Heavy Tails, Importance Sampling and Cross–Entropy

Søren Asmussen[*§] Dirk P. Kroese[†]
Reuven Y. Rubinstein[‡§]

August 18, 2003

Abstract

We consider the problem of estimating $\mathbb{P}(Y_1 + \cdots + Y_n > x)$ by importance sampling when the Y_i are i.i.d. and heavy-tailed. The idea is to exploit the cross-entropy method as a tool for choosing good parameters in the importance sampling distribution; in doing so, we use the asymptotic description that given $\mathbb{P}(Y_1 + \cdots + Y_n > x)$, $n-1$ of the Y_i have distribution F and one the conditional distribution of Y given $Y > x$. We show in some specific parametric examples (Pareto and Weibull) how this leads to precise answers which, as demonstrated numerically, are close to being variance minimal within the parametric class under consideration. Related problems for M/G/1 and GI/G/1 queues are also discussed.

Key Words ALGORITHMIC COMPLEXITY, CROSS-ENTROPY, GI/G/1 QUEUE, IMPORTANCE SAMPLING, MAXIMUM LIKELIHOOD, M/G/1 QUEUE, PARETO DISTRIBUTION, POLLACZEK-KHINTCHINE FORMULA, RANDOM WALK, RARE EVENT, SUBEXPONENTIAL DISTRIBUTION, WEIBULL DISTRIBUTION

[*]Department of Theoretical Statistics, Department of Mathematical Sciences, Aarhus University, Ny Munkegade, 8000 Aarhus C, Denmark; `asmus@imf.au.dk`; `home.imf.au.dk/asmus`

[†]Department of Mathematics, The University of Queensland, Brisbane 4072, Australia; `kroese@maths.uq.edu.au`; `www.maths.uq.edu.au/~kroese`

[‡]Faculty of Industrial Engineering and Management, Technion, Haifa, Israel; Research Supported by Israel Science Foundation under contract 191-565; `ierrr01@ie.technion.ac.il`; `iew3.technion.ac.il:8080/ierrr01.phtml`

[§]Partially supported by MaPhySto — A Network in Mathematical Physics and Stochastics, founded by the Danish National Research Foundation

1 Introduction

This paper is concerned with importance sampling (IS) and cross-entropy (CE) techniques for simulating small probabilities, in the presence of heavy-tailed distributions.

Despite the fact that performance evaluation with heavy tails has received considerable attention in recent years, the literature on simulation methods consists of just a handful papers, in contrast to the light–tailed case where the number of references is huge. Also, the models and problems for which satisfying solutions have been developed are quite simple, basically evaluating $\mathbb{P}(Y_1 + \cdots + Y_n > x)$ where $Y_1, \ldots, Y_n$ are i.i.d. with common distribution F concentrated on $(0, \infty)$ and heavy-tailed, and n a fixed integer or an independent random variable, and (closely related) evaluating the tail of the M/G/1 waiting time distribution; according to the Pollaczek-Khintchine (PK) formula, this corresponds to taking n above as a geometric r.v.

In the light–tailed case, the intuition behind most efficient algorithms is that one should perform an i.i.d. change of measure (twist of distribution; say of $Y_1, \ldots, Y_n$ in the above setting) motivated from an asymptotic description of the way in which the rare event in question occurs. Heavy tail asymptotics, however, usually involves just one or a few big random variables, with the rest being unaffected by the rare event, cf. e.g. (2) below, and therefore one would not apriori expect a good change of measure to be i.i.d. (in fact, the first efficient algorithm for heavy tails, given in [5], does not even use importance sampling but a different variance reduction method, namely, conditional Monte Carlo). Nevertheless, it is found in [6] that the most obvious non–i.i.d. IS schemes do not asymptotically improve the variance, and a further finding of [6] is that an i.i.d. change of measure may indeed be efficient. The IS distribution is taken independent of x in [6] but substantial performance improvements are obtained in [14] by choosing it dependent on x.

Both in [6] and [14], the change of measure which is asymptotically efficient (in a sense to be made precise in Section 2) is subject to choice within a rather broad class, in contrast to the light–tailed case where it is essentially unique, cf. [7] Theorem 17.7. Relevant questions are therefore how sensitive the performance is to the particular choice, and whether there are general principles allowing to identify the optimal choice. In Sections 3, 4, we present numerical examples illustrating the first, and suggest a more theoretical approach for the second; this has its starting point in the CE method [17] but also links up with the maximum likelihood method from statistics [18]. The setting of Section 3 is that of [14], hazard rate twisting, in the two specific examples of Pareto and Weibull distributions. In Section 4, we study the problem of scale twisting in the Pareto case which has not been considered so far in the literature. Our results essentially indicate that this change of measure has little promise of leading to algorithms which are more efficient than existing ones. However, the numerical results support what is maybe the main message of the paper, that choosing the IS distribution via minimal CE is a quick and systematic way to find a change of measure which is close to being variance minimal

The setting of $\mathbb{P}(Y_1 + \cdots + Y_n > x)$ is, as noted above, sufficient to deal with the M/G/1 queue. Nevertheless, a main challenge left by [6, 14] is to extend to more general models, in particular the GI/G/1 queue (an algorithm is proposed in [10] but unfortunately it applies essentially only to the Weibull distribution, not to the more standard class of regularly varying distributions, and further one may object that a truncation step is involved without explicit bounds allowing to control the error). In [18] and [15] it is discussed how parametric IS via the CE method can readily give an excellent speed up (variance reduction) for the GI/G/1 queue and more complex queueing models, for both light and heavy tail distributions. It was not clear from the numerical results, however, whether in the heavy tail case one gets polynomial complexity for the GI/G/1 queue. In Section 6 we complement the counterexamples of [6] by showing in fact the complexity is exponential. This does of course not contradict the main finding of the rest of the paper, that in a given setting the CE method does very well in finding the best change of measure.

The content of the rest of the paper is as follows. Section 2 is a short preliminary on rare events simulation, heavy tails and the cross-entropy method. Some cruder but sometimes more easily implemented alternative to the CE method in Section 3 are briefly discussed in Section 5.

2 Preliminaries

We refer to [7] and [12] for general surveys on rare events simulation and to [11, 3, 1, 20, 18, 15] for heavy tails. The set–up and facts that will be needed in the paper can be found in these references as well as an abundance of research articles, and we will therefore only give a brief summary.

2.1 Rare events simulation

We consider a family $\{A(x)\}$ of events defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and indexed by a parameter $x \in \mathbb{R}$, such that $z(x) = \mathbb{P}(A(x)) \to 0$ as $x \to \infty$. A Monte Carlo method estimate $\widehat{z}(x)$ of $z(x)$ is obtained by simulating N replicates $Z_1,\dots,Z_N$ of a random variable $Z(x)$ with $\mathbb{E}Z(x) = z(x)$ and letting $\widehat{z}(x)$ be the empirical mean. The traditional measure for the efficiency of the scheme is the relative error $\epsilon(x) = \bigl(\mathbb{V}\mathrm{ar}Z(x)\bigr)^{1/2}/z(x)$, and the family $\{Z(x)\}$ is called *logarithmically efficient*, or for brevity just *efficient*, if $\epsilon(x) = \bigl(\mathrm{o}(z(x)^{-\delta}\bigr)$ for any $\delta > 0$; often also the term *polynomial* or *polynomial time* is used.

