .1693, c_1 = .56$	$w_u = .0503, c_1 = .50$	$w_u = 0.0135, c_1 = .41$
	$x_u^\star = 23.38$	$x_u^\star = 23.38$	$x_u^\star = 23.38$
	$\nu_{x_u^\star} = .125$	$\nu_{x_u^\star} = .156,$	$\nu_{x_u^\star} = .099$
	$e^{-u\nu_{x_u^\star}} = 3.8E - 6$	$e^{-u\nu_{x_u^\star}} = .1.7E - 7$	$e^{-u\nu_{x_u^\star}} = 4.9E - 5$
200	$w_u = .1185, c_1 = .56$	$w_u = .0364, c_1 = .51$	$w_u = 0.0105, c_1 = .45$
	$x_u^\star = 30.95$	$x_u^\star = 30.95$	$x_u^\star = 30.95$
	$\nu_{x_u^\star} = .125$	$\nu_{x_u^\star} = .122$	$\nu_{x_u^\star} = 073$
	$e^{-u\nu_{x_u^\star}} = 1.4E - 11$	$e^{-u\nu_{x_u^\star}} = 2.6E - 11$	$e^{-u\nu_{x_u^\star}} = 4.3E - 7$
400	$w_u = .0827, c_1 = .55$	$w_u = .0261, c_1 = .52$	$w_u = 0.0079, c_1 = .47$
	$x_u^\star = 39.58$	$x_u^\star = 39.58$	$x_u^\star = 39.58$
	$\nu_{x_u^\star} = .125$	$\nu_{x_u^\star} = .101$	$\nu_{x_u^\star} = .059$
	$e^{-u\nu_{x_u^\star}} = 2.0E - 22$	$e^{-u\nu_{x_u^\star}} = 2.4E - 18$	$e^{-u\nu_{x_u^\star}} = 6.8E - 11$
800	$w_u = .058, c_1 = .55$	$w_u = .0186, c_1 = .53$	$w_u = 0.0058, c_1 = .49$
	$x_u^\star = 49.26$	$x_u^\star = 49.26$	$x_u^\star = 49.26$
	$\nu_{x_u^\star} = .125$	$\nu_{x_u^\star} = .088$	$\nu_{x_u^\star} = .0508$
	$e^{-u\nu_{x_u^\star}} = 4.0E - 44$	$e^{-u\nu_{x_u^\star}} = 2.4E - 31$	$e^{-u\nu_{x_u^\star}} = 4.9E - 18$

The values of the other parameters used by the algorithm are given in Table 3. They were determined using the heuristic approach described in Section 5. Note that for the general subexponential Weibull distribution (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 1 is easier to implement than the ones in [6] and [27], for service-time distributions for which the integrated-tail distribution is difficult to compute. However, it is usually far less efficient in terms of CPU time.

Table 4: Estimates of $P(W > u)$ for the M/GI/1 Queue with Weibull$(1, 1/2)$ Service Times. A1 uses techniques from this paper, AA is the asymptotic approximation and J-S denotes the estimator from [27] and is used to get relatively accurate estimates of $P(W > u)$. The number in the parenthesis besides the A1 estimate denotes the efficiency ratio over naive simulation. The number in the parenthesis besides the AA estimate denotes the relative bias of AA.

u		$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
100	A1	$2.31E-4 \pm 1.7\%$ $(2.4E2)$	$1.38E-3 \pm 3.4\%$ (3.8)	$1.68E-2 \pm 10.8\%$ (0.07)
	J-S	$2.30E-4 \pm 1.3\%$	$1.41E-3 \pm 1.3\%$	$1.89E-2 \pm .67\%$
	AA	$1.17E-4(49.1\%)$	$5.00E-4(64.5\%)$	$1.50E3(92.1\%)$
200	A1	$4.71E-6 \pm 2.0\%$ $(5.9E3)$	$2.46E-5 \pm 3.5\%$ $(1.5E2)$	$6.41E-4 \pm 38.8\%$ $(.46)$
	J-S	$4.61E-6 \pm 1.5\%$	$2.55E-5 \pm 3.15\%$	$7.37E-4 \pm 3.3\%$
	AA	$3.64E-6(21.0\%)$	$1.09E-5(57.3\%)$	$3.28E-5(95.5\%)$
400	A1	$1.65E-8 \pm 2.5\%$ $(9.2E5)$	$7.12E-8 \pm 3.1\%$ $(1.3E5)$	$1.53E-6 \pm 69.5\%$ (13.7)
	J-S	$1.66E-8 \pm 1.6\%$	$7.11E-8 \pm 2.75\%$	$1.62E-6 \pm 43.3\%$
	AA	$1.44E-8(13.3\%)$	$4.33E-8(39.1\%)$	$1.30E-7(92.0\%)$
800	A1	$5.54E-12 \pm 3.0\%$ $(1.8E9)$	$2.04E-11 \pm 3.1\%, (4.5E8)$	$1.27E-10 \pm 9.4\%$ $(1.5E6)$
	J-S	$5.45E-12 \pm 2.0\%$	$2.04E-11 \pm 1.8\%,$	$1.36E-10 \pm 7.9\%$
	AA	$5.08E-12(6.8\%)$	$1.52E-12(25.5\%)$	$4.57E-11(66.4\%)$

