THIELE CENTRE
FOR APPLIED MATHEMATICS IN NATURAL SCIENCE

Importance Sampling for Failure Probabilities in Computing and Data Transmission

Søren Asmussen

Research Report

No. 14 | November 2008

Importance Sampling for Failure Probabilities in Computing and Data Transmission

Søren Asmussen[*]

Abstract

We study efficient simulation algorithms for estimating $\mathbb{P}(X > x)$, where X is the total time of a job with ideal time T that needs to be restarted after a failure. The main tool is importance sampling where one tries to identify a good importance distribution via an asymptotic description of the conditional distribution of T given $X > x$.

If $T \equiv t$ is constant, the problem reduces to the efficient simulation of geometric sums, and a standard algorithm involving a Cramér type root $\gamma(t)$ is available. However, we also discuss an algorithm avoiding the rootfinding. If T is random, particular attention is given to T having either a gamma-like tail or a regularly varying tail, and to failures at Poisson times. Different type of conditional limits occur, in particular exponentially tilted Gumbel distributions and Pareto distributions. The algorithms based upon importance distributions for T using these asymptotical descriptions have bounded relative error as $x \to \infty$ when combined with the ideas used for a fixed t.

Nevertheless, the paper gives examples that algorithms carefully designed to enjoy bounded relative error may provide little or no asymptotic improvement of crude Monte Carlo simulation when the computational effort is taken into account. To resolve this problem, an alternative algorithm using two-sided Lundberg bounds is suggested.

Key words and phrases communications engineering, compound sum, computer reliability, conditioned limit theorem, Cramér root, exponential tilting, geometric sum, Gumbel distribution, integral asymptotics, Lundberg's inequality, rare event simulation, regular variation, RESTART.

[*] Dept. of Mathematical Sciences, Aarhus University, Ny Munkegade, DK-8000 Aarhus C, Denmark; `asmus@imf.au.dk, http://home.imf.au.dk/asmus/`

1 Introduction

Consider a task of length T that is subject to failures and must be restarted if a failure occurs before completion. For example, the task may be the execution of a computer program, the transmission of a file on a communications channel, or a conversation with a call center.

The distribution of the (ideal) task time T is throughout denoted by F, and the distribution of the failure time U by G. For convenience, the densities f, g are asssumed to exist except when $T \equiv t$ is constant. Due to the possibility of (multiple) failures, the total task time X can possibly be large (certainly, we always have $X \geq T$). We are here interested in the distribution H of X, more specifically in its tail $\overline{H}(x) = \mathbb{P}(X > x)$.

This problem has a long history in computer science where the model goes under the name of RESTART (see [5] for references). Nevertheless, a comprehensive description of the tail asymptotics of X was only recently provided by Sheahan *et al.* [16] and Asmussen *et al.* [5]. At about the same time (in part independently), Jelenković & Tan [14], [15] performed a related study in the communications engineering context; a main difference from [5] is an on-off assumption on the channel, which in the computer reliability context corresponds to incorporating repair times. Further aspects involve parallel computing, [2], and checkpointing (fragmentation), [6].

In the early work of Sheahan *et al.* [16], a numerical comparison of approximations and simulated values was performed. This turned out to be a computationally extremely demanding task, since $R = 10^8$ independent copies of X were needed to be generated to obtain sufficiently precise estimates of $\mathbb{P}(X > x)$ in the range of x-values under study.[1]

The present paper suggests and analyzes some more sophisticated algorithms designed to reduce the computational effort. Given the literature on rare event simulation (surveyed in, e.g., [4] Ch. VI), it is not unexpected that importance sampling is the main tool (though other ideas like conditional Monte Carlo and splitting have been used for specific purposes, see again *loc. cit.*). The classical idea when using importance sampling is to look for an asymptotical description of the conditional distribution given the rare event and use this as importance distribution. This is also the path we follow here and leads to some additional theoretical problems on the model, since we must analyze such problems as how failures accumulate within a long but fixed time horizon, and what is the asymptotics as $x \to \infty$ of T given that $X > x$. We will see some rather non-standard limit distributions arise.

For most applications, it would be of particular interest to assume G to be exponential, say at rate μ, and F to be either degenerate (say at t), gamma-like in the sense that

$$f(t) \sim ct^{\alpha-1}\mathrm{e}^{-\lambda t}, \quad t \to \infty, \tag{1.1}$$

(this incorporates as a special case the three distributions in the numerical example of [16]) or of power-form in the sense that $\log f(t)/t \to -\alpha - 1$; this covers as a

[1] [16] says $R = 10^6$ but this is a typo.

special case a regularly varying f,

$$f(t) \sim \frac{L(t)}{t^{\alpha+1}}, \quad t \to \infty, \tag{1.2}$$

with L slowly varying. We shall therefore pay particular attention to these specific cases.

The paper is organized as follows. In Section 2, we give the relevant preliminaries both on RESTART and rare event simulation. In particular, a crucial quantity for the rest of the paper is introduced, a Cramér-type root $\gamma(t)$. Section 3 and Appendix B studies the simulation problem when $T \equiv t$ is deterministic. This is fairly standard in its simplest formulation since as surveyed in Section 2, $X - T = X - t$ then admits a geometric sum representation, and it is folklore that the simulation of tails of light-tailed geometric sums is most efficiently carried out by exponential tilting; in the RESTART setting, this means involving $\gamma(t)$. However, we also discuss to which extent the evaluation of $\gamma(t)$ can be avoided.

The rest of the paper deals with the case of a random T. The asymptotic results of [5] exhibit a great diversity depending on the specific form of the tails of F and G, and for this reason one has to expect the same to be the case for the form of efficient rare event simulation algorithms. We consider two cases, in both of which G is taken to be exponential(μ). Section 4 studies the gamma-like case (1.1). Motivated from general principles for rare event simulation, the asymptotic behavior of T given $X > x$ is studied, and after appropriate centering, we obtain a non-standard limit, the exponentially tilted Gumbel distribution Q_β. The use of this as importance distribution is discussed, and an important message is that importance sampling on T alone is only modestly efficient — to do better, one has to combine with the more sophisticated algorithms for geometric sums.

In Section 5, a similar discussion is carried out for the regularly varying case (1.2). Here T given $X > x$ needs to be both centered and scaled (not just centered), and the limit is Pareto. However, using the Pareto (shifted and scaled back to T) as importance distribution one encounters an absolute continuity problem. This is resolved by combining with another importance sampling algorithm.

A maybe surprising feature of these algorithms is that even if the distribution of X is always heavy-tailed when T has unbounded support, then the ideas all come from the light-tailed area; in general, the methodologies for simulation of light versus heavy tails are intrinsically different, cf. [4] Ch. VI.

The algorithms just outlined enjoy bounded relative error, a concept in the center of the rare event simulation literature (the definition is given in Section 2.2), and generally considered to represent the ultimate improvement of crude Monte Carlo simulation one can hope for. However, focusing solely on bounded relative error as efficiency measure is misleading — one needs to consider also the computational effort. This is done in Section 6, and a considerably more diverse picture emerges. A partial solution to the problem based upon two-sided Lundberg type bounds is suggested in Section 7. Finally, Section 8 contains some numerical examples.

2 Preliminaries

2.1 The RESTART model

Consider a deterministic $T \equiv t$ and let $X(t)$ be the corresponding simple RESTART total time, $H_t(x) = \mathbb{P}(X(t) \le x)$.

As in [5], we can write $X(t) = t + S(t)$ where $S(t) = \sum_1^N U_i(t)$ is a geometric sum: $N, U_1(t), U_2(t), \ldots$ are independent such that $\mathbb{P}(N = n) = (1-\rho)\rho^n$ with $\rho = G(t)$, and the $U_i(t)$ have the distribution $G_{|t}$ defined as G conditioned to $(0,t)$. That is, the c.d.f. is $\mathbb{P}(U_i(t) \le s) = G(s)/G(t)$ for $s \le t$, $\mathbb{P}(U_i(t) \le s) = 1$ for $s > t$, and the density is $g(s)/G(t)$ for $s \le t$, 0 for $s > t$. By general theory for geometric sums, [17] (see also [5]), we know that

$$\mathbb{P}\bigl(S(t) > x\bigr) \sim C_1(t)\mathrm{e}^{-\gamma(t)x}\,, \tag{2.1}$$

where $\gamma(t)$ is the solution of

$$1 = \int_0^t \mathrm{e}^{\gamma u} g(u)\,\mathrm{d}u \tag{2.2}$$

and

$$C_1(t) = \frac{\overline{G}(t)}{\gamma(t)m(t)} \quad \text{where } m(t) = \int_0^t u\mathrm{e}^{\gamma(t)u} g(u)\,\mathrm{d}u\,. \tag{2.3}$$

Since $\mathbb{P}\bigl(X(t) > x\bigr) = \mathbb{P}\bigl(S(t) > x - t\bigr)$, we therefore have

$$\overline{H}_t(x) = \mathbb{P}\bigl(X(t) > x\bigr) \sim C_2(t)\mathrm{e}^{-\gamma(t)x} \quad \text{where } C_2(t) = \mathrm{e}^{\gamma(t)t}C_1(t)\,. \tag{2.4}$$

From [5], we also quote the two-sided Lundberg inequality:

$$\mathrm{e}^{-\gamma(t)x} \le \overline{H}_t(x) \le \mathrm{e}^{\gamma(t)t}\mathrm{e}^{-\gamma(t)x}\,. \tag{2.5}$$

It is shown in [5] that for a general G, $\gamma(t) \sim \mu\overline{G}(t)$ as $t \to \infty$ where μ is the mean of G. For the exponential case, we shall need certain refinements and related results that are proved/collected in Appendix A. In particular:

$$\mu\mathrm{e}^{-\mu t} \le \gamma(t) = \mu\mathrm{e}^{-\mu t} + \mu^2 t\mathrm{e}^{-2\mu t} + \mathrm{o}(t\mathrm{e}^{-2\mu t}) \quad \text{as } t \to \infty\,, \tag{2.6}$$

$$\gamma(t) = -\mu \log t/t\bigl(1 + \mathrm{o}(1)\bigr) \quad \text{as } t \downarrow 0\,. \tag{2.7}$$

2.2 Rare Event Simulation

Consider the probability $z(x)$ of an event $A(x)$ (in our case, $\{X > x\}$) that is rare in the sense that $z(x) \to 0$ as $x \to \infty$. As in [4], we denote by an estimator for $z(x)$ a r.v. $Z(x)$ that can be generated by simulation and is unbiased, $\mathbb{E}Z(x) = z(x)$. A family $\bigl\{Z(x)\bigr\}_{x>0}$ of such estimators (or just $Z(x)$) is said to have bounded relative error if $\mathbb{V}\mathrm{ar}Z(x) = \mathrm{O}\bigl(z(x)^2\bigr)$ as $x \to \infty$, and to be logarithmically efficient if $\mathbb{V}\mathrm{ar}Z(x) = \mathrm{O}\bigl(z(x)^{2-\varepsilon}\bigr)$ for all $\varepsilon > 0$ (cf. [4] p. 159). In practice, the estimate of $z(x)$ for a given x is obtained by averaging R replications of $Z(x)$, and Gaussian confidence intervals can be produced in a standard way by computing the empirical variance.