The crude Monte Carlo method (CMCM) corresponds to $Z(x) = I(A(x))$ and sampling from the given probability measure $\mathbb{P}$. It has relative error of order $z(x)^{-1/2}$ and the CMCM is therefore not efficient. Importance sampling corresponds to $Z(x) = WI(A(x))$, where now the sampling is done from a different probability measure $\widetilde{\mathbb{P}}$ (possibly dependent on x) and W is the likelihood ratio $\mathrm{d}\mathbb{P}/\mathrm{d}\widetilde{\mathbb{P}}$. Efficiency or even variance reduction is not guaranteed, but there are many examples in the literature where one can indeed obtain efficiency by an appropriate choice of $\widetilde{\mathbb{P}}$. The dominant method for producing such a $\widetilde{\mathbb{P}}$ is to take $\widetilde{\mathbb{P}}$ as close as possible to $\mathbb{P}^{(x)} = \mathbb{P}\bigl(\cdot\,|\,A(x)\bigr)$ (the conditional distribution given the rare event). In particular, this approach has proved fruitful for light tails where it most often leads to an exponential change of measure scheme.

2.2 Heavy tails

We consider here a heavy–tailed setting where some underlying distribution F is *subexponential*, meaning that the convolution tail $\overline{F}^{*n}(x)$ satisfies

$$\overline{F}^{*n}(x) \;=\; \mathbb{P}(Y_1+\cdots+Y_n > x) \;\sim\; n\overline{F}(x) \tag{1}$$

(here $Y_1, Y_2, \dots$ are i.i.d. with common distribution F, and $a(x) \sim b(x)$ means $a(x)/b(x) \to 1$ as $x \to \infty$). For the intutition behind much of this paper, it is crucial to note $A(x) = \{Y_1+\cdots+Y_n > x\}$ occurs by $n-1$ of the Y_i have distribution F and one the conditional distribution of Y given $Y > x$, and all components being independent. In terms of the order statistics $Y_{(1)} < \cdots < Y_{(n)}$,

$$\Bigl\|\mathbb{P}\bigl(Y_{(1)},\dots,Y_{(n)} \in \cdot\bigr)\,\big|\,A(x)\Bigr\| \;-\; \underbrace{F\otimes\cdots\otimes F}_{n-1}\otimes\,\mathbb{P}(Y\in\cdot\,|\,Y>x) \;\to\; 0, \tag{2}$$

see [3] Lemma 5.6 p. 278 ($\|\cdot\|$ = total variation distance).

Our main examples will be Pareto and Weibull distributions, where

$$\overline{F}(x) \;=\; \frac{1}{(1+x/\gamma)^{\alpha}}, \tag{3}$$

$$\overline{F}(x) \;=\; \mathrm{e}^{-(x/\gamma)^{\beta}}, \tag{4}$$

respectively; note that γ is just a scale parameter whereas α and β determine the degree of heavy–tailedness (one needs $\beta < 1$ for the Weibull distribution to be heavy–tailed).

In terms of $\Lambda(x) = -\log \overline{F}(x)$ and the hazard rate $\lambda(x) = \Lambda'(x)$, one may note that twisting the hazard rate to $\theta\lambda(x)$ as in [14] simply means changing α in (3) and γ in (4).

2.3 The cross–entropy method

The cross-entropy method originated from an adaptive method for estimating probabilities of rare events in complex stochastic networks [16], and has quickly evolved into a versatile and unified method for efficient simulation and combinatorial and multi-extremal continuous optimization, [17, 8, 9, 15, 13]. For our purposes we may view the CE method as a particular implementation of choosing a good change of measure by making the importance sampling distribution $\widetilde{\mathbb{P}}$ look as much alike $\mathbb{P}^{(x)}$ as possible. The idea is to take the Kullback–Leibler distance

$$\mathcal{D}\big(\mathbb{P}^{(x)}, \widetilde{\mathbb{P}}\big) \;=\; \mathbb{E}^{(x)} \log \frac{\mathrm{d}\mathbb{P}^{(x)}}{\mathrm{d}\widetilde{\mathbb{P}}} \tag{5}$$

as a measure of closeness and minimize with respect to $\widetilde{\mathbb{P}}$. The practical implementation in more complex models involves typically a (numerical) minimization problem

$$\min_\theta \mathcal{D}(\mathbb{P}^{(x)}, \mathbb{P}_\theta), \tag{6}$$

where we look for $\widetilde{\mathbb{P}} = \mathbb{P}_\theta$ not in the set of all absolutely continuous probability distributions but rather in a restricted parametric class $\{\mathbb{P}_\theta, \theta \in \Theta\}$. For example, for the estimation of $\mathbb{P}(Y_1 + \cdots + Y_n > x)$ with a Pareto distribution as in (3), it is natural to restrict to an i.i.d. change of measure where the new distribution of $Y_1, \ldots, Y_n$ is again Pareto, only with α, γ changed to $\widetilde{\alpha}, \widetilde{\gamma}$ (or possibly only one of the parameters changed). If, in general, $Y_1, \ldots, Y_n$ are i.i.d. random variables with common density $f_\theta(y)$ with respect to the Lebesgue measure, then minimization of (6) reduces to the maximization problem

$$\max_\theta \mathbb{E}^{(x)} \sum_{i=1}^{n} \log f_\theta(Y_i)\ . \tag{7}$$

With rare events, naive numerical optimization of (7) runs into difficulties because the tilted parameters will typically be far off the given ones, and the crux of the cross-entropy method is that it provides an adaptive optimization algorithm; we will not go into details since the examples of this paper are simple enough that we can deal directly with the minimization.

It is crucial for the following to note that entropy minimization, as in (6), is closely related to likelihood maximization in statistics, see [13] and [9]. In particular, if $Y_1, \ldots, Y_n$ are i.i.d. with common density $f_\theta(y)$, then the log likelihood is

$$\sum_{i=1}^{n} \log f_\theta(Y_i) \;=\; n \int \log f_\theta(y)\, \mathbb{P}_n(\mathrm{d}y) \;=\; -n\, \mathcal{D}\big(\mathbb{P}_n, \mathbb{P}_\theta\big) + \text{const}, \tag{8}$$

where $\mathbb{P}_n$ is the empirical distribution. Comparing the minimization problem (6) with the maximization of (8) shows that maximum likelihood results can be easily translated into minimum cross-entropy results, by replacing $\mathbb{P}_n$ with $\mathbb{P}^{(x)}$.

3 Parametric cross-entropy minimization — hazard rate twisting

In this and the next section, we study the estimation of $\mathbb{P}(Y_1 + \cdots + Y_n > x)$ where $Y_1, \ldots, Y_n$ are i.i.d. with common distribution F concentrated on $(0, \infty)$ and heavy-tailed. The method is importance sampling, where one does not look for the importance sampling distribution F^* within the class of all distributions on $(0, \infty)$ but restricts attention to a parametric class $(F_\theta)_{\theta \in \Theta}$ (θ may be multidimensional). That is, $F^* = F_{\theta^*}$ for some $\theta^* \in \Theta$.

Inspired by the classical optimality result in importance sampling, we try to choose F_{θ^*} such that $F_{\theta^*} \otimes \cdots \otimes F_{\theta^*}$ is as close as possible to the conditional distribution of $Y_1, \ldots, Y_n$ given $Y_1 + \cdots + Y_n > x$. We do this by maximum likelihood or equivalently minimum cross-entropy, plugging in the asymptotic form of the conditional distribution given by (2).