The results are presented in Table 4. The results from [27], denoted by J-S, were based on 10,000,000 replications, in order to get accurate estimates for comparison purposes. For A1, 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. Motivated by [25], we define the standard effort of any simulation algorithm as the variance per simulation replication times the CPU time per simulation replication. The numbers in the parenthesis besides the A1 estimator denote the *efficiency ratio*, which is the ratio of the standard effort of naive simulation and the standard effort of A1. For naive simulation the standard effort is estimated by using the estimate of $P(W > u)$ from J-S 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 $k_0(u)$ (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 more CPU time 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 [27], as that algorithm can only be used for the special case of M/GI/1 systems. Indeed, as mentioned above, for the M/GI/1 case the algorithms in [6] and [27] are much better. The number in the parenthesis besides the AA denote the relative bias of AA, i.e., 100% × $|\hat{\alpha}(u) - \alpha_a(u)|/\hat{\alpha}(u)$, where $\hat{\alpha}(u)$ is the accurate simulation estimate from J-S. Estimates in Table 4 for high values of ρ are not accurate for low u and the given number of simulation replications. This is also the case for J-S and $u = 400$. However, for large u the asymptotics take effect and the accuracy improves. From Table 4 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 $k_0(u)$ (to decrease the relative bias) and/or run more simulation replications to improve the estimates from A1.

Table 5: Estimates of $P(W > u)$ for the $D/GI/1$ Queue with Weibull$(1, 1/2)$ Service Times. The number in the parenthesis besides the A1 estimate denotes the efficiency ratio. The number in the parenthesis besides the AA estimate denotes the relative bias of AA.

$u = 150$	$\rho = 0.5$
A1	$1.12E - 4 \pm 2.5\%(80.7)$
naive simulation	$1.10E - 4 \pm 2.5\%$
AA	$6.36E - 5(42.2\%)$

In Table 5 we present an example where we use deterministic interarrival times. The values of the parameters are the same as for the case of Poisson arrivals. We use 300,000 replications for A1 and 100,000,000 replications for naive simulation. The large number of replications for naive simulation were necessary in order to get sufficiently accurate estimates for comparison purposes. The results are also compared with AA. For the relative bias of AA, we compare AA with the relatively accurate naive simulation estimate.

7.2 Lognormal service times

For the lognormal distribution we take $\alpha = 0$, i.e., the density is given by

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

with $\sigma^2 = \log(4)$. One can check that $E[X]$ is again 2. It is difficult to implement hazard rate twisting, since we have no explicit expression for f_{θ_u}. As an alternative, we apply another form of subexponential twisting (see [27]) that involves using

$$f_{\theta_u}(x) = \frac{1}{x\sqrt{2\pi\sigma_{\theta_u}^2}} e^{-\frac{1}{2}\left[\frac{\log x}{\sigma_{\theta_u}}\right]^2},$$

where $\sigma_{\theta_u} = \sigma/(1 - \theta_u)$. We take $w_u = 0$ (see Appendix B for a motivation for this choice) and like in the Weibull case, we take $k_0(u) = \max\{-a(u) \log \delta^\beta/\mu, 50\}$. We also use $b = 2.1$. To compare answers we use the order statistics conditioning algorithm from [4], which is based on the Pollaczek-Khintchine transformation. We denote this algorithm by A-B. We refer the reader to that paper for details on the algorithm.

Table 6: Estimates of $P(W > u)$ for the M/GI/1 Queue with lognormal$(0, \log 4)$ Service Times. A1 uses techniques from this paper, AA is the asymptotic approximation and A-B denotes the estimator from [4] and is used to get relatively accurate estimates of $P(W > u)$. The number in the parenthesis besides the A1 estimate denotes the efficiency ratio. The number in the parenthesis besides the AA estimate denotes the relative bias of AA.