If we (in a non-standard terminology!) define the *logarithmic efficiency factor* of an estimator $Z(x)$ as

$$\sup\Bigl\{p > 0 : \frac{\mathbb{V}\mathrm{ar}\, Z(x)}{z(x)^p} \to 0\Bigr\},$$

then crude Monte Carlo method has logarithmic efficiency factor 1 and an estimator that is logarithmically efficient or has bounded relative error has logarithmic efficiency factor at least 2.

The traditional approach to exhibiting estimators with logarithmic efficiency factor > 1 via importance sampling is to provide an asymptotic description of the conditional distribution $\mathbb{P}\bigl(\cdot \mid A(x)\bigr)$ given the rare event $A(x)$, and to use this as importance distribution. The philosophy is that sampling from $\mathbb{P}\bigl(\cdot \mid A(x)\bigr)$ yields a zero-variance estimator, so that an importance distribution that is close hopefully has a small variance.

As already touched upon in Section 1, also computational effort needs to be taken into account; this is often neglected in the rare event simulation literature. We defer the discussion of this to Section 6.

3 Simulation Algorithms for a Deterministic $T \equiv t$

In this section, we discuss efficient algorithms for simulation of $z(x) = \mathbb{P}\bigl(S(t) > x\bigr)$ for a fixed t. One of them (Algorithm 1) has bounded relative error. Replacing x by $x - t$ gives algorithms with bounded relative error for simulation of $\overline{H}_t(x)$ (the case of a random T is the subject of the rest of the paper and requires more work). The other approach, Algorithm 2, is conceptually simpler and reduces variance with an exponential factor, but does not have bounded relative error.

We will allow G to be general, not necessarily exponential. We write

$$S_n = U_1 + \cdots + U_n\,, \quad \tau(x) = \inf\{n : S_n > x\}\,.$$

Recall from Section 2 that $G_{|t}$ denotes G conditioned to $(0, t)$, and define $G_{\gamma(t)}$ as the distribution on $(0, t)$ with density $g_{\gamma(t)}(y) = \mathrm{e}^{\gamma(t)y} g(y)$, $0 < y < t$.

The first algorithm is a special case of the one given in [4], Exercise 2.3 p. 172, for general geometric sums (see also Blanchet & Li [9]). An outline of the approach is given in Appendix B. One needs to determine a certain root and to define a corresponding exponentially tilted distribution. When specialized to the RESTART setting, it is easy to see that the root is precisely $\gamma(t)$ and that the exponentially tilted distribution bevomes $G_{\gamma(t)}$ (see Remark 8.1). This yields the following algorithm:

Algorithm 1 *Generate $U_1, U_2, \ldots$ from $G_{\gamma(t)}$. Stop the simulation at $\tau(x)$ and return the estimator $Z_1(x) = \mathrm{e}^{-\gamma(t)S_{\tau(x)}}$.*

From Appendix B we have at once:

Theorem 3.1 *The estimator $Z_1(x)$ is unbiased for $z(x)$ and has bounded relative error. That is, $\mathbb{V}\mathrm{ar}_{\gamma(t)}\, Z_1(x) = \mathrm{O}\bigl(z(x)^2\bigr)$ as $x \to \infty$.*

Random variate generation from $G_{\gamma(t)}$ as well as the rootfinding may sometimes be tedious. A simpler idea is to take advantage of the special feature of bounded support (that is not available for general geometric sums) and simulate using the distribution $G_{|t}$. This leads to:

Algorithm 2 *Generate* $U_1(t), U_2(t), \ldots$ *from* $G_{|t}$. *Stop the simulation at* $\tau(x)$ *and return the estimator* $Z_2(x) = G(t)^{\tau(x)}$.

Proposition 3.1 *The estimator* $Z_2(x)$ *is unbiased for* $z(x)$. *Further,* $\mathbb{V}\mathrm{ar}_{|t} Z_2(x)$ *is of order* $\mathrm{e}^{-(\gamma(t)+\xi(t))x}$ *where* $\xi(t)$ *is the solution of*

$$1 = G(t) \int_0^t \mathrm{e}^{(\gamma(t)+\xi(t))u} g(u)\, \mathrm{d}u \tag{3.1}$$

and satisfies $0 < \xi(t) < \gamma(t)$. *That is, the logaritmic efficiency factor is* $1 + \xi(t)/\gamma(t) \in (1, 2)$.

Proof. For $u < t$, we have

$$\mathbb{P}_{|t}(U_1 \in \mathrm{d}u) \;=\; \frac{g(u)\mathrm{d}u}{\overline{G}(t)} \;=\; \frac{\mathrm{e}^{-\gamma(t)u}}{\overline{G}(t)} \mathbb{P}_{\gamma(t)}(U_1 \in \mathrm{d}u)\,,$$

and it follows by a standard extension to stopping times (e.g., [4] pp. 131–132) that

$$\mathbb{E}_{|t} Z_2(x) = \mathbb{E}_{\gamma(t)}\Big[\frac{\mathrm{e}^{-\gamma(t)S_{\tau(x)}}}{\overline{G}(t)^{\tau(x)}} Z_2(x)\Big] = \mathbb{E}_{\gamma(t)} \mathrm{e}^{-\gamma(t)S_{\tau(x)}} = \mathbb{E}_{\gamma(t)} Z_1(x) = z(x)\,,$$

showing unbiasedness.

Since (3.1) can be rewritten as $1 = G(t)\mathbb{E}_{\gamma(t)} \mathrm{e}^{\xi(t)U_1}$, it follows in a similar way that

$$\begin{aligned}\mathbb{E}_{|t} Z_2(x)^2 = \mathbb{E}_{\gamma(t)}\Big[\frac{\mathrm{e}^{-\gamma(t)S_{\tau(x)}}}{\overline{G}(t)^{\tau(x)}} Z_2(x)^2\Big] \;&=\; \mathbb{E}_{\gamma(t)}\big[\mathrm{e}^{-\gamma(t)S_{\tau(x)}} G(t)^{\tau(x)}\big] \\ = \mathbb{E}_{\gamma(t)}\Big[\mathrm{e}^{-(\gamma(t)+\xi(t))S_{\tau(x)}} &\mathrm{e}^{\xi(t)S_{\tau(x)}} \big/ \big[\mathbb{E}_{\gamma(t)} \mathrm{e}^{\xi(t)U_1}\big]^{\tau(x)}\Big]\,.\end{aligned}$$

Using $|S_{\tau(x)} - x| \le t$ shows that this expression is bounded up and below by a constant times

$$\mathrm{e}^{-(\gamma(t)+\xi(t))x} \cdot \mathbb{E}_{\gamma(t)}\Big[\mathrm{e}^{\xi(t)S_{\tau(x)}} \big/ \big[\mathbb{E}_{\gamma(t)} \mathrm{e}^{\xi(t)U_1}\big]^{\tau(x)}\Big]\,.$$

But the expectation is the expectation of the Wald martingale stopped at $\tau(x)$. The condition for optional stopping ([3] p. 362) is trivially satisfied because by positivity, $\tau(x)$ is automatically finite for any exponential tilting of $\mathbb{P}_{\gamma(t)}$. Thus the expectation is indeed one, so that the order of $\mathbb{V}\mathrm{ar}_{|t} Z_2(x)$ is as asserted.

To complete the proof, it remains to show that $0 < \xi(t) < \gamma(t)$. Clearly, the r.h.s. of (3.1) is increasing in $\xi(t)$. The value at $\xi(t) = 0$ is $G(t) < 1$ because of the definition of $\gamma(t)$. This implies $\xi(t) > 0$. Similarly, $\xi(t) < \gamma(t)$ will follow if we can show that the value at $\gamma(t)$ is > 1. But this value is

$$G(t) \int_0^t \mathrm{e}^{2\gamma(t)u} g(u)\, \mathrm{d}u = G(t)^2 \mathbb{E}_{|t} \mathrm{e}^{2\gamma(t)U_1} > G(t)^2 \big[\mathbb{E}_{|t} \mathrm{e}^{\gamma(t)U_1}\big]^2 = 1\,,$$

where the last step used that the definition of $\gamma(t)$ can be rewritten

$$1=\int_0^t \mathrm{e}^{\gamma(t)u}g(u)\,\mathrm{d}u = G(t)\mathbb{E}_{|t}\mathrm{e}^{\gamma(t)U_1}\,. \qquad \square$$

The last part of Proposition 3.1 shows that indeed Algorithm 2 provides exponential variance reduction (at rate $\xi(t)$) but does not have bounded relative error (for this $\xi(t)\ge\gamma(t)$ would have needed). However, the loss of efficiency vanishes as $t\to\infty$:

Proposition 3.2 *Assume* $\widehat{G}[\epsilon]=\int_0^\infty \mathrm{e}^{\epsilon t}\,G(\mathrm{d}t)<\infty$ *for some* $\epsilon>0$. *Then* $\xi(t)\sim\gamma(t)\sim\mu\mathrm{e}^{-\mu t}$ *as* $t\to\infty$. *That is, the logarithmic efficiency factor of Algorithm 2 goes to* 2 *as* $t\to\infty$.

For the proof, see Appendix A.