3.1 Pareto with $\gamma = 1$ fixed

We now take $\overline{F}(x) = (1+x)^{-\alpha}$. Equivalently, the density is $f(x) = \alpha(1+x)^{-\alpha-1}$. We look for F^* as another distribution of this form, with parameter α^*, say.

First, we need to compute the MLE $\widehat{\alpha}$. The log likelihood is $n \log \alpha - \alpha \sum_1^n \log(1+y_i)$, which in a straightforward way yields

$$\widehat{\alpha} \;=\; \frac{n}{\sum_1^n \log(1+y_i)} \;=\; \Big(\int_0^\infty \log(1+y) F_n(dy)\Big)^{-1},$$

where F_n is the empirical distribution.

The conditional distribution of Y given $Y > x$ has density

$$\frac{\alpha(1+x)^\alpha}{(1+y)^{\alpha+1}}, \quad y > x.$$

Thus, we take $\alpha^* = 1/J_x$, where

$$\begin{aligned} J_x &= \int_0^\infty \log(1+y)\Big(\frac{n-1}{n}\frac{\alpha}{(1+y)^{\alpha+1}} + \frac{1}{n}\frac{\alpha(1+x)^\alpha}{(1+y)^{\alpha+1}} I(y > x)\Big)\,\mathrm{d}y \\ &= \frac{n-1}{n\,\alpha} + \frac{1}{n}\left(\log(1+x) + \frac{1}{\alpha}\right) \\ &= \frac{\log(1+x)}{n} + \frac{1}{\alpha}\,. \end{aligned}$$

It follows that for large x

$$\alpha^* \approx \alpha_0^* := \frac{n}{\log(1+x)}. \tag{9}$$

This is to be compared with the suggestion of [14] to take $\alpha^* = b/\log(1+x)$, with b unspecified but arbitrary, and with that of [6] to take $\overline{F}^*(x) = 1/\log(1+x)$ which has a heavier tail and may be consider as a particular instance of the boundary case $b = 0$.

To illustrate the sensitivity to the particular choice of α^*, we performed a simulation study, taking $n = 2$ and $\alpha = 3/2$ (that is, in the range of finite mean but infinite variance which is often argued to be the one of primary interest). We considered $x = 4\,m, 16\,m, 64\,m, 256\,m$ where $m = \mathbb{E}Y = 1/(\alpha-1) = 2$ and candidates α^* of the form $2^{t/2}\,\alpha_0^*$, $t \in \{-6, \ldots, 5\}$, where α_0^* is as in (9). For each combination of values of (x, t), $R = 10,000$ replicates of (Y_1, Y_2) were produced (by inversion of the α^*–c.d.f. and using common random numbers for fixed x). The IS estimates for $\mathbb{P}(Y_1 + Y_2 > x)$ and the corresponding 95% confidence intervals are given in Figure 1 with t on the horizontal axis and $\mathbb{P}(Y_1 + Y_2 > x)$ on the vertical; the four panels correspond to the four x values in lexicographical order. The extra tick on the t axis correspond to the t–value making $\alpha^* = \alpha$, that is, to crude Monte Carlo simulation.

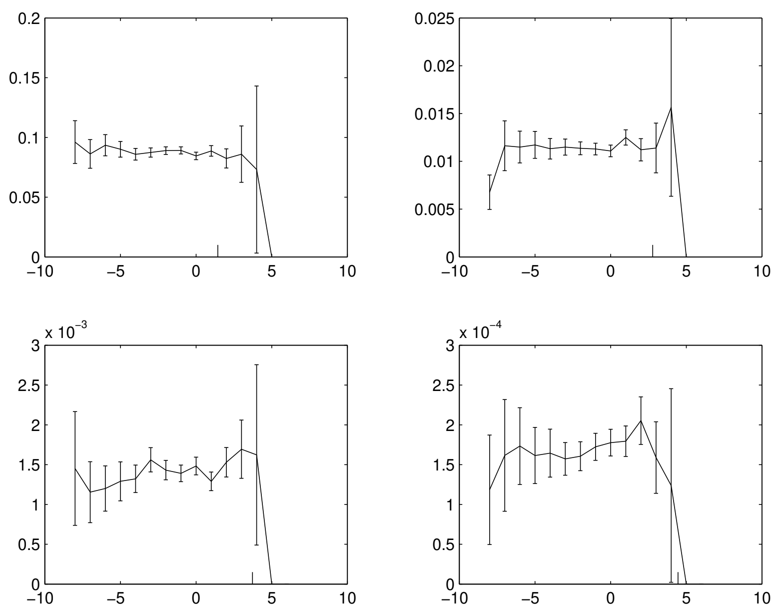

Figure 1: Estimates of $\mathbb{P}(Y_1 + Y_2 > x)$ for the Pareto case with fixed $\gamma = 1$.

A number of conclusions to be drawn from this figure are expected: the efficiency of the IS algorithm deteriorates as α^* approaches the crude Monte Carlo value α and goes beyond, and for high values of α^* the simulation estimates come out as 0, corresponding to no exceedance of x in the R replications. Also, the growing width of the confidence intervals as α^* becomes small certainly supports some of our (unpublished) numerical studies, that the choice $\overline{F}^*(x) = 1/\log(1+x)$ of [6] may well be efficient asymptotically but not in practical situations.

However, for the present purposes the main conclusion is that indeed choosing α^* by (asymptotic) minimal CE appears to be very close to variance minimality; this is of course crucial for justifying the adaptive CE algorithm in more comlex situations. Of main interest is also the degree of robustness of the choice of α^*: it is seen that there is no essential performance degradation in the interval $t \in [-3, 2]$ (at least), meaning $\alpha^* \in [0.4\alpha_0^*, 2\alpha_0^*]$.

Remark 3.1 The connection to maximum likelihood is suggestive, but of course entropy minimization can be carried out directly. In this example, the details are as follows. By taking derivatives, the solution α^* to (7) is given as the solution to

$$\mathbb{E}^{(x)} \frac{\mathrm{d}}{\mathrm{d}\alpha} \left(n \log \alpha - \alpha \sum_{1}^{n} \log(1 + Y_i) \right) = 0,$$

which is

$$\alpha^* = \frac{n}{\mathbb{E}^{(x)} \sum_i \log(1+Y_i)} \ . \qquad \square$$

3.2 Weibull with β fixed

We consider the Weibull case $\overline{F}(x) = \mathrm{e}^{-x^\beta}$ or equivalently with density $f(x) = \beta x^{\beta-1}\mathrm{e}^{-x^\beta}$ for some $0 < \beta < 1$. We write this F as F_1 where F_θ has tail $\mathrm{e}^{-\theta x^\beta}$ and look for F^* within this class of distributions.

We first need to compute the MLE $\widehat{\theta}$ of θ based upon observations $y_1, \ldots, y_n$. The density of F_θ is $\theta\beta x^{\beta-1}\mathrm{e}^{-\theta x^\beta}$ so that the log likelihood is

$$n\log\theta + n\log\beta + (\beta-1)\sum_{i=1}^n \log y_i - \theta\sum_{i=1}^n y_i^\beta .$$

Differentiating with respect to θ and letting the resulting expression equal to 0, we obtain in a straightforward way that

$$\widehat{\theta} \ = \ \frac{n}{\sum_1^n y_i^\beta} \ .$$

This can be written as

$$\Bigl(\int_0^\infty y^\beta F_n(\mathrm{d}y)\Bigr)^{-1}$$

where as above F_n is the empirical distribution.