u		$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
100	A1	$3.14E-4 \pm 10.4\%$ (5.6)	$1.06E-3 \pm 7.0\%$ (.8)	$7.14E-3 \pm 26.6\%$ (0.03)
	A-B	$3.06E-4 \pm 2.3\%$	$1.15E-3 \pm 2.3\%$	$7.59E-3 \pm 1.8\%$
	AA	$2.78E-4(9.2\%)$	$8.33E-4(27.6\%)$	$2.49E-3(67.2\%)$
200	A1	$3.48E-5 \pm 9.3\%$ $(31.4E2)$	$1.26E-4 \pm 7.2\%$ (26.7)	$5.93E-4 \pm 32.7\%$ (6.9)
	A-B	$3.72E-5 \pm 1.5\%$	$1.23E-4 \pm 1.6\%$	$6.15E-4 \pm 3.7\%$
	AA	$3.53E-5(5.2\%)$	$1.06E-4(13.8\%)$	$3.18E-4(48.3\%)$
400	A1	$3.29E-6 \pm 10.1\%$ $(3.9E2)$	$1.04E-5 \pm 6.3\%$ $(3.0E2)$	$4.64E-5 \pm 29.3\%$ $(1.1E2)$
	A-B	$3.38E-6 \pm 1.3\%$	$1.08E-5 \pm 4.1\%$	$3.91E-5 \pm 4.0\%$
	AA	$3.27E-6(3.3\%)$	$9.81E-6(9.2\%)$	$2.94E-5(24.8\%)$
800	A1	$2.19E-7 \pm 5.2\%$ $(5.4E3)$	$7.00E-7 \pm 8.1\%, (2.2E3)$	$2.13E-6 \pm 12.8\%$ $(3.4E2)$
	A-B	$2.22E-7 \pm 1.0\%$	$6.90E-7 \pm 1.7\%,$	$2.28E-6 \pm 2.1\%$
	AA	$2.19E-7(1.4\%)$	$6.58E-7(4.6\%)$	$1.97E-6(13.6\%)$

For A-B, in order to draw from F_I we use numerical integration. Since this method is rather time consuming, we use only 1,000,000 replications for A-B. For A1 we use 100,000 replications for $\rho = .25$ and $\rho = .5$ and 30,000 replications for $\rho = .75$. The results are presented in Table 6. Underestimation of A1 is more severe than for the Weibull case, but for high values of u and moderate values of ρ, the estimates seem to be pretty good. Note that this is also a region where the asymptotic estimation performs quite well. Hence for the case of lognormal distributions, the only advantage of using A1 over AA is that one can reduce the relative error by using larger number of replications, and/or reduce the relative bias by increasing $k_0(u)$, while the relative bias of AA is beyond our control.

In Table 7 we present an example where we use deterministic interarrival times. The values of the parameters are the same as for the case of Poisson arrivals. We use 300,000 replications for A1 and 20,000,000 replications for naive simulation. The results are compared with naive simulation and AA. Once again, for the relative bias of AA, we compare AA with the relatively accurate naive simulation estimate.

Table 7: Estimates of $P(W > u)$ for the $D/GI/1$ Queue with lognormal$(0, \log 4)$ Service Times. The number in the parenthesis besides the A1 estimate denotes the efficiency ratio. The number in the parenthesis besides the AA estimate denotes the relative bias of AA.

$u = 200$	$\rho = 0.5$
A1	$1.10E - 4 \pm 4.5\%(98.2)$
naive simulation	$1.10E - 4 \pm 5.5\%$
AA	$1.06E - 4(3.8\%)$

Table 8: Estimates of $P(W > u)$ for the M/GI/1 Queue with Pareto$(2, 4)$ Service Times. AA is the asymptotic approximation and A-B denotes the estimator from [4] and is used to get relatively accurate estimates of $P(W > u)$. The number in the parenthesis besides the AA estimate denotes the relative bias of AA.

u		$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
100	A-B	$5.24E - 4 \pm .40\%$	$1.79E - 3 \pm .37\%$	$8.92E - 3 \pm .35\%$
	AA	$4.93E - 4(6.0\%)$	$1.48E - 3(17.4\%)$	$4.44E - 3(50.2\%)$
200	A-B	$1.32E - 4 \pm .40\%$	$4.21E - 4 \pm .41\%$	$1.59E - 3 \pm .43\%$
	AA	$1.28E - 4(2.9\%)$	$3.84E - 4(8.8\%)$	$1.15E - 3(27.3\%)$
400	A-B	$3.31E - 5 \pm .30\%$	$1.03E - 4 \pm .44\%$	$3.39E - 4 \pm .44\%$
	AA	$2.27E - 5(1.4\%)$	$9.8E - 5(4.7\%)$	$2.94E - 4(13.3\%)$
800	A-B	$8.31E - 6 \pm .34\%$	$2.53E - 5 \pm .37\%$	$7.91E - 5 \pm .50\%$
	AA	$8.25E - 6(0.7\%)$	$2.48E - 5(2.0\%)$	$7.43E - 5(6.2\%)$

7.3 Regularly-varying service times

To get a complete picture we present some results from the M/GI/1 queue with Pareto service times. Although A1 does not work well in this case (see Section 6.2), we can still compare estimates obtained from A-B with the asymptotic approximation AA for the case of Poisson arrivals. Like our experiments for Weibull and lognormal service times, we take the mean service time equal to 2. The results are given in Table 8 . We use 1,000,000 replications for A-B to get accurate estimates. From the Tables 4, 6 and 8 we can clearly see that AA gives better approximations for Pareto service times than for Weibull and lognormal service times. Hence the need for fast simulation techniques in this case is less essential than in the Weibull or lognormal cases.