Algorithm 2 is simpler than Algorithm 1 by avoiding the rootfinding and the exponential tilting. However, for G exponential the exponentially tilted distribution is truncated exponential. So, both algorithms require simulation from an exponential distribution truncated to $(0,t)$ (but with different parameters $\mu_1=\mu-\gamma(t), \mu_2=\mu$). This can easily be done by inversion: generate the r.v. as $-\log\bigl((1-\mathrm{e}^{-\mu_i t})V/\mu_i\bigr)$ with V uniform on $(0,1)$, cf. [4], Remark 2.4 p. 39. Another way is acceptance-rejection: use the exponential(μ_i) distribution as proposal and reject values $>t$.

4 Simulation Algorithms for a Gamma-Like T

If T is random, one expects a large X to occur as consequence of a large T. Thus the general principles of importance sampling surveyed in Section 2.2 suggest to look for the conditional distribution of T given $X>x$. Our result is:

Theorem 4.1 *Assume that* F *is gamma-like as in* (1.1). *Then the conditional distribution of* $Y=Y(x)=\mu T-\log x-\log\mu$ *given* $X>x$ *has a limit in distribution as* $x\to\infty$, *namely the distribution* Q_β *with density*

$$q(y)=\exp\bigl\{-\mathrm{e}^{-y}-\beta y\bigr\}/\Gamma(\beta)\,,\ -\infty<y<\infty \quad \textit{where } \beta=\lambda/\mu\,. \tag{4.1}$$

One simple message is that T given $X>x$ is of order $\log x/\mu$. When $\lambda=\mu$, $Q=Q_1$ is the Gumbel distribution familiar from extreme value theory (also known as the Fisher-Tippet distribution). The c.d.f. at y is $\exp\bigl\{-\mathrm{e}^{-y}\bigr\}$. When $\lambda\ne\mu$, Q_β is an exponentially tilted Gumbel distribution, and the properties are less standard. We return to this at the end of the section.

Proof of Theorem 4.1. Specializing Corollary 1.1 (or Theorem 2.2) of [5], we get

$$\mathbb{P}(X>x)\sim\frac{c\Gamma(\beta)}{\mu^{\alpha+\beta}}\,\frac{\log^{\alpha-1}x}{x^\beta}\,. \tag{4.2}$$

Let $f(t;x)$ be the density of T on the event $X>x$, that is,

$$f(t;x)\,\mathrm{d}t=\mathbb{P}(T\in\mathrm{d}t,\ X>x)\,,$$

and let $t(x,y) = (\log x + \log \mu + y)/\mu$. Then the density $q(y|x)$ at y of Y given $X > x$ is $f\bigl(t(x,y);x\bigr)/\mu\mathbb{P}(X > x)$. Using (2.6) gives

$$\gamma\bigl(t(x,y)\bigr) = \mathrm{e}^{-y}/x + \mathrm{O}(\log x/x^2)\,.$$

It follows then from the two-sided Lundberg inequality (2.5) that

$$\mathbb{P}\bigl(X > x \,\big|\, T = t(x,y)\bigr) \sim \exp\{-\mathrm{e}^{-y}\}\,,$$

and so

$$\begin{aligned} q(y|x) &= \frac{1}{\mu\mathbb{P}(X > x)} f\bigl(t(x,y);x\bigr) = \frac{1}{\mu\mathbb{P}(X > x)} f\bigl(t(x,y)\bigr)\mathbb{P}\bigl(X > x \,\big|\, T = t(x,y)\bigr) \\ &\sim \frac{\mu^{\alpha+\beta-1}}{c\Gamma(\beta)} \frac{x^\beta}{\log^{\alpha-1} x}\, ct(x,y)^{\alpha-1}\mathrm{e}^{-\lambda t(x,y)} \exp\{-\mathrm{e}^{-y}\} \\ &\sim \frac{\mu^{\alpha+\beta-1}}{c\Gamma(\beta)} \frac{x^\beta}{\log^{\alpha-1} x}\, c\frac{\log^{\alpha-1} x}{\mu^{\alpha-1}} x^{-\beta}\mu^{-\beta}\mathrm{e}^{-\beta y} \exp\{-\mathrm{e}^{-y}\} = q(y)\,. \end{aligned}$$

But Scheffé's theorem ([7]) states that convergence of densities implies convergence in distribution. □

Theorem 4.1 suggests that in the case of a gamma-like F as in (1.1), one should proceed as follows in order to simulate $z(x) = \mathbb{P}(X > x)$:

Algorithm 3 *Generate Y from the density q in (4.1) and let $T = t = (\log x + \log \mu + Y)/\mu$. If $T \le 0$, return the estimator $Z_3(x) = 0$. Otherwise calculate the likelihood ratio*

$$W = f(T)/\mu q(\mu T - \log x - \log \mu) = f(T)x^{-\beta}\mu^{-1-\beta}\Gamma(\beta)\exp\bigl\{\mu\mathrm{e}^{-\mu T}x + \lambda T\bigr\}\,,$$

compute the crude Monte Carlo estimator $Z_0(x-t)$ for $\mathbb{P}\bigl(S(t) > x-t\bigr) = \mathbb{P}(X(t) > x)$, and return the estimator $Z_3(x) = WZ_0(x-t)$.

The algorithm is motivated from the general principle of rare event simulation, that one should use a distribution close to the conditional distribution given the rare event (here $X > x$) as importance distribution, cf. [4] Example 1.3 p. 128. Indeed, the suggested importance distribution for T corresponds to the asymptotic description of this conditional distribution provided by Theorem 4.1 and the event $X > x$ is not rare when T is simulated from q. The following result shows that the algorithm indeed has a substantial smaller asymptotic variance than the Crude Monte Carlo method, but does not get close to bounded relative error or logarithmic efficency:

Proposition 4.1 *The estimator $Z_3(x)$ has logarithmic efficiency factor at most $3/2$, and exactly equal to $3/2$ provided $\int_0^{t_0} f(t)^2\,\mathrm{d}t < \infty$ for all $t_0 < \infty$.*

In the proof, we shall need the following analytical result:

Lemma 4.1 *For any $t_0 > 0$, $\displaystyle\int_{t_0}^{\infty} \exp\bigl\{-k\mathrm{e}^{-\eta t}x\bigr\}ct^{\delta-1}\mathrm{e}^{-\lambda t}\,\mathrm{d}t \sim \frac{\Gamma(\lambda/\eta)}{\eta^\delta k^{\lambda/\eta}}\frac{\log^{\delta-1} x}{x^{\lambda/\eta}}$ as $x \to \infty$.*

The Lemma is of the same type as a crucial step in the proof of (4.2) in [5], but since the proof is short, we reproduce it here: substituting $s = \mathrm{e}^{-\eta t}$, the integral becomes

$$\int_0^{\mathrm{e}^{-\eta t_0}} \mathrm{e}^{-ksx} \frac{(-\log s)^{\delta-1}}{\eta^\delta} s^{\lambda/\eta-1}\, \mathrm{d}s\,,$$

and Karamata's Tauberian theorem ([8] Theorems 1.5.11 and 1.7.1) implies that this has the asserted asymptotics. □

Proof of Proposition 4.1. From $\mathbb{E}Z_0(x-t)^2 = \mathbb{P}\big(S(t) > x-t\big) = \overline{H}_t(x)$, we get by conditioning upon $T = t$ that

$$\begin{aligned}
&\mathbb{E}Z_3(x)^2 \\
&\quad = \int_0^\infty \overline{H}_t(x) \frac{f(t)^2}{\mu^2 q(\mu t - \log x - \log \mu)^2} \mu q(\mu t - \log x - \log\mu)/\mu \,\mathrm{d}t && (4.3)\\
&\quad = x^{-\beta} \int_0^\infty \overline{H}_t(x) f(t)^2 \Gamma(\beta)\mu^{-1-\beta} \exp\big\{\mu \mathrm{e}^{-\mu t} x + \lambda t\big\}\,\mathrm{d}t && (4.4)\\
&\quad \ge k_1 x^{-\beta} \int_{t_0}^\infty \mathrm{e}^{-\gamma(t)x} t^{2\alpha-2} \mathrm{e}^{-2\lambda t} \exp\big\{\mu \mathrm{e}^{-\mu t}x + \lambda t\big\}\,\mathrm{d}t\\
&\quad \ge k_1 x^{-\beta} \int_{t_0}^\infty t^{2\alpha-2} \exp\big\{-\mathrm{O}\big(t\mathrm{e}^{-2\mu t}\big)x - \lambda t\big\}\,\mathrm{d}t\\
&\quad \ge k_1 x^{-\beta} \int_{t_0}^\infty t^{2\alpha-2} \exp\big\{-k_2(\epsilon)\mathrm{e}^{-(2-\epsilon)\mu t}x - \lambda t\big\}\,\mathrm{d}t\\
&\quad \sim k_3(\epsilon) \frac{\log^{2\alpha-2} x}{x^{\lambda/(2-\epsilon)\mu}} = k_3(\epsilon) x^{-\beta} \frac{\log^{2\alpha-2} x}{x^{\beta(1+1/(2-\epsilon))}}\,,
\end{aligned}$$

where we used the lower Lundberg bound in (2.5), the r.h. inequality in (2.6) and Lemma 4.1. Combining with (4.2) shows that the logarithmic efficiency factor is as most $1 + 1/(2-\epsilon)$ and therefore at most $3/2$.

For the lower bound, first note that the upper Lundberg bound implies that (4.4) can be bounded by

$$k_5 x^{-\beta} \int_0^\infty f(t)^2 \exp\big\{\psi(t,x) - \lambda t\big\}\,\mathrm{d}t$$

where $\psi(t,x) = \gamma(t)t - \gamma(t)x + \mu\mathrm{e}^{-\mu t}x$. Let I_1, I_2, denote the contributions to this integral from the intervals $0 < t \le t_0$, resp. $t > t_1$, where t_0, t_1 will be specified later. Then, with $k_7 = \sup_{t>t_0} \gamma(t)t$, we have by the r.h.s. of (2.6) that

$$\begin{aligned}
I_2 &\le k_6 \int_{t_1}^\infty t^{2\alpha-2} \exp\big\{k_7 - k_8 t\mathrm{e}^{-2\mu t}x - \lambda t\big\}\,\mathrm{d}t\\
&\le k_9 \int_{t_1}^\infty t^{2\alpha-2} \exp\big\{-k_8 t_1 \mathrm{e}^{-2\mu t}x - \lambda t\big\}\,\mathrm{d}t \sim k_{10} \frac{\log^{2\alpha-2} x}{x^{\lambda/2\mu}}\,.
\end{aligned}$$

For $x \ge 1$, we can bound I_1 by

$$\int_0^{t_0} f(t)^2 \exp\big\{\psi(t)x\big\}\,\mathrm{d}t$$

where $\psi(t) = \gamma(t)t - \gamma(t) + \mu \mathrm{e}^{-\mu t}$. Using (2.7) yields

$$\psi(t) \le -\mu \log t(1-1/t)\big(1+\mathrm{O}(1)\big) + \mu$$

as $t \downarrow 0$. This shows that is t_0 is small enough, then $\psi(t) < 0$ uniformly in $0 < t \le t_0$. Hence using the assumption on f^2 shows that I_1 goes to 0 exponentially fast as $x \to \infty$.