The conditional distribution of Y given $Y > x$ has density

$$\beta y^{\beta-1}\mathrm{e}^{-(y^\beta - x^\beta)}, \quad y > x.$$

Thus, we take $\theta^* = 1/I_x$ where

$$\begin{aligned} I_x \ &= \ \int_0^\infty y^\beta \Bigl(\frac{n-1}{n}\beta y^{\beta-1}\mathrm{e}^{-y^\beta} + \frac{1}{n}\beta y^{\beta-1}\mathrm{e}^{-(y^\beta - x^\beta)} I(y > x)\Bigr)\,\mathrm{d}y \\ &= \ \frac{n-1}{n}\,c + \frac{1}{n}\,c_x \end{aligned}$$

where

$$\begin{aligned} c \ &= \ \int_0^\infty y^\beta \beta y^{\beta-1}\mathrm{e}^{-y^\beta}\,\mathrm{d}y \ = \ 1, \\ c_x \ &= \ \int_x^\infty y^\beta \beta y^{\beta-1}\mathrm{e}^{-(y^\beta - x^\beta)}\,\mathrm{d}y \ = \ x^\beta + 1 \ . \end{aligned}$$

It follows that for large x

$$\theta^* \ \approx \ \frac{n}{x^\beta} \ . \tag{10}$$

This is to be compared with the suggestion of [14] to take $\beta^* = b/x^\beta$, with b unspecified but arbitrary, and with that of [6] to take $\overline{F}^*(x)$ regularly varying which has a heavier tail and may be consider as a particular instance of the boundary case $b = 0$.

We performed a similar simulation study as for the Pareto case, only replacing the Pareto($\alpha = 3$) distribution with the Weibull($\beta = 1/3$) distribution (note that here $m = \Gamma(\beta)/\beta$). The results are in Figure 2 and the conclusions are much the same as for the Pareto case. In particular, the θ^* picked by the CE argument appears to be very close to variance minimal.

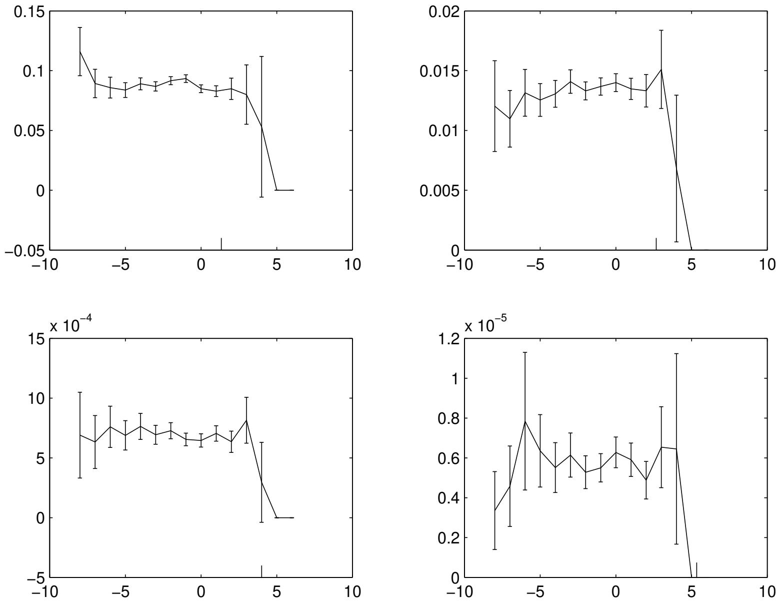

Figure 2: Estimates of $\mathbb{P}(Y_1 + Y_2 > x)$ for the Weibull case with fixed β.

4 Parametric cross-entropy minimization — scale twisting in the Pareto case

Let $\overline{F}(x) = (1 + x)^{-\alpha}$ with $\alpha > 1$ fixed, that is, with density $f(x) = \alpha(1 + x)^{-(\alpha+1)}$. We look for a change of measure with density $\alpha(1 + x/\gamma)^{-(\alpha+1)}/\gamma$. The log likelihood is

$$n \log \alpha - n \log \gamma - (\alpha + 1) \sum \log(1 + y_i/\gamma),$$

so that the MLE $\widehat{\gamma}$ is determined by

$$-\frac{n}{\widehat{\gamma}} + (\alpha+1)\sum \frac{y_i/\widehat{\gamma}^2}{1+y_i/\widehat{\gamma}} \;=\; 0\;;$$

that is,

$$\frac{1}{1+\alpha} \;=\; \frac{1}{n}\sum \frac{y_i/\widehat{\gamma}}{1+y_i/\widehat{\gamma}} \;=\; \int_0^\infty \frac{y/\widehat{\gamma}}{1+y/\widehat{\gamma}}\,F_n(\mathrm{d}y)$$

(note that the r.h.s. is a decreasing function of $\widehat{\gamma}$ with limits 1 and 0 at 0, resp. ∞, so that a solution always exists). Since $F^{(x)}$ has density $\alpha(1+x)^\alpha(1+y)^{-\alpha-1}$, the γ^* suggested by cross–entropy is determined by

$$\frac{1}{1+\alpha} \;=\; \frac{n-1}{n}\int_0^\infty \frac{y/\gamma^*}{1+y/\gamma^*}\,F(\mathrm{d}y) \;+\; \frac{1}{n}\int_x^\infty \frac{y/\gamma^*}{1+y/\gamma^*}\frac{\alpha(1+x)^\alpha}{(1+y)^{\alpha+1}}\mathrm{d}y \qquad (11)$$

There appears to be no closed solution but we computed the numerical one for $\alpha = 3/2$, $n = 2$ and the same x–values as in Section 3.1. These are given in Table 1.

x	8	32	128	512
γ^*	7.4	20.2	64.6	233.3

Table 1: Optimal scale parameters for the Pareto case with fixed $\alpha = 3/2$.

Table 1 suggests that $\gamma^* \sim \gamma_0^*$ where $\gamma_0^* = cx$, and we will verify that indeed the solution of (11) is asymptotically of this form with c the solution of

$$\int_1^\infty \frac{1}{c+u}\frac{\alpha}{u^\alpha}\,\mathrm{d}u \;=\; \frac{n}{1+\alpha} \qquad (12)$$

provided that $n < 1+\alpha$ (as in our example). To this end, note first that the first integral in (11) goes to 0 as γ^* goes to ∞. Taking $\gamma^* = cx$ and substituting $y = x + xz$, the second integral becomes

$$\int_x^\infty \frac{y}{cx+y}\frac{\alpha(1+x)^\alpha}{(1+y)^{\alpha+1}}\mathrm{d}y \;=\; \int_0^\infty \frac{1+z}{1+c+z}\frac{x\alpha(1+x)^\alpha}{(1+x+xz)^{\alpha+1}}\,\mathrm{d}z$$
$$\sim\; \int_0^\infty \frac{1+z}{1+c+z}\frac{\alpha}{(1+z)^{\alpha+1}}\,\mathrm{d}z \;=\; \int_1^\infty \frac{1}{c+u}\frac{\alpha}{u^\alpha}\,\mathrm{d}u.$$

Now just note that a similar consideration as above shows that this can be put equal to $n/(1+\alpha)$ for some c if and only if $n < 1+\alpha$.