8 Further research directions

We are currently trying to extend the algorithms and results of this paper to queues with Markov modulated arrival processes, as well as to the estimation of other probabilities in the insurance risk context, for example, the estimation of the finite horizon ruin probability $\psi(u,T)$ (see Section 2.2).

A Appendix: Proof of Theorem 4.1 and Proposition 4.2

Let $X_{(1)}^{k,n} \leq \cdots \leq X_{(k)}^{k,n}$ be the order statistics of the service times conditioned on the event A_n^k. Conditioned on the event A_n^k, we can write the likelihood-ratio Z as

$$Z = (1+w_u)^k \left(\frac{\bar{F}_{\theta_u}(x_u^\star)}{(1-\theta_u)(1+w_u-F(x_u^\star))} \right)^n e^{-\theta_u \left(\sum_{i=k-n+1}^k \Lambda\left(X_{(i)}^{k,n}\right)\right)}, \tag{28}$$

using (15). To derive a deterministic upper bound on Z we have to derive a deterministic lower bound on

$$D_n^k := \sum_{i=k-n+1}^{k} \Lambda\left(X_{(i)}^{k,n}\right).$$

Let $\underline{x}$ be the minimum value after which $\lambda(x)$ is decreasing.

Lemma A.1 *If Assumption 2 holds, then for $x_u^\star \geq \underline{x}$,*

$$D_n^k \geq (n-1)\Lambda(x_u^\star) + \Lambda\left(\{u-kx_u^\star\}^+ + x_u^\star\right). \tag{29}$$

Proof. It is clear that

$$D_n^k \geq \min_{\substack{\sum_{i=1}^k z_i \geq u \\ z_1,\ldots,z_{k-n}\in[0,x_u^\star], z_{k-n+1},\ldots,z_k \geq x_u^\star}} \sum_{i=k-n+1}^{k} \Lambda(z_i) \geq \min_{\substack{\sum_{i=k-n+1}^k z_i \geq u-(k-n)x_u^\star \\ z_{k-n+1},\ldots,z_k \geq x_u^\star}} \sum_{i=k-n+1}^{k} \Lambda(z_i) =: \bar{D}_n^k. \tag{30}$$

It is actually insightful to consider $\sum_{i=k-n+1}^k \Lambda(z_i)$ as a cost function that you have to minimize under some constraints. In this case the cost function is a sum of n increasing and concave (due to Assumption 2) cost functions $\Lambda_{n-k+1},\ldots,\Lambda_k$, all of them being identical to Λ. Now there can be two cases.

Case 1: $u-(k-n)x_u^\star \leq nx_u^\star$ or $u-kx_u^\star \leq 0$

Since the cost functions $\Lambda_{k-n+1},\ldots,\Lambda_k$ are increasing, it is clear that the optimal solution is to set $z_{k-n+1} = \cdots = z_k = x_u^\star$. Then the contraint $\sum_{i=k-n+1}^k z_i \geq u-(k-n)x_u^\star$ is automatically satisfied. Thus the $\bar{D}_n^k$ is $n\Lambda(x_u^\star) = (n-1)\Lambda(x_u^\star) + \Lambda(x_u^\star)$.

Case 2: $u-(k-n)x_u^\star \geq nx_u^\star$ or $u-kx_u^\star \geq 0$

Since the cost functions $\Lambda_{k-n+1},\ldots,\Lambda_k$ are increasing, the optimal solutions will satisfy $\sum_{i=k-n+1}^k z_i = u-(k-n)x_u^\star \geq nx_u^\star$. Since the cost functions $\Lambda_{k-n+1},\ldots,\Lambda_k$ are identical, concave and increasing, it is clear that one of the optimal solutions is to set $z_{k-n+1} = \cdots = z_{k-1} = x_u^\star$ and to set z_k to be the rest, i.e., $z_k = u - (k-n)x_u^\star - (n-1)x_u^\star = u-(k-1)x_u^\star$. Thus the $\bar{D}_n^k$ is $(n-1)\Lambda(x_u^\star) + \Lambda\left(u-(k-1)x_u^\star\right) = (n-1)\Lambda(x_u^\star) + \Lambda\left(u - kx_u^\star + x_u^\star\right)$

One can combine the expressions for Case 1 and Case 2 and write $\bar{D}_n^k$ as $(n-1)\Lambda(x_u^\star) + \Lambda\left(\{u-kx_u^\star\}^+ + x_u^\star\right)$. Hence from (30), we get (29). □

Lemma A.2 *For u large enough and $k \leq k_0(u)$, the expectation $E[ZI(A^k)]$ can be upper bounded by*

$$E[ZI(A^k)] \leq (1+w_u)^k \frac{\Lambda(u)}{w_u + \Lambda(u)^{-b}} e^{-(1-\theta_u)\Lambda(x_u^\star)} e^{-\theta_u \Lambda\left(\{u-k_0(u)x_u^\star\}^+ + x_u^\star\right)} P^{(u)}(\tau(u)=k)P(\tau(u)<\infty). \tag{31}$$