Replacing t_1 by a smaller value, we may assume $t_1 \le t_0$ and then (4.4) is bounded by $x^{-\beta}(I_1 + I_2)$, which in turn by the above estimates is $\mathrm{O}(x^{-\delta})$ for all $\delta < 3\beta/2$. This completes the proof. □

Remark 4.1 An essential ingredient of the proof is informally to replace $\mathbb{P}\big(S(t-x) > x\big)$ for a large t by its Cramér-Lundberg approximation $C_2(t)\mathrm{e}^{-\gamma(t)x}$, note that $C_2(t) \sim 1$ and $\gamma(t) \sim \mu \mathrm{e}^{-\mu t}$ as $t \to \infty$, so that the final approximation is $\exp\{-\mu \mathrm{e}^{-\mu t} x\}$; to justify this, Lundberg's inequality (and in part more refined estimates like (2.7)) were used.. The same procedure will be used later in the paper in Section 5, but since we have carefully given the details for the present case, we will not do so there. □

To improve Algorithm 3, we involve further properties of the conditional distribution given the rare event, namely the behaviour of $U_1(t), \dots, U_{N(t)}(t)$ as used in Algorithms 1, 2 (it is not apriori obvious that this will help since since the event $X > x$ is not rare when T is simulated from q!)

Algorithm 4 *Generate Y from the density q in* (4.1) *and let $T = t = (\log x + \log \mu + Y)/\mu$. If $T \le 0$, return $Z_4(x) = 0$. Otherwise calculate the likelihood ratio*

$$W \;=\; f(t)/\mu q(\mu t - \log x - \log \mu)\,,$$

compute one of the two estimators $Z_i(x-t)$ of Section 3 ($i = 1$ or 2), and return $Z_4(x) = W Z_i(x)$.

Theorem 4.2 *The estimator $Z_4(x)$ has bounded relative error provided $\int_0^{t_0} f(t)^2 \,\mathrm{d}t < \infty$ for all $t_0 < \infty$.*

Proof. The proof is a small variant of the last part of the proof of Proposition 4.1. Let first $i = 1$. From $\mathbb{E}Z_1(x-t)^2 \le \mathrm{e}^{-2\gamma(t)(x-t)}$, we get by conditioning upon $T = t$ and replacing $\overline{H}(t)$ by $\mathrm{e}^{-2\gamma(t)(x-t)}$ in (4.4) that

$$\mathbb{E}Z_4(x)^2 \le x^{-\beta} \int_0^\infty \mathrm{e}^{2\gamma(t)t} f(t)^2 \exp\big\{-2\gamma(t)x + \mu \mathrm{e}^{-\mu t}x + \lambda t\big\}\,\mathrm{d}t$$

Let again I_1, I_2, denote the contributions to this integral from the intervals $0 < t \le t_0$, resp. $t > t_1$. The proof that I_1 goes to 0 exponentially fast follows the same lines as above. Further,

$$I_2 \le k_{12} \int_0^\infty \mathrm{e}^{2k_7} t^{2\alpha-2} \exp\big\{-\mu \mathrm{e}^{-\mu t}x - \lambda t\big\}\,\mathrm{d}t \sim k_{13}\frac{\log^{2\alpha-2} x}{x^\beta}\,.$$

This shows the assertion for $i = 1$. For $i = 2$, we have

$$\mathbb{E}Z_4(x)^2 \le k_{14} \int_0^\infty \mathrm{e}^{\gamma(t)t+\xi(t)t} f(t)^2 \exp\bigl\{-\gamma(t)x - \xi(t)x + \mu\mathrm{e}^{-\mu t}x + \lambda t\bigr\}\,\mathrm{d}t$$

For I_1, we insert $\xi(t) \ge 0$ and are then back to the same integral as above. For I_2, we use $\xi(t) \ge k_{15}\gamma(t)$ for $t \ge t_1$ and can then use just the same estimates. □

For the implementation of Algorithms 3, 4, we note the following results:

Proposition 4.2 *The distribution Q_β in* (4.1) *has c.d.f.*

$$Q_\beta(y) = \frac{1}{\Gamma(\beta)} \int_{\mathrm{e}^{-y}}^\infty u^{\beta-1}\mathrm{e}^{-u}\,\mathrm{d}u\,.$$

Proof. In the identity $Q_\beta(y) = \int_{-\infty}^y q(v)\,\mathrm{d}v$, substitue $u = \mathrm{e}^{-v}$. □

Corollary 4.1 *Assume $\beta > 1$. Then a r.v. Y with distribution Q_β can be generated as $Y = -\log Z_\beta$ with Z_β gamma with density $z^{\beta-1}\mathrm{e}^{-z}/\Gamma(\beta)$.*

Proof. $\mathbb{P}(-\log Z_\beta \le y) = \mathbb{P}(Z_\beta \ge \mathrm{e}^{-y}) = \mathbb{P}(Y \le y)$. □

5 Simulation Algorithms for Heavy-Tailed F

In this section, we assume that F is regularly varying, cf. (1.2). As in Section 4, the first step in the design of simulation algorithms is to look for the conditional distribution of T given $X > x$, that is, for an analogue of Theorem 4.1. We then face the difficulty that the results of [5] (more precisely part (2:1) of Theorem 2.1 of [5]) only gives logarithmic asymptotics. Part (i) of the following result improves this to sharp asymptotics:

Theorem 5.1 *Assume that G is exponential with rate μ and that $f(t) = L(t)/t^{\alpha+1}$ with $\alpha > 0$ and $L(x)$ slowly varying as $t \to \infty$. Then:*
(i) $\overline{H}(x) \sim \dfrac{L(\log x)\mu^\alpha}{\alpha\log^\alpha x}$;
(ii) $\mathbb{P}(X > x,\, T > \log x/\mu) \sim \dfrac{L(\log x)\mu^\alpha}{\alpha\log^\alpha x}$;
(iii) $\mathbb{P}(X > x,\, T \le \log x/\mu) \sim \dfrac{L(\log x)\mu^\alpha E_1(\mu)}{\log^{\alpha+1} x}$.

Here $E_1(z) = \int_z^\infty v^{-1}\mathrm{e}^{v}\,\mathrm{d}v$ denotes the exponential integral, cf. [1].

Note that the asymptotics in (i) and (ii) are the same, whereas the one in (iii) exhibits a lighter tail. Thus, the main contribution to $\mathbb{P}(X > x)$ comes from the event $T > \log x/\mu$.

Proof of Theorem 5.1. Obviously, (i) is a trivial consequence of (ii), (iii), so it suffices to prove (ii), (iii).

Consider first (ii). Appealing to Remark 4.1 and substituting $t = \log x/\mu + y\log x/\mu$, we get

$$\begin{aligned}
\frac{1}{L(\log x)}\mathbb{P}(X > x,\, T > \log x/\mu) &\sim \frac{1}{L(\log x)}\int_{\log x/\mu}^{\infty} \exp\{-\mu \mathrm{e}^{-\mu t}x\}\frac{L(t)}{t^{\alpha+1}}\,\mathrm{d}t \\
&= \frac{1}{L(\log x)}\int_0^{\infty} \exp\{-\mu \mathrm{e}^{-y\log x}\}\frac{L\big(\log x(1/\mu + y/\mu)\big)}{(\log x(1/\mu + y/\mu)^{\alpha+1}}\frac{\log x}{\mu}\,\mathrm{d}y \\
&\sim \frac{\mu^{\alpha}}{\log^{\alpha} x}\int_0^{\infty} \frac{R(x,y)}{(1+y)^{\alpha+1}}\,\mathrm{d}y\,. \qquad (5.1)
\end{aligned}$$

where $R(x,y) = L(\log x(1/\mu + y/\mu\big)/L(\log x)$. Choose $0 < \delta < \alpha$. By the Potter bounds ([8] p. 25) there exists k and y_0 such that $R(x,y) \le ky^{\delta}$ for all $y > y_0$, and by the uniform convergence theorem for slowly varying functions ([8] p. 22 with $\rho = 0$ and $a = 1/\mu$), $R(x,y) \to 1$ uniformly on $(0, y_0)$. Since $R(x,y) \to 1$ also on (y_0, ∞), dominated convergence applies to the integral over this interval, and we conclude that (5.1) asymptotically behaves like

$$\frac{\mu^{\alpha}}{\log^{\alpha} x}\int_0^{\infty} \frac{1}{(1+y)^{\alpha+1}}\,\mathrm{d}y = \frac{\mu^{\alpha}}{\alpha\log^{\alpha} x}$$

as claimed.

For (iii), $\mathbb{P}(X > x,\, T \le t_0)$ goes to 0 exponentially fast (at rate at least $\gamma(t_0)$) and can be neglected. Further (cf. again Remark 4.1)

$$\begin{aligned}
\mathbb{P}(X > x,\, t_0 \le T \le \log x/\mu) &\sim \int_{t_0}^{\log x/\mu} \exp\{-\mu \mathrm{e}^{-\mu t}x\}\frac{L(t)}{t^{\alpha+1}}\,\mathrm{d}t \\
&= \int_0^{\log x/\mu - t_0} \exp\{-\mu e^{y}\}\frac{L(\log x/\mu - y)}{(\log x/\mu - y)^{\alpha+1}}\,\mathrm{d}y \\
&\sim \frac{L(\log x)\mu^{\alpha+1}}{\log^{\alpha+1} x}\int_0^{\infty} \exp\{-\mu e^{y}\}\,\mathrm{d}y\,,
\end{aligned}$$

where the last step used similar arguments as in the proof of (ii). But substituting $v = \mathrm{e}^{y}$, the integral becomes $E_1(\mu)/\mu$. □

Theorem 5.2 *Assume that F is regularly varying as in* (1.2). *Then the conditional distribution of $Y = \mu T/\log x - 1$ given $X > x$ has a limit in distribution as $x \to \infty$, namely the Pareto(α) distribution P_{α} with density $p_{\alpha}(y) = \alpha/(1+y)^{\alpha+1}$, $y > 0$.*

It follows that given $X > x$, the order of T is again $\log x/\mu$. However, whereas the deviation of T from $\log x/\mu$ remained of constant order in the gamma case, it now has to be scaled by $\log x$.