If $n > 1+\alpha$, γ^* does surprisingly not go to ∞ but to γ_0^*, the solution of

$$\frac{1}{1+\alpha} \;=\; \frac{1}{n} + \frac{n-1}{n}\int_0^\infty \frac{y/\gamma_0^*}{1+y/\gamma_0^*}\,F(\mathrm{d}y) \tag{13}$$

This follows simply because the second integral in (11) goes to 1 as $x \to \infty$ with γ^* fixed. As example, we took $\alpha = 1/2$, $n = 2$. Since the mean is infinite, we cannot use the same x–values as above but considered 10^i, $i = 1, 2, 3, 4$. The results are displayed in Table 2.

x	10	100	1,000	10,000
γ^*	4.7	9.2	11.3	11.6

Table 2: Optimal scale parameters for the Pareto case with fixed $\alpha = 1/2$.

Numerical examples for the two examples are given in Figure 3 ($\alpha = 3/2$) and Figure 4 ($\alpha = 1/2$). They once more shows that minimizing the cross–entropy works very well for selecting a good IS parameter.

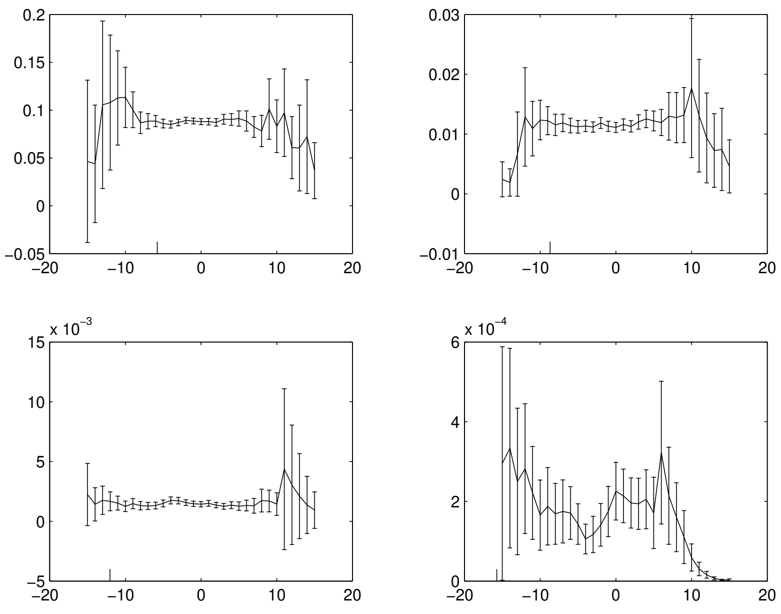

Figure 3: Estimates of $\mathbb{P}(Y_1+Y_2 > x)$ for the Pareto case with fixed $\alpha = 3/2$.

There appears to be no theoretical results in the literature concerning complexity properties of IS using a twist of γ. We next present a set of results in this direction; the first explains in particular the strange (at a first look) suggestion of the CE method, to take γ_0^* bounded if $n > \alpha + 1$.

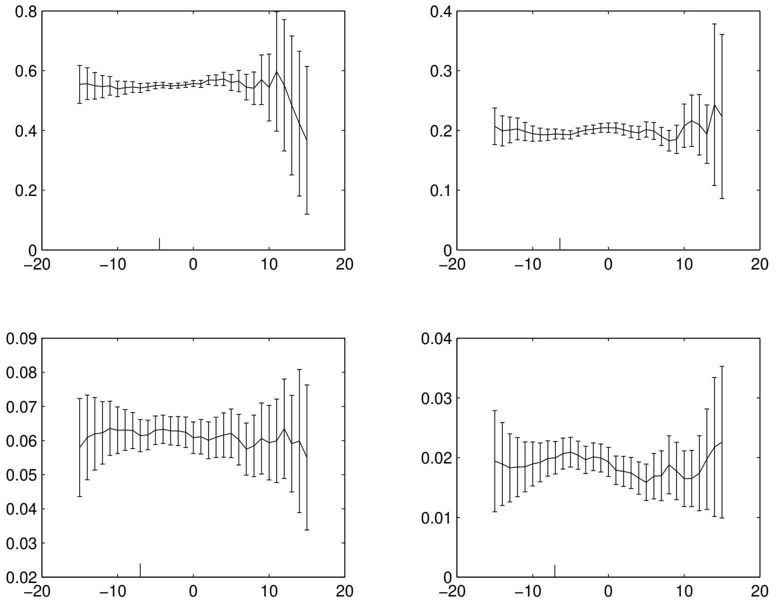

Figure 4: Estimates of $\mathbb{P}(Y_1+Y_2 > x)$ for the Pareto case with fixed $\alpha = 1/2$.

Proposition 4.1 *Consider an IS scheme given by twisting γ from 1 to $\gamma(x)$ for each x and let $Z(x,\gamma(x))$ be the corresponding estimators. Assume $n > \alpha + 1$ and that the IS is asymptotically no worse than crude Monte Carlo simulation in the sense that*

$$\limsup_{x\to\infty} \frac{\mathbb{V}\mathrm{ar}\, Z\big(x,\gamma(x)\big)}{\mathbb{V}\mathrm{ar}\, Z(x,1)} \ < \ \infty.$$

Then

$$\limsup_{x\to\infty} \gamma(x) < \infty, \quad \liminf_{x\to\infty} \gamma(x) > 0.$$

Proof. Assume first that the liminf is 0. By passing to a subsequence if necessary, one may then assume $\gamma(x) \to 0$. From

$$\mathbb{E}Z(x,\gamma(x))^2 \ = \ \int\cdots\int_{\{y_1+\cdots+y_n>x\}} \prod_{i=1}^{n} \frac{\gamma(x)(1+y_i/\gamma(x))^{\alpha+1}}{(1+y_i)^{2\alpha+2}} dy_i \qquad (14)$$

it follows that

$$\begin{aligned}\mathbb{E}Z(x,\gamma(x))^2 \ &\geq \ \int\cdots\int_{\{y_1+\cdots+y_n>x\}} \prod_{i=1}^{n} \frac{\gamma(x)(y_i/\gamma(x))^{\alpha+1}}{(1+y_i)^{2\alpha+2}} dy_i \\ &= \ \frac{1}{\gamma(x)^{n\alpha}} \int\cdots\int_{\{y_1+\cdots+y_n>x\}} \prod_{i=1}^{n} \frac{y_i^{\alpha+1}}{(1+y_i)^{2\alpha+2}} dy_i.\end{aligned}$$

Considering $\mathbb{P}(Y_1+\cdots+Y_n > x)$ for the regularly varying distribution G with density proportional to $y^{\alpha+1}/(1+y)^{2\alpha+2}$ (hence tail of order $x^{-\alpha}$) shows that the last integral is of order $x^{-\alpha}$ which is again of the same order as $\mathbb{E}Z(x,1)^2$. Hence $\mathbb{E}Z(x,\gamma(x))^2/\ \mathbb{E}Z(x,1)^2 \to \infty$. [Note that this part of the proof does not require $n > \alpha+1$].