Proof. Using Lemma A.1, we find

$$E[ZI(A_n^k)] \leq (1+w_u)^k \left(\frac{\bar{F}_{\theta_u}(x_u^\star)}{(1-\theta_u)(1+w_u-F(x_u^\star))}\right)^n e^{-\theta_u\left((n-1)\Lambda(x_u^\star) + \Lambda\left(\{u-kx_u^\star\}^+ + x_u^\star\right)\right)} P(A_n^k). \tag{32}$$

Some rewriting gives

$$\frac{\bar{F}_{\theta_u}(x_u^\star)e^{-\theta_u\Lambda(x_u^\star)}}{(1-\theta_u)(1+w_u-F(x_u^\star))} = \frac{e^{-(1-\theta_u)\Lambda(x_u^\star)}\Lambda(u)e^{-\theta_u\Lambda(x_u^\star)}}{w_u + e^{-\Lambda(x_u^\star)}} = \frac{\Lambda(u)^{-b+1}}{w_u + \Lambda(u)^{-b}}. \tag{33}$$

Using (33), for $n \geq 1$ and u large enough

$$E[ZI(A_n^k)] \leq (1+w_u)^k \left[\frac{\bar{F}_{\theta_u}(x_u^\star)e^{-\theta_u\Lambda(x_u^\star)}}{(1-\theta_u)(1+w_u-F(x_u^\star))}\right] e^{-\theta_u\Lambda\left(\{u-kx_u^\star\}^+ + x_u^\star\right)} e^{\theta_u\Lambda(x_u^\star)} P(A_n^k), \tag{34}$$

since the right-hand side of (33) goes to zero as u goes to infinity, because of Assumption 3. Using (33) in (34) and using the fact that $\Lambda(u)^{-b} = e^{-\Lambda(x_u^\star)}$ (by definition of $x_u^\star$), we get

$$E[ZI(A_n^k)] \leq (1+w_u)^k \frac{\Lambda(u)}{w_u + \Lambda(u)^{-b}} e^{-(1-\theta_u)\Lambda(x_u^\star)} e^{-\theta_u\Lambda\left(\{u-kx_u^\star\}^+ + x_u^\star\right)} P(A_n^k). \tag{35}$$

Summing up the left-hand and right-hand side of (35) from $n=1$ to $n=k$ and upper bounding $P(A^k)$ by $P(\tau(u)=k) = P^{(u)}(\tau(u)=k)P(\tau(u)<\infty)$ yield

$$E[ZI(A^k)] \leq (1+w_u)^k \frac{\Lambda(u)}{w_u + \Lambda(u)^{-b}} e^{-(1-\theta_u)\Lambda(x_u^\star)} e^{-\theta_u\Lambda\left(\{u-kx_u^\star\}^+ + x_u^\star\right)} P^{(u)}(\tau(u)=k)P(\tau(u)<\infty).$$

Finally, for $k \leq k_0(u)$, we get (31). □

Lemma A.3 *For u large enough,*

$$E[ZI(A)I(\tau(u) \leq k_0(u))] \leq \frac{a(u)}{\mu c_1 \delta^{c_1}} \Lambda(u) e^{-\theta_u\Lambda\left(\{u-k_0(u)x_u^\star\}^+ + x_u^\star\right)} P(\tau(u)<\infty). \tag{36}$$

Proof. From (31), the fact that $\Lambda(u)/(w_u + \Lambda(u)^{-b}) \leq \Lambda(u)/w_u = \Lambda(u)a(u)/c_1\mu$ (using $w_u > 0$ and $\Lambda(u)^{-b} > 0$) and the fact that $e^{-(1-\theta_u)\Lambda(x_u^\star)} \leq 1$, one finds

$$\begin{aligned}
& E[ZI(A)I(\tau(u) \leq k_0(u))] \\
\leq \quad & \frac{a(u)}{c_1\mu}\Lambda(u)P(\tau(u) < \infty)\sum_{k=1}^{k_0(u)} e^{-\theta_u\Lambda\left(\{u-k_0(u)x_u^\star\}^+ + x_u^\star\right)}(1+w_u)^k P^{(u)}(\tau(u) = k) \\
\leq \quad & \frac{a(u)}{c_1\mu}\Lambda(u)P(\tau(u) < \infty)e^{-\theta_u\Lambda\left(\{u-k_0(u)x_u^\star\}^+ + x_u^\star\right)}(1+w_u)^{k_0(u)}P^{(u)}(\tau(u) \leq k_0(u)). \qquad (37)
\end{aligned}$$

Then (36) follows from the fact that $P^{(u)}(\tau(u) \leq k_0(u)) \leq 1$ and

$$(1+w_u)^{k_0(u)} = \left(1 + \frac{c_1\mu}{a(u)}\right)^{-\frac{a(u)\log\delta}{\mu}} = \left\{\left(1 + \frac{c_1\mu}{a(u)}\right)^{\frac{a(u)}{c_1\mu}}\right\}^{-c_1\log\delta} \leq e^{-c_1\log\delta} = \delta^{-c_1}$$