Proof of Theorem 5.2. We recall from Theorem 5.1(i) that

$$\mathbb{P}(X > x) \sim \frac{L(\log x)\mu^{\alpha}}{\alpha\log^{\alpha} x}\,.$$

Let $t(x,y) = \log x(1+y)/\mu$. Then

$$\gamma\big(t(x,y)\big) \sim \mu\mathrm{e}^{-\mu t(x,y)} = \mu\mathrm{e}^{-y\log x}/x$$

and therefore (cf. Remark 4.1)

$$\mathbb{P}\big(S(t(x,y)) > x - t(x,y)\big) \sim \exp\big\{-\gamma\big(t(x,y)\big)x\big\} \to 1.$$

With $f(t;x)$ as in the proof of Theorem 4.1, we therefore have it follows that the density of Y given $X > x$ is

$$\begin{aligned}
&\frac{\log x}{\mu\mathbb{P}(X > x)} f\big(t(x,y);x\big)\mathbb{P}\big(S(t(x,y)) > x - t(x,y)\big) \\
&\quad\sim \frac{\alpha \log^{\alpha+1} x}{\mu^{\alpha+1} L(\log x)} \frac{L\big(\log x(1+y)/\mu\big)\mu^{\alpha+1}}{(1+y)^{\alpha+1}\log^{\alpha+1} x} \\
&\quad\sim \frac{\alpha}{L(\log x)} \frac{L(\log x)}{(1+y)^{\alpha+1}} \;=\; p_\alpha(y)\,.
\end{aligned}$$

□

For simulation of $\mathbb{P}(X > x)$, Theorem 5.2 suggest to use the distribution of $T(Y) = (Y \log x + \log x)/\mu$ as importance distribution for T. This choice meets the difficulty that the support of $T(Y)$ is $(\log x/\mu, \infty)$, so that absolute continuity fails and the algorithm can only estimate $\mathbb{P}(X > x, T > \log x/\mu)$:

Algorithm 5 *Generate Y from the Pareto density p_α and let $T = t = (Y \log x + \log x)/\mu$. Calculate the likelihood ratio*

$$W = \frac{\mu f(t)}{\log x\, p_\alpha(\mu t/\log x - 1)} = \frac{f(t)\mu^{\alpha+2}\, t^{\alpha+1}}{\alpha \log^{\alpha+2} x}\,.$$

Compute the crude Monte Carlo estimator $Z_0(x-t)$ for $\mathbb{P}\big(S(t) > x - t\big)$. Return the estimator $Z_5(x) = WZ_0(x-t)$ for $\mathbb{P}(X > x, T > \log x/\mu)$.

That only the crude Monte Carlo estimator of $\mathbb{P}\big(S(t) > x - t\big)$ needs to be used comes of course from the fact that the event $X > x$ is not rare even in the whole support of $T(Y)$, and indeed:

Theorem 5.3 *Algorithm 5 has bounded relative error for estimating $\mathbb{P}(X > x, T > \log x/\mu)$.*

Proof. Appealing to Remark 4.1, we get

$$\begin{aligned}
\mathbb{E}Z_5(x)^2 &\sim \int_{\log x/\mu}^{\infty} \mathbb{P}\big(S(t) > x - t\big)\frac{f(t)^2\mu^\alpha t^{\alpha+1}}{\alpha \log^\alpha x}\,\mathrm{d}t \\
&\le \frac{k_{15}}{\log^\alpha x}\int_{\log x/\mu}^{\infty} \exp\big\{-\mu \mathrm{e}^{-\mu t}x\big\}\frac{L(t)^2}{t^{2\alpha+2}} \cdot t^{\alpha+1}\,\mathrm{d}t \\
&\le \frac{k_{15}}{\log^\alpha x}\int_{\log x/\mu}^{\infty} \frac{L(t)^2}{t^{\alpha+1}}\,\mathrm{d}t \sim \frac{k_{15}L(\log x/\mu)^2}{\log^{2\alpha} x} \\
&\sim \frac{k_{15}L(\log x)^2}{\log^{2\alpha} x} \sim k_{16}\mathbb{P}(X > x, T > \log x/\mu)^2\,,
\end{aligned}$$

where we used Karamata's theorem for the integral asymptotics and (in the last step) Theorem 5.1(i). □

To provide an unbiased estimate of $\mathbb{P}(X > x)$, we thus need an estimator of $\mathbb{P}(X > x, T \le \log x/\mu)$. We first note:

Theorem 5.4 *The conditional distribution of $Y = \log x - \mu T$ given $X > x$ and $T \le \log x/\mu$ has a limit in distribution as $x \to \infty$, namely the distribution R_μ with density $r_\mu(y) = \exp\{-\mu e^y\}/E_1(\mu)$, $y > 0$.*

It follows that given $X > x$ and $T < \log x/\mu$, the order of T is again $\log x/\mu$. However, whereas the deviation of T from $\log x/\mu$ had to be scaled by $\log x$ when T was unrestricted as in Theorem 5.2, it now remains constant.

Proof of Theorem 5.4. Let $t(x,y) = \log x/\mu - y/\mu$. Since $T = \log x/\mu - Y/\mu$, it follows that the density of Y given $X > x$ and $T \le \log x/\mu$ is asymptotically

$$\begin{aligned}
&\frac{1}{\mu \mathbb{P}(X > x, T \le \log x/\mu)} f\bigl(t(x,y)\bigr)\mathbb{P}\bigl(S(t(x,y)) > x - t(x,y)\bigr) \\
&\sim \frac{\log^{\alpha+1} x}{\mu^{\alpha+1} L(\log x) E_1(\mu)} f(\log x/\mu - y/\mu) \exp\bigl\{-\mu e^{-\mu(\log x/\mu - y/\mu)} x\bigr\} \\
&\sim \frac{\exp\{-\mu e^y\}}{E_1(\mu)}
\end{aligned}$$

□

We are now lead to the following algorithm for eatimating $\mathbb{P}(X > x, T \le \log x/\mu)$:

Algorithm 6 *Generate Y from the density r_μ and let $T = t = \log x/\mu - Y/\mu$. Calculate the likelihood ratio*

$$W \;=\; f(t)/\mu r_\mu(\log x - \mu t) \;=\; E_1(\mu) f(t) \exp\bigl\{\mu e^{-\mu t} x\bigr\}/\mu\,.$$

Compute one of the two estimators $Z_i(x - t)$ of Section 3 ($i = 1$ or 2). Return the estimator $Z_6(x) = W Z_i(x - t)$ of $\mathbb{P}(X > x, T \le \log x/\mu)$.

Theorem 5.5 *Algorithm 6 has bounded relative error for estimating $\mathbb{P}(X > x, T \le \log x/\mu)$.*

Let first $i = 1$. Then

$$\begin{aligned}
\mathbb{E} Z_6(x)^2 &= \int_0^{\log x/\mu} \mathbb{E} Z_1(x-t)^2 E_1(\mu) f(t)^2 \exp\bigl\{\mu e^{-\mu t} x\bigr\}/\mu \,\mathrm{d}t \\
&\le\; k_{17} \int_0^{\log x/\mu} e^{-2\gamma(t)x} f(t)^2 \exp\bigl\{\mu e^{-\mu t} x\bigr\}/\mu \,\mathrm{d}t\,.
\end{aligned}$$

A similar argument as in the proof of Theorem 5.3 together with the bound (2.6) for $\gamma(t)$ shows that this asymptotically is bounded by

$$\begin{aligned}
&k_{17} \int_{t_0}^{\log x/\mu} e^{-2\gamma(t)x} \frac{L(t)^2}{t^{2\alpha+2}} \exp\bigl\{\mu e^{-\mu t} x\bigr\}\,\mathrm{d}t \\
&\le k_{17} \int_{t_0}^{\log x/\mu} \exp\bigl\{-\mu e^{-\mu t} x\bigr\} \frac{L(t)^2}{t^{2\alpha+2}}\,\mathrm{d}t \\
&= k_{18} \int_0^{\log x - \mu t_0} \exp\bigl\{-\mu e^y\bigr\} \frac{L(\log x/\mu - y/\mu)^2}{(1 + \log x/\mu - y/\mu)^{2\alpha+2}}\,\mathrm{d}y \\
&\sim k_{19} \frac{L(\log x)^2}{\log^{2\alpha+2} x} \;\sim\; k_{20} \mathbb{P}(X > x)^2\,.
\end{aligned}$$

We omit the details for $i = 2$. □

6 Computational Effort

It was noted alreday by Hammersley & Handscombe [13] that considering variance alone as performance measure of an algorithm may be misleading: one needs also to consider the computational effort. They even quantified this effect in the statement that “The efficiency of a Monte Carlo process may be taken as inversely proportional to the product of the sampling variance and the amount of labour expended in obtaining this estimate.”

The philosophy behind this is the fact that the “inverse efficiency” $\mathbb{V}\mathrm{ar}Z \cdot \mathbb{T}\mathrm{ime}\, Z$ of a simulation estimator Z can be identified with the variance per unit computer time; here $\mathbb{T}\mathrm{ime}\, Z$ is the expected computer time to generate Z. See Glynn & Whitt [12] and [4] III.10. To identify $\mathbb{T}\mathrm{ime}\, Z$ in a mathematical rigorous way may of course be difficult, but in many situations a natural definition suggests itself. For the huge majority of standard rare event simulation algorithms $\mathbb{T}\mathrm{ime}\, Z(x)$, however, grows logaritmically in $\mathbb{V}\mathrm{ar}Z(x)$, and so involving $\mathbb{T}\mathrm{ime}\, Z(x)$ makes little difference. The situation in this paper will now be seen to be quite different.