If the limsup is ∞, we may similarly assume $\gamma(x) \to \infty$. Using (14) we get

$$
\begin{aligned}
&\mathbb{E}Z(x,\gamma(x))^2 \\
&\quad \geq \int\cdots\int_{\{y_1+\cdots+y_n>x\}} \frac{\gamma(x)(y_1/\gamma(x))^{\alpha+1}}{(1+y_1)^{2\alpha+2}}dy_1 \prod_{i=2}^{n} \frac{\gamma(x)}{(1+y_i)^{2\alpha+2}}dy_i \\
&\quad = \gamma(x)^{n-\alpha-1}\int\cdots\int_{\{y_1+\cdots+y_n>x\}} \frac{y_1^{\alpha+1}}{(1+y_1)^{2\alpha+2}}dy_1 \prod_{i=2}^{n} \frac{1}{(1+y_i)^{2\alpha+2}}dy_i.
\end{aligned}
$$

Considering $\mathbb{P}(Y_1+\cdots+Y_n > x)$ where Y_1 follows the distribution G above and $Y_2,\ldots,Y_n$ follow the lighter–tailed regularly varying distribution H with density $(2\alpha+1)/(1+y)^{2\alpha+2}$ shows that the last integral is of order $\mathbb{P}(Y_1 > x)$ (cf. [3], Lemma 1.8 p. 255) which in turn has the common order of $x^{-\alpha}$ and $\mathbb{E}Z(x,1)^2$. Hence $\mathbb{E}Z(x,\gamma(x))^2/\mathbb{E}Z(x,1)^2 \to \infty$. □

The next results supports the findings of the CE method in the case $n < \alpha+1$, to take $\gamma(x)$ of order x.

Corollary 4.1 *Consider the setting of Proposition 4.1 with $n < \alpha+1$. Then the choice $\gamma(x) = cx$ is asymptotically optimal in the sense that whenever $\limsup \gamma(x)/x = \infty$ or $\liminf \gamma(x)/x = 0$, then*

$$\limsup_{x\to\infty} \frac{\mathbb{V}\mathrm{ar}\, Z\big(x,\gamma(x)\big)}{\mathbb{V}\mathrm{ar}\, Z(x,cx)} = \infty.$$

Furthermore, $\mathbb{V}\mathrm{ar}\, Z(x,cx) \sim d(c)/x^{2\alpha+1-n}$ for some $d(c)$.

Proof. The key step is to show that $\mathbb{V}\mathrm{ar}\, Z(x,\gamma(x))$ is of order $h(x)$ where

$$h(x) \;=\; \frac{1}{\gamma(x)^{\alpha+1-n}x^{\alpha}} + \frac{\gamma(x)^n}{x^{2\alpha+1}}. \tag{15}$$

Indeed, this immediately gives the statement on $\mathbb{V}\mathrm{ar}\, z(x,cx)$ since both terms in (15) are of order $1/x^{2\alpha+1-n}$ when $\gamma(x)$ is of order x, and further, the first term is of higher order when $\liminf \gamma(x)/x = 0$ and the second of lower order when $\limsup \gamma(x)/x = \infty$.

Combining the two lower bounds in the proof of Proposition 4.1 gives $\liminf \mathbb{V}\mathrm{ar}\, Z(x,\gamma(x))/h(x) > 0$. To get $\limsup < \infty$, we use the c_r inequality

$(a+b)^r \le 2^r(a^r+b^r)$ with $r=\alpha+1$, $a=1$, $b=y_i/\gamma(x)$ to conclude as in the last part of the proof of Proposition 4.1 that

$$\mathbb{E}Z(x,\gamma(x))^2 \ \le\ \gamma(x)^n \sum_{k=0}^{n} c_k \gamma(x)^{-k(\alpha+1)} \mathbb{P}_k(Y_1+\cdots+Y_n > x)$$

where $Y_1,\ldots,Y_n$ are i.i.d. under $\mathbb{P}_k$ with distribution G of $Y_1,\ldots,Y_k$ and H of $Y_{k+1},\ldots,Y_n$. The result now follows by noting that (se again [3]) $\mathbb{P}_k(Y_1+\cdots+Y_n > x)$ is of order $x^{-2\alpha-1}$ for $k=0$ and $x^{-\alpha}$ for $k>0$ (thus the $k=2,\ldots,n$ terms are dominated by the $k=1$ term). $\square$

The results above support the usefulness of the CE method in picking a good change of measure also for IS using twist of γ. However, for the idea of twisting γ they are pessimistic since one only can achieve variance reduction under the condition $n<\alpha+1$ which is rather unnatural for any given α, not least in the important range $\alpha<2$ (infinite variance). Furthermore, even if $n<\alpha+1$ the order of $\mathbb{V}\mathrm{ar}\, Z(x,cx)$ is always higher than $x^{-2\alpha}$ in the non–trivial case $n>1$ so that the complexity can never be polynomial. These negative observations are further supported by:

Corollary 4.2 *Consider IS for the M/Pareto/1 queue using simulation from the PK formula with twisted γ and let $Z_{\mathrm{PK}}\big(x,\gamma(x)\big)$ denote the corresponding estimator. Then no choice of the $\gamma(x)$ can achieve asymptotic variance reduction. That is, one always has*

$$\liminf_{x\to\infty} \frac{\mathbb{V}\mathrm{ar}_{\mathrm{PK}}(Z(x,\gamma(x))}{\mathbb{V}\mathrm{ar}_{\mathrm{PK}}(Z(x,1)} \ >\ 0.$$

Proof. Just note that the algorithm means estimating the tail $\mathbb{P}(W>x)$ of the stationary waiting time W by $Z_{\mathrm{PK}}(x,1)=I(Y_1+\cdots+Y_N>x)$ where the Y_i follow the integrated tail distribution (which is Pareto with α changed to $\alpha-1$) and N is an independent geometric r.v. Thus, from above we have that the contribution to $\mathbb{V}\mathrm{ar}_{\mathrm{PK}}(Z(x,\gamma(x))$ from the event $N>1+\alpha$ is of the same order as $\mathbb{V}\mathrm{ar}_{\mathrm{PK}}(Z(x,1))$. $\square$

In conclusion, twisting γ may provide some modest variance reduction for a given x but a twist of α appears the more promising approach.

5 Other ideas for selecting IS parameters

A familiar idea from statistics is to replace ML estimation by the often simpler device of moment fitting. As a simple example, consider the Pareto case with

$\gamma = 1$ fixed as in Section 3.1. Here

$$\mathbb{E}_\alpha Y \;=\; \frac{1}{\alpha - 1}, \quad \mathbb{E}_\alpha[Y \,|\, Y > x] \;=\; \frac{\alpha x + 1}{\alpha - 1},$$

so that the moment method suggest to determine the α^* for importance sampling by means of

$$\begin{aligned} \frac{1}{\alpha^* - 1} \;&=\; \mathbb{E}_{\alpha^*} Y \;=\; \frac{n-1}{n}\mathbb{E}_\alpha Y \;+\; \frac{1}{n}\mathbb{E}_\alpha[Y \,|\, Y > x] \\ &=\; \frac{n-1}{n}\frac{1}{\alpha - 1} \;+\; \frac{1}{n}\frac{\alpha x + 1}{\alpha - 1}, \end{aligned}$$

i.e.

$$\alpha^* \;=\; \frac{\alpha(n + x)}{n + \alpha x}.$$

Thus $\alpha^* \to 1$ which cannot lead to polynomiality. The simplicity of this example thus indicates that the moment method is unlikely to become useful.