The last inequality follows because $(1 + x^{-1})^x \leq e$ for $x > 0$. □

Proof of Proposition 4.2. For (i), note that from Lemma A.3 and (5), for u large enough,

$$E[ZI(A)I(\tau(u) \leq k_0(u))] \leq \frac{2\rho}{(1-\rho)\mu c_1\delta^{c_1}}a(u)\Lambda(u)e^{-\Lambda_I(u)}e^{-\theta_u\Lambda\left(\{u-k_0(u)x_u^\star\}^+ + x_u^\star\right)}. \qquad (38)$$

Then from Assumption 4, we have that $u - k_0(u)x_u^\star > 0$ for u large enough, proving (i). For (ii), note that for u large enough,

$$x_u^\star k_0(u) = \frac{-x_u^\star a(u)\log\delta}{\mu} < u,$$

because of Assumption 4. This implies that for u large enough,

$$A' \cap \{\tau(u) \leq k_0(u)\} = \emptyset. \qquad (39)$$

□

Proof of Theorem 4.1. First we prove that

$$\lim_{u\to\infty}\frac{\Lambda(u - x_u^\star k_0(u) + x_u^\star)}{\Lambda(u)} = 1, \qquad (40)$$

Since $\lambda(x)$ is decreasing for $x \geq \underline{x}$, $\Lambda(x)$ is concave for $x \geq \underline{x}$. Also, using Assumption 4 (which says that $x_u^\star k_0(u)/u$ goes to zero as u goes to infinity), $u - x_u^\star k_0(u) + x_u^\star \geq \underline{x}$ for u large enough. Hence,

$$\Lambda(u) \geq \Lambda\left(u - x_u^\star k_0(u) + x_u^\star\right) \geq \Lambda(\underline{x}) + \frac{(u - x_u^\star k_0(u) + x_u^\star) - \underline{x}}{u - \underline{x}}(\Lambda(u) - \Lambda(\underline{x})). \qquad (41)$$

The second inequality in (41) says that due to concavity, the slope of the tangent of the function $\Lambda(u)$ between $\underline{x}$ and $u - x_u^\star k_0(u) + x_u^\star$ is at least as high as the slope of the tangent of the function from $\underline{x}$ to u. Dividing (41) throughout by $\Lambda(u)$, letting $u \to \infty$, and using the fact that $x_u^\star k_0(u)/u$ goes to zero as $u \to \infty$, we get (40).

From (5) we see that $\gamma(u) \geq \tilde{K_1} \exp(-\Lambda_I(u))\rho/(1-\rho)$ for some positive constant $\tilde{K_1}$ and u large enough. Also $\text{work}(u) \leq \tilde{K_2} k_0(u) \leq K_2 a(u)$ where K_2 and $\tilde{K_2}$ are constants. Using (40) and Proposition 4.2, it follows that for u large enough,

$$\begin{aligned} & \frac{\log\left(\text{work}(u) \times \text{Var}\left[\hat{\gamma}(u)\right]\right)}{\log(\gamma^2(u))} \\ \geq \quad & \frac{\log K_2 + 2\log a(u) + \log K_1 + \log\Lambda(u) - \Lambda_I(u) - (1-\Lambda^{-1}(u))\Lambda(u - x_u^\star k_0(u) + x_u^\star)}{2\left[\log\tilde{K_1} + \log(\rho/(1-\rho)) - \Lambda_I(u)\right]} \\ \sim \quad & \frac{2\log a(u) + \log\Lambda(u) - \Lambda_I(u) - \Lambda(u)}{-2\Lambda_I(u)} \sim 1. \end{aligned} \tag{42}$$

To prove the last limit, all we need to show is that

$$\frac{\Lambda(u)}{\Lambda_I(u)} \to 1, \ \frac{\log\Lambda(u)}{\Lambda_I(u)} \to 0, \text{ and } \frac{\log a(u)}{\Lambda_I(u)} \to 0,$$

as $u \to \infty$. Assumption 3 implies that $\lim_{u\to\infty} \Lambda(u)^{-b+1} a(u) = 0$ and hence $\lim_{u\to\infty} a(u)^{(b-1)^{-1}}/\Lambda(u) = 0$. From the last limit we can conclude that

$$\lim_{u\to\infty} \frac{\log a(u)}{\Lambda(u)} = 0. \tag{43}$$

From (43) and the fact that

$$\frac{\log a(u)}{\Lambda(u)} \sim \frac{\Lambda(u) - \Lambda_I(u) + \log E[X]}{\Lambda(u)} = 1 - \frac{\Lambda_I(u)}{\Lambda(u)} + \frac{\log E[X]}{\Lambda(u)},$$

we see that

$$\lim_{u\to\infty} \frac{\Lambda(u)}{\Lambda_I(u)} = 1. \tag{44}$$

From (43) and (44) we conclude that $\lim_{u\to\infty} \log a(u)/\Lambda_I(u) = 0$. Finally, from (44), we get $\lim_{u\to\infty} \log\Lambda(u)/\Lambda_I(u) \to 0$. □

B Appendix: Justification for using Algorithm 1 for the case of log-normal service times

We will need to use the notation and the lemmas of Appendix A.