Crude Monte Carlo simulation of $\mathbb{P}(X > x)$ was implemented in [16] by generating X and returning $Z(x) = 1\{X > x\}$. The effort in generating X is roughly proportional to the number of restarts, which in turn is roughly proportional to X. Thus we take $\mathbb{T}\mathrm{ime}\, Z(x) = \mathbb{E}X$ and get

$$\mathbb{V}\mathrm{ar}Z(x) \cdot \mathbb{T}\mathrm{ime}\, Z(x) = \mathbb{P}(X > x)\big(1 - \mathbb{P}(X > x)\big) \cdot \mathbb{E}X \approx \mathbb{P}(X > x) \cdot \mathbb{E}X\,. \quad (\mathrm{CMC}_0)$$

For the algorithms considered sofar in this paper, we can take $\mathbb{T}\mathrm{ime}\, Z(x) \approx x$ and the bounded relative error property implies

$$\mathbb{V}\mathrm{ar}Z(x) \cdot \mathbb{T}\mathrm{ime}\, Z(x) \approx \mathbb{P}(X > x)^2 \cdot x\,. \quad (\mathrm{IS})$$

To ompare these two expressions, we consider the case of F being exponential (λ) and G exponential(μ). With $\beta = \lambda/\mu$, the probability of no restarts is $\int_0^\infty \lambda e^{-(\mu+\lambda)t}\,\mathrm{d}t = \beta/(1+\beta)$. Thus we have many restarts for β small and many for β large. Further $\mathbb{P}(X > x)$ is of order $x^{-\beta}$ by (4.2). In particular, $\mathbb{E}X = \infty$ when $\beta \le 1$, and then the advantage of (IS) over (CMC$_0$) if of course enormous. However, for $\beta > 1$ (CMC$_0$) is of order $x^{-\beta}$, (IS) of order $x^{1-2\beta}$ which is only notably better if β is large.

The calculation does, however, not pay full justice to crude Monte Carlo simulation because in order to simulate $\mathbb{P}(X > x)$ it is not necessary to generate X but only $X1\{X \le x\}$. Thus, when $\beta < 1$, (CMC$_0$) has to be replaced by

$$\mathbb{V}\mathrm{ar}Z(x) \cdot \mathbb{T}\mathrm{ime}\, Z(x) \approx \mathbb{P}(X > x) \cdot \mathbb{E}[X1\{X \le x\}] \approx x^{1-2\beta}\,. \quad (\mathrm{CMC}_1)$$

where the last step used

$$\mathbb{E}[X1\{X \le x\}] = \int_0^x \mathbb{P}(X > s)\,\mathrm{d}s \approx \int_{x_0}^x s^{-\beta}\,\mathrm{d}s \ \approx\ x^{1-\beta}\,.$$

Thus, the order is of the same magnitude as (IS). In order words, the importance sampling algorithm does not lead to any asymptotic improvement in the work-corrected variance!

This raises the problem of finding a complexity O(1) but still efficient estimator of $\mathbb{P}\bigl(S(t) > x - t\bigr)$. One may note that this probability is simply the probability that the largest interevent time of a Poisson(μ) process M on the interval $[0, x-t]$ is at most t (counting 0 and $x - t$ as epochs). An explicit expression for this is known (Fisher [11]; see also Davis [10]) given the number $m = M(x-t)$ of Poisson epochs, but is an alternating series with order m terms, so using this formula would not reduce the complexity from O(x) and could potentially be numerically unstable. We have therefore not pursued this approach, but suggest a different solution in the next section.

7 An Algorithm Exploiting Lundberg's Inequality

The problem in the analysis of Section 6 is the order of increase in $\mathbb{T}\mathrm{ime}\, Z(x)$ in x. We now suggest an alternative estimator having the property $\mathbb{T}\mathrm{ime}\, Z(x) = \mathrm{O}(1)$. The estimator may lead to increased confidence bands, in particular for small x, but the problem vanishes as $x \to \infty$.

The idea is to avoid the O(x) simulation of $\mathbb{P}\bigl(S(t) > x - t\bigr)$ by just replacing this probability by its upper and lower Lundberg bounds. For example, in the gamma-exponential setting of Section 4:

Algorithm 7 *Generate Y from the density q in* (4.1) *and let $T = t = (\log x + \log \mu + Y)/\mu$. If $T \le 0$, return the estimator $Z_3(x) = 0$. Otherwise calculate $\gamma(t)$ and the likelihood ratio*

$$W = f(T)/\mu q(\mu T - \log x - \log \mu) = f(T)x^{-\beta}\mu^{-1-\beta}\Gamma(\beta)\exp\bigl\{\mu \mathrm{e}^{-\mu T}x + \lambda T\bigr\}\,,$$

and let $Z_8'(x) = W\mathrm{e}^{-\gamma(t)x}$, $Z_8''(x) = W\mathrm{e}^{-\gamma(t)(x-t)}$. Repeat R times and compute the empricial means $z_8'(x), z_8''(x)$ and variances $s_7'(x)^2, s_7''(x)^2$. Return the interval

$$\bigl(z_8'(x) - 1.96 s_8'(x)/R^{1/2}, z_8''(x) + 1.96 s_8''(x)/R^{1/2}\bigr)\,. \tag{7.1}$$

It follows immediately that:

Theorem 7.1 *The interval* (7.1) *is an asymptotic 95% confidence interval for $\mathbb{P}(X > x)$. That is, as $R \to \infty$ it contains $\mathbb{P}(X > x)$ with probability at least 95%.*

The limiting probability that (7.1) contains $\mathbb{P}(X > x)$ is of course somewhat larger than 95%. How much depends on how tight the Lundberg bounds are, but as noted above, these bounds are asymptotically tight as $t \to \infty$.

8 Numerical Examples

We took F as exponential(1) and G as exponential(0.8). Thus we are in the setting of Section 4 with $\alpha = 1$, $\beta = 1.25$. We consider 10 x-values $10^{i/2}$, $i = 1, \ldots, 10$; in this range, $z(x) = \mathbb{P}(X > x)$ varies approximately from 10^{-1} to 10^{-6}.

We implemented first Algorithm 4 with $R = 1000$ replications.

Figure 1 shows the 95% twosided confidence band (the scale is log_{10}–log_{10} as for all figures except Figure 4). As is seen, the precision is excellent even with the modest value $R = 1000$, except for small values of x. The error appears to be decreasing in x and this is further confirmed by Figure 2 that gives the relative error of the algorithm, as defined by the halfwidth of the confidence band divided by the simulated values. It may even look as if the relative error goes to zero, even if our theoretical analysis rather suggest it has a limit. This could be explained by Algorithm 2 becoming more and more efficient as $t \to \infty$ because $\xi(t) \uparrow \gamma(t)$ as $t \to \infty$, cf. Proposition 3.2, and that the limit $\gamma(t)$ is not yet attained in the range of x-values under consideration.

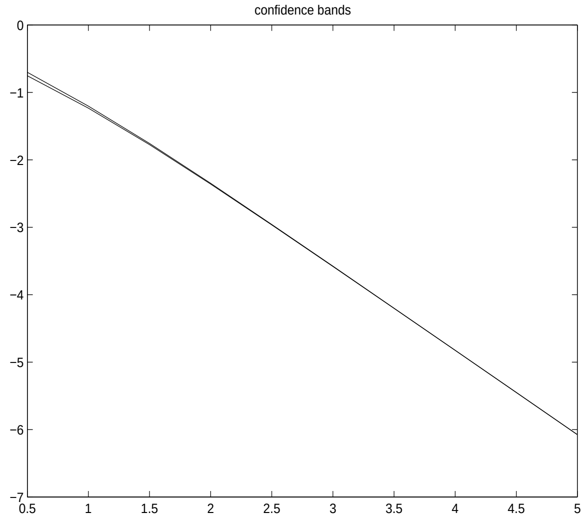

Figure 1: Confidence bands for Algorithm 4

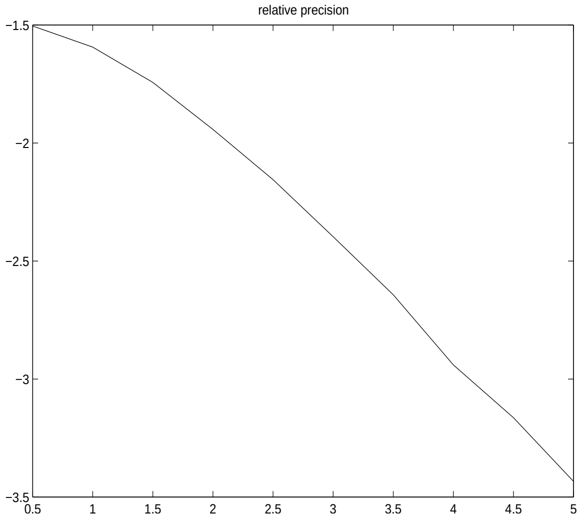

Figure 2: Relative precision of Algorithm 4

In comparison to Figure 1, the confidence bands produced by Algorithm 7 are given in Figure 3.

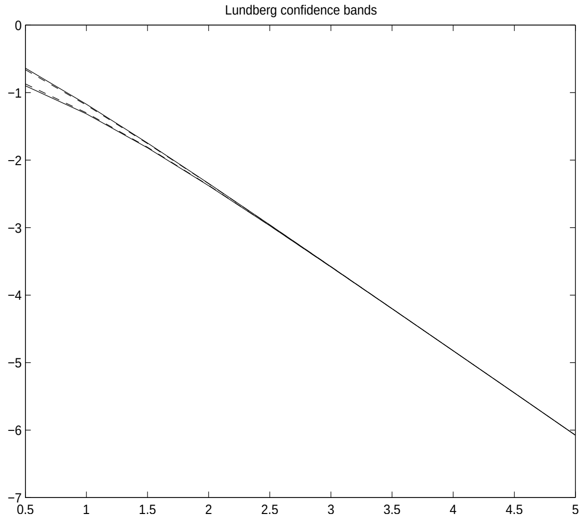

Figure 3: Confidence bands for Algorithm 7

The precision is comparable to Algorithm 4 except for the smallest values of x. Of course, one expects this to be due to the inaccuracy of the Lundberg bounds for small x, and this is confirmed by Figure 4 that shows the upper and lower Lundberg bounds divided by the simulated values.