Yet another idea is to make the $\mathbb{P}_{\alpha^*}$-distribution of $Y_1, \ldots, Y_n$ alike the $\mathbb{P}^{(x)}$–distribution by equating to 1 the expected number of Y_i with $Y_i > x$. In the same Pareto example, this gives $n/x^{\alpha^*} = 1$, i.e. $\alpha^* = \log n / \log x$. For the Weibull example in Section 3.2, one gets $\gamma^* = \log n / x^\beta$. Thus, in both cases the asymptotic forms are $b/\log \overline{F}(x)$ as in [14] so that polynomiality holds. However, the numerical results above indicate that cross–entropy minimization is superior in terms of finding the optimal b.

Finally, in the Pareto scale example in Section 4, one gets $\gamma^* = x/(n^{1/\alpha} - 1)$.

6 Exponential complexity for $\mathbb{P}(\tau(x) < \infty)$ for the GI/G/1 queue

Let $S_n = X_1 + \cdots + X_n$ be a random walk (RW) such that $X_k = U_k - T_k$ where the U_k are i.i.d. with tail $(1+x)^{-\alpha}$ or equivalently with common density $f_\alpha(x) = \alpha/(1+x)^{\alpha+1}$ for some $\alpha > 1$ and the T_k are i.i.d. (and independent of the U_k) with mean $\mathbb{E}T > 1/(\alpha - 1)$ so that $\mathbb{E}X < 0$ and $\mathbb{P}(\tau(x) < \infty) \sim c/x^{\alpha-1}$ where $\tau = \tau(\gamma) = \inf\{n : S_n > \gamma\}$, see e.g. [4] Theorem 9.1 p. 296 ($\mathbb{P}(\tau(x) < \infty)$ is also the probability that the waiting time exceeds x in the GI/G/1 queue).

Let $\alpha_* = \alpha^*(x)$ be candidates for the IS parameter, satisfying $\mathbb{P}_{\alpha_*}(\tau < \infty) = 1$ (that is, $\alpha_* \le \alpha_0$ where $\alpha_0 = 1 + 1/\mathbb{E}T$). The IS estimator is

$$Z_* \;=\; Z_*(x) \;=\; \prod_{n=1}^{\tau} \frac{f_\alpha(U_n)}{f_{\alpha_*}(U_n)}.$$

Theorem 6.1 *Assume that $\mathbb{E}e^{rT} < \infty$ for some $r > 0$. Then the estimator Z_* cannot be polynomial for any choice of $\alpha_* = \alpha_*(x) \le \alpha_0$.*

Lemma 6.1 $\mathbb{E}_{\alpha_*} Z_*^2 \;=\; \mathbb{E}_{2\alpha-\alpha_*}\bigl[c^\tau;\, \tau < \infty\bigr]$ *where* $c \;=\; \dfrac{\alpha^2}{\alpha_*(2\alpha-\alpha_*)}$.

Proof. The argument is a small extension of similar steps in [6], [14] and [19], but is given here for the sake of completeness. Let $\mathbb{E}^{\mathbf{t}}$ be the conditional expectation given $T_1 = t_1, T_2 = t_2, \ldots$ and

$$A_k \;=\; \Bigl\{(u_1,\ldots,u_k) : \sum_{n=1}^{k}(u_n - t_n) > x,\; \sum_{n=1}^{\ell}(u_n - t_n) \le x \text{ for } \ell < k\Bigr\}.$$

Then

$$\begin{aligned}
\mathbb{E}^{\mathbf{t}}_{\alpha_*} Z_*^2 \;&=\; \sum_{k=1}^{\infty} \mathbb{E}^{\mathbf{t}}_{\alpha_*}\Bigl[\prod_{n=1}^{k} \frac{f_\alpha^2(U_n)}{f_{\alpha_*}^2(U_n)};\, \tau = k\Bigr] \\
&=\; \sum_{k=1}^{\infty} \int\!\cdots\!\int_{A_k} \prod_{n=1}^{k} \frac{f_\alpha^2(u_n)}{f_{\alpha_*}^2(u_n)}\, f_{\alpha_*}(u_1)\ldots f_{\alpha_*}(u_k)\, du_1 \ldots du_k] \\
&=\; \sum_{k=1}^{\infty} \int\!\cdots\!\int_{A_k} \prod_{n=1}^{k} \frac{\alpha^2}{\alpha_*(1+u_n)^{2\alpha-\alpha^*+1}}\, du_1 \ldots du_k] \\
&=\; \sum_{k=0}^{\infty} c^k \int\!\cdots\!\int_{A_k} f_{2\alpha-\alpha_*}(u_1)\ldots f_{2\alpha-\alpha_*}(u_k)\, du_1 \ldots du_k \\
&=\; \sum_{k=1}^{\infty} c^k \mathbb{P}^{\mathbf{t}}_{2\alpha-\alpha_*}(\tau = k) \;=\; \mathbb{E}^{\mathbf{t}}_{2\alpha-\alpha_*}\bigl[c^\tau;\, \tau < \infty\bigr].
\end{aligned}$$

Integrating $T_1 = t_1, T_2 = t_2, \ldots$ out, the result follows. □

Proof of Theorem 6.1. Let $\mathcal{G}_n = \sigma\bigl(U_1,\ldots,U_{n-1},T_1,\ldots,T_n\bigr)$, $B_k = \{T_1 + \cdots + T_k \le k\mu\}$. Then for each k,

$$\begin{aligned}
\mathbb{E}_{\alpha_*} Z_*^2 \;&\ge\; c^k \mathbb{P}_{2\alpha-\alpha_*}(\tau = k) \\
&=\; c^k \mathbb{P}_{2\alpha-\alpha_*}\bigl(\tau > k-1, U_1 + \cdots + U_k > x + T_1 + \cdots + T_k\bigr) \\
&\ge\; c^k \mathbb{P}_{2\alpha-\alpha_*}\bigl(\tau > k-1, U_1 > x + T_1 + \cdots + T_k, B_k\bigr) \\
&\ge\; c^k \overline{F}(x + k\mu)\mathbb{P}_{2\alpha-\alpha_*}\bigl(\tau > k-1, B_k\bigr) \\
&\ge\; c^k \overline{F}_{2\alpha-\alpha_*}(x + k\mu)\bigl[\mathbb{P}_{2\alpha-\alpha_*}(\tau > k-1) - \mathbb{P}(B_k^c)\bigr].
\end{aligned}$$

Choose $\mu > \mathbb{E}T$. Then the assumption $\mathbb{E}e^{rT} < \infty$ implies by standard large deviations estimates (e.g. [4] p. 355) that $\mathbb{P}(B_k^c)$ goes to 0 exponentially fast.

Further, since $2\alpha - \alpha_* \geq 2\alpha - \alpha_0 \geq \alpha$, $P_{2\alpha-\alpha_*}(\tau > k-1)$ is bounded from below by $\mathbb{P}_\alpha(\tau > k-1) \geq \mathbb{P}_\alpha(\tau = \infty)$ which goes to 1 as $k \to \infty$. Taking $k = k(x) = x$, we get the asymptotic lower bound

$$c^x \overline{F}_{2\alpha-\alpha_*}\big(x(1+\mu)\big) \;\geq\; c^x \frac{1}{\big(1 + x(1+\mu)\big)^{2\alpha}}$$

for $\mathbb{E}Z_*^2$ which rules out polynomiality since $c > 1$ (note that the quadratic $\alpha_*(2\alpha - \alpha_*)$ attains it maximum α^2 at $\alpha_* = \alpha$ so that a lower bound for the α_* in question is $\alpha_0(2\alpha - \alpha_0) < \alpha^2$). □

For the switching regenerative estimator, consider now $\tau^s = \tau \wedge \tau_-$ where $\tau_- = \inf\{n > 0 : S_n \leq 0\}$ is the descending ladder epoch.