First, we show that Assumption 4 is satisfied. Using l'Hospital's rule, we find as in [27],

$$\lim_{x\to\infty} \frac{2\sigma^2\Lambda(x)}{\log^2(x)} = \lim_{x\to\infty} \frac{\sigma^2 x\lambda(x)}{\log x} = 1.$$

Thus

$$\Lambda(x) \sim \log^2(x)/2\sigma^2. \tag{45}$$

Recall that the auxiliary function $a(u)$ is given by

$$a(u) = \frac{\sigma^2 u}{\log u - \alpha}. \tag{46}$$

From (45) and (22) we find

$$x_u^\star \sim e^{\sqrt{2\sigma^2 b \log\left[\frac{\log^2(u)}{2\sigma^2}\right]}}. \tag{47}$$

Since for fixed $\epsilon > 0$ and for x large enough, $\sqrt{x} \leq \epsilon x$, we have also for u large enough,

$$\sqrt{2\sigma^2 b \log\left[\frac{\log^2(u)}{2\sigma^2}\right]} \leq \epsilon 2\sigma^2 b \log\left[\frac{\log^2(u)}{2\sigma^2}\right] = \log\left[\frac{\log^2(u)}{2\sigma^2}\right]^{2\sigma^2 b\epsilon}.$$

It follows that

$$e^{\sqrt{2\sigma^2 b \log\left[\frac{\log^2(u)}{2\sigma^2}\right]}} \leq \left[\frac{\log^2(u)}{2\sigma^2}\right]^{2\sigma^2 b\epsilon} = \frac{\log^{4\sigma^2 b\epsilon}(u)}{(2\sigma^2)^{2\sigma^2 b\epsilon}}. \tag{48}$$

Substituting (48) in (47), it follows from (47) and (46) that for u large enough,

$$\frac{a(u)x_u^\star}{u} \leq \frac{\sigma^2 \log^{4\sigma^2 b\epsilon}(u)}{(\log u - \alpha)\,(2\sigma^2)^{2\sigma^2 b\epsilon}} \to 0 \ (u \to \infty), \tag{49}$$

for $\epsilon < (4\sigma^2 b)^{-1}$. Thus Assumption 4 is satisfied. It is easy to check that Assumption 3 does not hold.

We still heuristically argue that Algorithm 1 works for lognormal service times. From (32) and (33),

$$\begin{aligned} E[Z \mid A_n^k] &= (1+w_u)^k \left[\frac{\Lambda(u)^{-b+1}}{w_u + \Lambda(u)^{-b}}\right]^n e^{\theta_u \Lambda(x_u^\star)} e^{-\theta_u \Lambda\left(\{u - kx_u^\star\}^+ + x_u^\star\right)} \\ &\approx (1+w_u)^k\, \Lambda(u)^n e^{\theta_u \Lambda(x_u^\star)} e^{-\theta_u \Lambda\left(\{u - kx_u^\star\}^+ + x_u^\star\right)}, \end{aligned} \tag{50}$$

since for large u, $\Lambda(u) \gg w_u$. From (50) it seems plausible to take $w_u \equiv 0$ and we obtain for u large enough

$$E[Z \mid A_n^k] \leq \Lambda(u)^n e^{\theta_u \Lambda(x_u^\star)} e^{-\theta_u \Lambda\left(\{u - kx_u^\star\}^+ + x_u^\star\right)}.$$

We find for u large enough,

$$\begin{aligned} E[I(A^k)Z] &= \sum_{n=1}^k E[Z \mid A_n^k] P(A_n^k) \\ &\leq \sum_{n=1}^k \Lambda(u)^n e^{\theta_u \Lambda(x_u^\star)} e^{-\theta_u \Lambda\left(\{u - kx_u^\star\}^+ + x_u^\star\right)} P(A_n^k) \\ &= e^{-\theta_u \Lambda\left(\{u - kx_u^\star\}^+ + x_u^\star\right)} \sum_{n=1}^k \Lambda(u)^{n + b\theta_u} P(A_n^k). \end{aligned}$$

Hence,

$$\begin{aligned} E[ZI(A)I(\tau(u) \leq k_0(u))] &= \sum_{k=1}^{k_0(u)} E[I(A^k)Z] \\ &\leq \sum_{k=1}^{k_0(u)} e^{-\theta_u \Lambda\left(\{u-kx_u^\star\}^+ + x_u^\star\right)} \sum_{n=1}^{k} \Lambda(u)^{n+b\theta_u} P(A_n^k). \end{aligned} \tag{51}$$

Similar as in the proof of Theorem 4.1, we can show with (49) and (51) that for u large enough,

$$E[ZI(A)I(\tau(u) \leq k_0(u))] \leq P(\tau(u) < \infty) e^{-\theta_u \Lambda(u)} \sum_{k=1}^{k_0(u)} \sum_{n=1}^{k} \Lambda(u)^{n+b\theta_u} P^{(u)}(A_n^k).$$