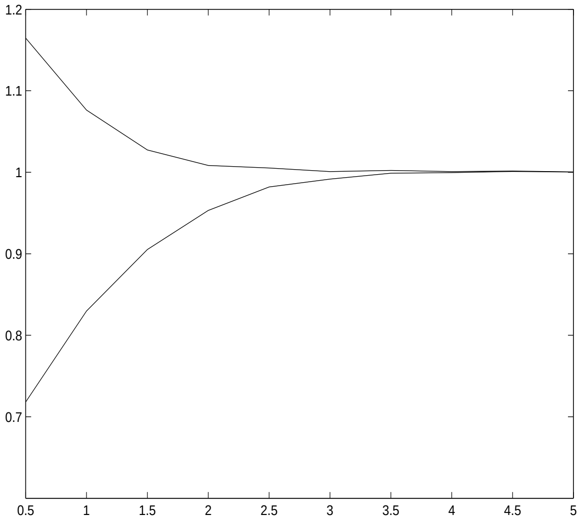

Figure 4: Lundberg bounds

The following table gives a comparison of the running time for Algorithms 4 and 7, more precisely the ration between the one for Algorithm 7 and the one for Algorithm 4 as produced by Matlab's `tic` and `toc` commands.

x	$10^{1/2}$	10^1	$10^{3/2}$	10^2	$10^{5/2}$	10^3	$10^{7/2}$	10^4	$10^{9/2}$	10^5
	407	377	228	157	50	18	4.8	1.2	0.35	0.12

It is seen that indeed the rootfinding in Algorithm 7 (implemented via Matlab's `fsolve`) is much more expensive than the $\mathrm{O}(x)$ complexity of Algorithm 4 for small or moderate x. The overall picture when comparing this with the precision as discussed above is that Algorithm 4 is preferable for small or moderate x, but Algorithm 7 for large x.

We finally took the opportunity to use our Matlab program to check the accuracy of the approximations of [5], in this specific setting (4.2). Figure 5 shows the simulated values versus approximations, and Figure 6 the relative error of the approximations, as defined by the absolute value of the difference between the simulated value and the approximation divided by the simulated value. The relative error indeed appears to go to 0, as should be, and the roughly linear shape of Figure 6 (cf. the log–log scale) suggests a roughly power-like rate of decrease.

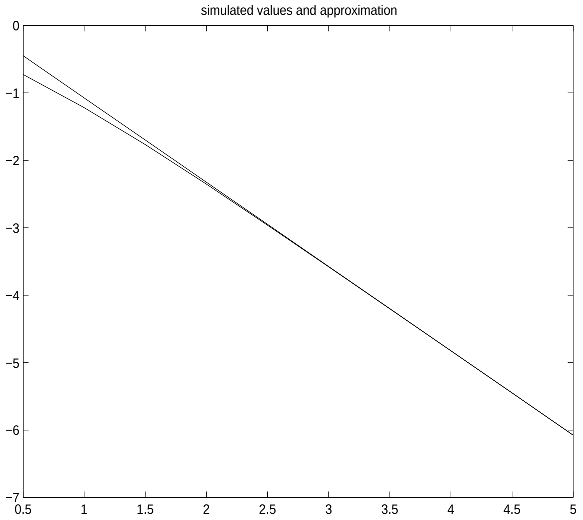

Figure 5: Simulated values versus approximations

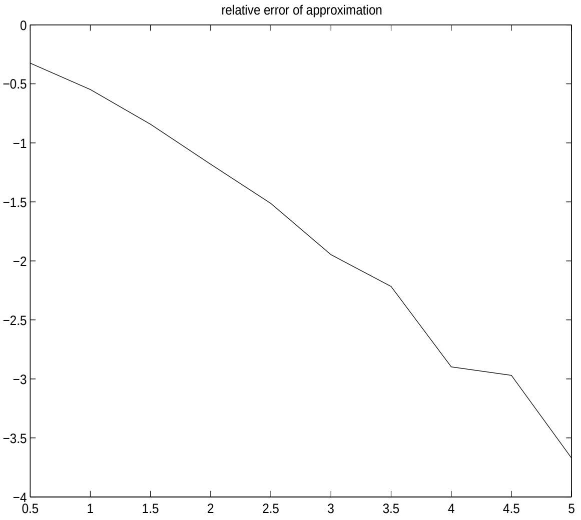

Figure 6: Relative error of approximation

Appendix A: Root Properties

It is shown in [5] that for a general G, $\gamma(t) \sim \mu\overline{G}(t)$ as $t \to \infty$. If G is exponential(μ), as assumed in the following, we shall need certain refinements and related results. First note that the defining equation (2.2) for $\gamma(t)$ means

$$1 = \varphi\bigl(\gamma(t)\bigr)\,, \quad \text{where } \varphi(\gamma) = \frac{\mu}{\gamma-\mu}\bigl(\mathrm{e}^{(\gamma-\mu)t}-1\bigr)\,. \tag{A.1}$$

Proof of (2.6). The r.h.s. of (2.2) (or, equivalently, of $\varphi(\gamma)$) is an increasing function of γ. Taking $\gamma = \mu\mathrm{e}^{-\mu t}$, this r.h.s. becomes

$$\frac{\mu}{\mu-\mu\mathrm{e}^{-\mu t}}\bigl(1-\exp\bigl\{(\mu\mathrm{e}^{-\mu t}t-\mu t)\bigr\} < \frac{\mu}{\mu-\mu\mathrm{e}^{-\mu t}}(1-\mathrm{e}^{-\mu t}) = 1\,.$$

Therefore the desired solution $\gamma(t)$ must be $> \mu\mathrm{e}^{-\mu t}$.

Since

$$\begin{aligned}\sum_{k=2}^{\infty}\frac{\gamma(t)^k}{k!}\int_0^t y^k\mu\mathrm{e}^{-\mu y}\,\mathrm{d}y &\le \gamma(t)^2\sum_{k=0}^{\infty}\frac{\gamma(t)^k}{k!}\int_0^\infty y^k\mu\mathrm{e}^{-\mu y}\,\mathrm{d}y\\ &= \gamma(t)^2\int_0^\infty \mu\mathrm{e}^{\gamma(t)y-\mu y}\,\mathrm{d}y \sim \gamma(t)^2\int_0^\infty \mu\mathrm{e}^{-\mu y}\,\mathrm{d}y = \mathrm{O}\bigl(\gamma(t)^2\bigr)\end{aligned}$$

as $t \to \infty$, we further get

$$\begin{aligned}1 &= \int_0^t\bigl(1+\gamma(t)y\bigr)\mu\mathrm{e}^{-\mu y}\,\mathrm{d}y + \mathrm{O}\bigl(\gamma(t)^2\bigr)\\ &= 1-\mathrm{e}^{-\mu t}+\gamma(t)\frac{1}{\mu}-\gamma(t)\int_t^\infty \mu y\mathrm{e}^{-\mu y}\,\mathrm{d}y + \mathrm{O}\bigl(\gamma(t)^2\bigr)\\ &= 1-\mathrm{e}^{-\mu t}+\gamma(t)\frac{1}{\mu}-\gamma(t)t\mathrm{e}^{-\mu t}\gamma(t)+\frac{\gamma(t)}{\mu}\mathrm{e}^{-\mu t}+\mathrm{O}\bigl(\gamma(t)^2\bigr)\\ &= 1-\mathrm{e}^{-\mu t}+\gamma(t)\frac{1}{\mu}-\gamma(t)t\mathrm{e}^{-\mu t}\bigl(1+\mathrm{o}(1)\bigr)+\mathrm{O}\bigl(\gamma(t)^2\bigr)\,.\end{aligned}$$

This implies the r.h. inequality in (2.6). □

Equivalent form of (A.1) are

$$\gamma(t) = \mu\mathrm{e}^{\gamma(t)t-\mu t}\,, \tag{A.2}$$

$$\gamma(t) = \mu + \frac{\log\gamma(t)-\log\mu}{t} \tag{A.3}$$

[indeed, (A.2) follows from (A.1) by trivial algebra, and (A.3) from (A.2) by taking logarithms]. Obviously, there is no explicit solution. Since some of our algorithms require computation of $\gamma(t)$ for a large number of t, an efficient numerical scheme is needed. In our numerical examples, we used Matlab's routine `fsolve`. Another possibility is involving the Lambert W function (the root of $\theta\mathrm{e}^{-\theta} = y$, in terms of which the solution of (A.2) can be expressed). In software environments, where general rootfinding algorithms are unavailable, one may use traditional Newton-Raphson iteration $\gamma_{n+1} = \gamma_n - \varphi(\gamma_n)/\varphi'(\gamma_n)$ or iterative schemes based upon (A.2), (A.3):

Proposition 8.1 *We have $\gamma(t) > \mu$, $\gamma(t) = \mu$ or $\gamma(t) < \mu$ according as $\mu t < 1$, $\mu t = 1$ or $\mu t > 1$. Further $\gamma = \gamma(t)$ can be computed as $\gamma = \lim_{n\to\infty} \gamma_n$, where in the case $\mu t > 1$*

$$\gamma_{n+1} \;=\; \mu \mathrm{e}^{\gamma_n t - \mu t}$$

and the initial value γ_0 is chosen with $\gamma_0 < \mu$, and in the case $\mu t < 1$

$$\gamma_{n+1} = \mu + \frac{\log \gamma_n - \log \mu}{t}$$

and the initial value γ_0 is chosen with $\gamma_0 > \mu$.

The need to distinguish between the cases $\mu t < 1$ and $\mu t > 1$ arises because (A.2), (A.3) have the additional fixpoint μ, and $\gamma(t)$ is attractive for (A.2) and μ repulsive when $\mu t > 1$, but repulsive when $\mu t < 1$ (similar remarks apply to (A.3)), see Fig. 7.