Theorem 6.2 *Assume that $\mathbb{E}e^{rT} < \infty$ for some $r > 0$. Then the estimator*

$$Z_*^s \;=\; I(\tau < \tau_-) \prod_{n=1}^{\tau^s} \frac{f_\alpha(U_n)}{f_{\alpha_*}(U_n)}$$

for $\mathbb{P}(\tau < \tau_-)$ cannot be polynomial for any choice of $\alpha_* = \alpha_*(x) \leq \alpha_0$.

Proof. It is shown in [2] that $\mathbb{P}(\tau < \tau_-) \sim \mathbb{E}\tau_-/(1+x)^\alpha$. Exactly as above,

$$\begin{aligned}
\mathbb{E}_{\alpha^*} Z_*^s \;&\geq\; c^k \overline{F}(\gamma + k\mu)\mathbb{P}_{2\alpha-\alpha_*}\big(\tau^s > k-1, B_k\big) \\
&\geq\; c^k \overline{F}(\gamma + k\mu)\big[\mathbb{P}_{2\alpha-\alpha_*}(\tau > k-1) - P_{2\alpha-\alpha_*}(\tau_- > k-1) - \mathbb{P}(B_k^c)\big].
\end{aligned}$$

Here the second term in $[\cdots]$ is uniformly small in α^* for large k, and the proof is completed exactly as above. □

Let next $F = F_1$ where F_θ is the Weibull distribution with tail $e^{-\theta x^\beta}$ where $0 < \beta < 1$ is fixed. Let $\theta_0 < 1$ correspond to 0 drift and consider a change of measure where the IS distribution is $F_{\theta_*} = F_{\theta_*(x)}$ where $\theta_* \leq \theta_0$.

Theorem 6.3 *The IS scheme given by the θ_* cannot be polynomial, neither for $\mathbb{P}(\tau < \infty)$ nor for $\mathbb{P}(\tau < \tau_-)$.*

Proof. From $f_\theta(x) = \theta\beta x^{\beta-1}e^{-\theta x^\beta}$ we get

$$\frac{f_1^2(x)}{f_{\theta_*}(x)} \;=\; \frac{1}{\theta_*}\beta x^{\beta-1}e^{-(2-\theta_*)x^\beta} \;=\; \frac{1}{\theta_*(2-\theta_*)} f_{2-\theta_*}(x).$$

With $c = 1/\theta_*/(2-\theta_*)$, we have $c \leq 1/\theta_0/(2-\theta_0) < 1$ because of $\theta_0 < 1$ and get as before

$$\begin{aligned}
\mathbb{E}_{\alpha_*} Z_*^2 \;&=\; \mathbb{E}_{2-\theta_*}\big[c^\tau;\, \tau < \infty\big] \;\geq\; c^k \mathbb{P}_{2\theta-\theta_*}(\tau = k) \\
&\geq\; c^k \overline{F}_{2\theta-\theta_*}(x + k\mu)\big[\mathbb{P}_{2\theta-\theta_*}(\tau > k-1) - \mathbb{P}(B_k^c)\big] \\
&\geq\; c^k e^{-(2\theta-\theta_*)(x+k\mu)^\beta}\big[\mathbb{P}_{2\theta-\theta_*}(\tau > k-1) - \mathbb{P}(B_k^c)\big].
\end{aligned}$$

Taking $k = x^\alpha$ where $1-\beta < \alpha < 1$, one has $\mathbb{P}_{2\theta-\theta_*}(\tau > k-1) \to 1$, cf. [4] Th. 6.5 p. 405, and the rest of the argument is now precisely as for the Pareto case. □

Remark 6.1 [10] shows polynomiality of the same algorithm truncated to terminate at latest at time $cx^{1-\beta}$ for some large c. Of course, this is no contradiction, cf. the way k was chosen in the proof. □

References

[1] R.J. Adler, R. Feldman & M.S. Taqqu (1998). *A User's Guide to Heavy Tails.* Birkhäuser.

[2] S. Asmussen (1998). Subexponential asymptotics for stochastic processes: extremal behaviour, stationary distributions and first passage probabilities. *Ann. Appl. Probab.* **8**, 354–372.

[3] S. Asmussen (2000). *Ruin Probabilities.* World Scientific.

[4] S. Asmussen (2003). *Applied Probability and Queues.* Springer–Verlag.

[5] S. Asmussen & K. Binswanger (1997). Simulation of ruin probabilities for subexponential claims. *ASTIN Bulletin* **27**, 297–318.

[6] S. Asmussen, K. Binswanger & B. Højgaard (2000). Rare events simulation for heavy–tailed distributions. *Bernoulli* **6**, 303–322.

[7] S. Asmussen & R.Y. Rubinstein (1995). Steady–state rare events simulation in queueing models and its complexity properties. *Advances in Queueing: Models, Methods and Problems* (J. Dshalalow, ed.), 429-466. CRC Press

[8] P. T. de Boer, D. P. Kroese & R. Y. Rubinstein (2003). Estimating buffer overflows in three stages using cross-entropy. In *Proceedings of the 2002 Winter Simulation Conference, San Diego*, 301–309.

[9] P.T. de Boer, D.P. Kroese, S. Mannor & R.Y. Rubinstein (2003). A tutorial on the cross-entropy method. *Annals of Operations Research.* Submitted.

[10] N.K. Boots & P. Shahabuddin (2002). Simulating tail probabilities in GI/G/1 queues and insurance risk processes with subexponential distributions. Submitted to *Opns. Res.*. Short version published in *Proceedings of the 2001 Winter Simulation Conference, San Diego*, pp. 468–476.

[11] P. Embrechts, C. Klüppelberg & T. Mikosch (1997). *Modeling Extremal Events for Finance and Insurance.* Springer–Verlag.

[12] P. Heidelberger (1995). Fast simulation of rare events in queueing and reliability models. *ACM TOMACS 6*, 43-85.

[13] T. Homem de Mello & R.Y. Rubinstein (2002). Rare event probability estimation for static models via cross-entropy and importance sampling. Submitted.

[14] S. Juneja & P. Shahabuddin (2002). Simulating heavy tailed processes using delayed hazard rate twisting. *ACM TOMACS* **12**, 94–118.

[15] D.P. Kroese & R.Y. Rubinstein (2003). The transform likelihood ratio method for rare event simulation with heavy tails. *QUESTA* (submitted).

[16] R. Y. Rubinstein (1997). Optimization of computer simulation models with rare events. *European Journal of Operations Research*, **99**, 89–112.

[17] R. Y. Rubinstein (1999). The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, **2**, 127–190.

[18] R.Y. Rubinstein & D.P. Kroese (2003) *The Cross-Entropy Method. A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning.* Book manuscript.

[19] R.Y. Rubinstein and B. Melamed (1998). *Modern Simulation and Modeling.* Wiley.

[20] K. Sigman (1999). A primer on heavy–tailed distributions. *QUESTA* **33**, 261–275.