Hence, we have to keep the term

$$\sum_{k=1}^{k_0(u)} \sum_{n=1}^{k} \Lambda(u)^{n+b\theta_u} P^{(u)}(A_n^k)$$

as small as possible. Since $\Lambda(u)^{n+b\theta_u}$ is blowing up for growing n, we want $P^{(u)}(A_n^k)$ to decay fast. Note that this is not a problem in the Weibull case, since there we do not have problems with factors like $\Lambda(u)^{n+b\theta_u}$ because Assumption 4 holds for Weibull service times. One can increase the decay of $P^{(u)}(A_n^k)$ by raising b. However, as a side effect, $\Lambda(u)^{n+b\theta_u}$ grows faster in that case. So there is a kind of trade-off in choosing b. Unfortunately, we do not know anything about $P^{(u)}(A_n^k)$, since this is a very complicated probability, which requires detailed insights in the probabilistic behavior of the random walk (M_n).

Acknowledgments

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

References

[1] Abate, J., G.L. Choudhury and W. Whitt, "Waiting-time probabilities in queues with long-tail service-time distributions", *Queueing Systems* 16 (1994), 311-338.

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

[3] Asmussen, S., *Ruin Probabilities*, World Scientific, Singapore, New Jersey, London, Hong Kong (2000).

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

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

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

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

[8] Asmussen, S., P. Frantz, M. Jobmann, H.P. Schwefel, "Large deviations and fast simulation in the presence of boundaries", working paper (2000), see http://www.maths.lth.se/matstat/staff/asmus/pspapers.html.

[9] Barlow, R.E., and F. Proschan, *Statistical Theory of Reliability and Life Testing*, Holt, Reinhart and Winston, Inc. (1975).

[10] Boxma, O.J. and Cohen J.W., "Heavy-traffic analysis for the GI/G/1 queue with heavy-tailed distributions", *Queueing systems* 33 (1-3) 1999, 177-204.

[11] Boxma, O.J. and Cohen J.W., "The M/G/1 queue with heavy-tailed service time distribution", *IEEE Journal on Selected Areas in Communications* 16 (5) 1998, 749-763.

[12] Boxma, O.J., and V. Dumas, "Fluid queues with long-tailed activity period distributions", *Computer Communications*, 21 (1998), 1509-1529.

[13] Bucklew, J.A., *Large Deviations Techniques in Decision, Simulation, and Estimation*, John Wiley & Sons, Inc. (1990).

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

[15] Cohen, J.W. *The Single Server Queue*, North-Holland, Amsterdam (1982).

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

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

[18] Embrechts, P., C. Klüppelberg and T. Mikosch, *Modelling Extremal Events*, Springer-Verlag, Berlin Heidelberg (1997).

[19] Embrechts, P. and C. Klüppelberg , "Some aspects of insurance mathematics", *Theory Probab. Appl.*, 38 (2) 1993, 262-295.

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

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

[22] Feller, W., *An Introduction to Probability Theory and its Applications, Volume II*, John Wiley & Sons, Inc. (1966).

[23] Glasserman, P., P. Heidelberger, P. Shahabuddin and T. Zajic, "Multilevel splitting for estimating rare event probabilities ", *Operations Research*, 47, (1999), 585-600.

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

[25] Glynn, P.W., and Whitt, W., "The asymptotic efficiency of simulation estimators", *Operations Research*, 40 (1992), 505-520.

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

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

[28] Heidelberger, P., "Fast simulation of rare events in queueing and reliability models", *ACM Transactions on Modeling and Computer Simulation*, 6 (1995), 43-85.

[29] Lehtohnen, T., and H. Nyrhinen, "Simulating level-crossing probabilities by importance sampling", *Advances in Applied Probability*, 24 (1992), 858-874.

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

[31] Minh, D.L., and R. Sorli, "Simulating the GI/GI/1 queue in heavy traffic," *Operations Research* 31 (1983), 966-971.

[32] Pakes, A.G., "On the tails of waiting time distributions", *Journal of Applied Probability*, 12 (1975), 555-564.

[33] Parekh, S., and J. Walrand, "A quick simulation method for excessive backlogs in network of queues", *IEEE Transactions on Automatic Control*, 34 (1989), 54-56.

[34] Prabhu, N.U., *Stochastic storage processes: queues, insurance risk, dams and data communication*, Second Edition, Springer Verlag, New York (1998).

[35] Resnick, S., and G. Samorodnitsky, "A heavy traffic approximation for workload processes with heavy tailed service requirements", Research Report, Dept. of Operations Research and Industrial Engineering, Cornell University, (1999). To appear in *Management Science.*

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

[37] Shahabuddin, P., "Importance sampling for the simulation of highly reliable Markovian systems", *Management Science*, 40 (1994), 333-352.

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