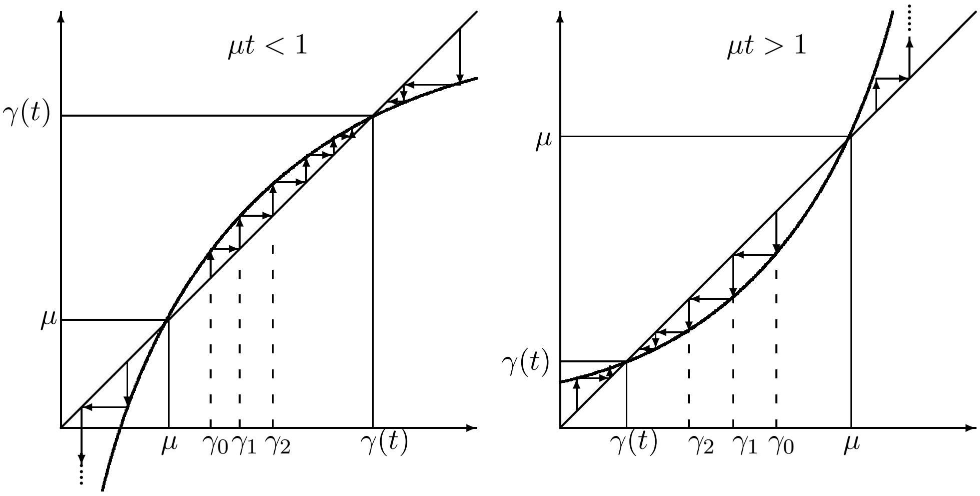

Figure 7: Iterations for $\gamma(t)$

Proof of Proposition 8.1. The first statement follows immediately since the r.h.s. of (2.2) equals μt when $\gamma = \mu$ and is increasing in γ.

The convergence properties follow by standard arguments based upon convexity, resp. concavity, see again Fig. 7. □

Proof of (2.7). $\gamma(t) = -\mu \log t/t\big(1+\mathrm{o}(1)\big)$ *as* $t \downarrow 0$. Define $\gamma_\delta = \mu - \delta \log t/t - \log \mu$. Then the r.h.s. of (A.2) is of order $t^{-\delta}$ for $\gamma = \gamma_\delta$, whereas the l.h.s. is of order $|\log t|/t$. If $\delta > 1$, $t^{-\delta}$ increases faster than $|\log t/t|$, and since the r.h.s. of (A.2) is convex and the l.h.s. affine, the desired solution $\gamma(t)$ must be $< \gamma_\delta$. A similar argument shows that $\gamma(t) > \gamma_\delta$ when $\delta < 1$. □

We finally give the

Proof of Proposition 3.2. That $\gamma(t) \sim \mu\overline{G}(t)$ is shown in [5]. For $\xi(t) \sim \gamma(t)$, note that the definition of $\xi(t)$ means

$$\begin{aligned} 1 &= G(t)\int_0^t \mathrm{e}^{(\gamma(t)+\xi(t))u} g(u)\mathrm{d}u \\ &= \big(1-\overline{G}(t)\big)\int_0^t \mathrm{e}^{\gamma(t)u}\big[(1+\xi(t)u+\xi(t)\mathrm{O}\big(t^2\xi(t)\big)\big]g(u)\mathrm{d}u \\ &= \big(1-\overline{G}(t)\big)\big[1+\xi(t)/\mu+\mathrm{o}\big(\xi(t)\big)\big], \end{aligned} \tag{A.4}$$

where we used $\xi(t) < \gamma(t) \sim \mu\overline{G}(t)$ together with $\widehat{G}[\epsilon] < \infty$ to infer that $t^2\xi(t) \to 0$, and $\widehat{G}[\epsilon] < \infty$ and dominated convergence to infer that

$$\int_0^t u\mathrm{e}^{\gamma(t)u} g(u)\,\mathrm{d}u = 1/\mu + \mathrm{o}(1)\,.$$

However, (A.4) is only, possible if $\xi(t) \sim \mu\overline{G}(t)$ □

Appendix B: Simulation of Geometric Sums

Let $U_1^*, U_2^*, \ldots$ be i.i.d. with common distribution G^* concentrated on $(0,\infty)$ and let N be an independent geometric r.v., $\mathbb{P}(N = n) = (1-\rho)\rho^n$ for $n = 0, 1, \ldots$ Define further

$$S_n^* = U_1^* + \cdots + U_n^*\,, \quad z(x) = \mathbb{P}(S_N^* > x)\,, \quad \tau^*(x) = \inf\{n : U_1^* + \cdots + U_n^* > x\}\,.$$

In [4], Exercise 2.3 p. 172 (see also Blanchet & Li [9]), the following algorithm is suggested for simulation of $z(x)$ and it is claimed that it has bounded relative error as $x \to \infty$.[2] As set-up, compute γ^* as solution of

$$1 = \rho \int_0^\infty \mathrm{e}^{\gamma^* y}\, G^*(\mathrm{d}y)\,. \tag{B.1}$$

Let G^* be the distribution defined by $\mathrm{d}G_{\gamma^*}/\mathrm{d}G^*(y) = \rho\mathrm{e}^{\gamma^* y}$, and to generate one replication of the estimator, proceed as follows:

Algorithm 8 *Generate $U_1^*, U_2^*, \ldots$ from G_{γ^*}. Stop the simulation at $\tau^*(x)$ and return the estimator $Z^*(x) = \mathrm{e}^{-\gamma^* S_{\tau^*(x)}}$.*

To understand the algorithm, note first that $z(x) = \mathbb{P}\big(\tau^*(x) \le N\big)$. Next let $\mathbb{P}_{\gamma^*}$ be the probability measure where the U_i^* are i.i.d. with distribution G_{γ^*} and N remains independent and geometric. Then by the definition of G_{γ^*},

$$\mathbb{P}(U_1^* \in \mathrm{d}u) = \frac{1}{\rho}\mathbb{E}_{\gamma^*}\big[\mathrm{e}^{-\gamma^* U_1^*};\, U_1^* \in \mathrm{d}u\big]\,.$$

By a standard extension to stopping times (see, e.g., [4] pp. 131–132), this implies

$$z(x) = \mathbb{E}_{\gamma^*}\big[\frac{1}{\rho^{\tau^*(x)}}\mathrm{e}^{-\gamma^* S_{\tau^*(x)}};\, \tau^*(x) \le N\big] = \mathbb{E}_{\gamma^*}\mathrm{e}^{-\gamma^* S^*_{\tau(x)}}\,,$$

where we used that N remains geometric and independent of the U_i^* under $\mathbb{P}_{\gamma^*}$. I.e., the estimator $Z^*(x)$ is unbiased.

Further

$$\mathbb{E}_{\gamma^*} Z^*(x)^2 = \mathbb{E}_{\gamma^*}\mathrm{e}^{-2\gamma^* S_{\tau(x)}} \le \mathrm{e}^{-2\gamma^* x} = \mathrm{O}\big(z(x)^2\big)\,,$$

where the last step used the standard Cramér-Lundberg asymptotics $z(x) \sim C^* e^{-2\gamma^* x}$ valid with $0 < C^* < \infty$ under weak additional assumptions. This shows that $Z^*(x)$ has bounded relative error.

[2] Note that the expression for the estimator in *loc. cit.* contains typos, corrected here.

Remark 8.1 For the geometric sum occuring in RESTART with $T \equiv t$ as discussed in Section 2, we have $\rho = G(t)$ and G^* is the distribution with density $g(y)/G(t)$, $0 < y < t$. Therefore γ^* is the root $\gamma(t)$ defined in (2.2), and G_{γ^*} is the distribution with density $\mathrm{e}^{\gamma(t)y} g(y)$, $0 < y < t$. □

Remark 8.2 Algorithm 8 may appear rather different from the best algorithm known for Poisson (rather than geometric) sums discussed in [4] VI.2d, where one exponentially tiltes the whole distribution of S_N^*, leading to a new compound sum with changed Poisson parameter and exponentially tilted increment distribution. The tilting parameter $\theta = \theta(x)$ is determined by $\mathbb{E}S_N^* \mathrm{e}^{\theta S_N^*}/\mathbb{E}\mathrm{e}^{\theta S_N^*} = x$. Performing the same operation for a geometric sum S_N^*, one can easily check that the relevant θ has limit γ^* so that the two algorithms asymptotically coincide. □

References

[1] M. Abramowitz & I.A. Stegun, eds. (1972) *Handbook of Mathematical Functions.* Dover.

[2] L.N. Andersen & S. Asmussen (2008) Parallel computing, failure recovery and extreme values. *J. Statist. Meth. Appl.* **2**, 279–292;

[3] S. Asmussen (2003) *Applied Probability and Queues* (2nd ed.). Springer-Verlag.

[4] S. Asmussen & P.W. Glynn (2007) *Stochastic Simulation: Algorithms and Analysis.* Springer-Verlag.

[5] S. Asmussen, P. Fiorini, L. Lipsky, T. Rolski & R. Sheahan (2008) On the distribution of total task times for tasks that must restart from the beginning if failure occurs. First version available from `www.thiele.au.dk`, revision to appear in *Math. Oper. Res.* Nov. 2008.

[6] S. Asmussen & L. Lipsky (2008) Failure recovery in computing and data transmission: limit theorems for checkpointing. *Working paper.*

[7] P. Billingsley (1968) *Convergence of Probability Measures.* Wiley.

[8] N.H. Bingham, C.M. Goldie, & J.L. Teugels (1987) *Regular Variation.* Cambridge University Press.

[9] J.H. Blanchet & C. Li (2006) Efficient rare-event simulation for geometric sums. *Proc. RESIM*, Bamberg, Germany.

[10] H.A. David (1970) *Order Statistics.* Wiley.

[11] R.A. Fisher (1929) *Proc. Roy. Soc.* **A125**, 54–59.

[12] P.W. Glynn & W. Whitt (1992) The asymptotic efficiency of simulation estimators. *Oper. Res.* **40**, 505–520.

[13] J.M. Hammersley & D.C. Handscomb (1964) *Monte Carlo Methods.* Methuen.

[14] P. Jelenković & J. Tan (2007) Can retransmissions of superexponential documents cause subexponential delays? *Proceedings of IEEE INFOCMO'07*, pp. 892–900, Anchorage.

[15] P. Jelenković & J. Tan (2007) Characterizing heavy-tailed distributions induced by retransmissions. *Workshop on Transient and Analytic Analysis of Queues*, EURANDOM, Eindhoven, October 17–19 2007.

[16] R. Sheahan, L. Lipsky, P. Fiorini & S. Asmussen (2006) On the distribution of task completion times for tasks that must restart from the beginning if failure occurs. *SIGMETRICS Performance Evaluation Review* **34**, 24–26.

[17] G. Willmot & X. Liu (2001) *Lundberg Approximations for Compound Distributions with Insurance Applications.* Lecture Notes in Statistics **156**. Springer-Verlag.