MaPhySto
The Danish National Research Foundation:
Network in Mathematical Physics and Stochastics

Research Report
no. 3 February 2004

Søren Asmussen
and Dirk P. Kroese:

Improved Algorithms for Rare Event Simulation with Heavy Tails

---

ISSN 1398-2699

Improved algorithms for rare event simulation with heavy tails

Søren Asmussen* ◇ Dirk P. Kroese†

Abstract

The estimation of $\mathbb{P}(S_n > u)$ by simulation where S_n is a sum of i.i.d. r.v.'s $Y_1, \ldots, Y_n$ is of importance in many applications. We propose two simulation estimators based upon the identity $\mathbb{P}(S_n > u) = n\,\mathbb{P}(S_n > u, M_n = Y_n)$ where $M_n = \max(Y_1, \ldots, Y_n)$. One estimator uses importance sampling (for Y_n only), the other conditional Monte Carlo conditioning upon $Y_1, \ldots, Y_{n-1}$. Properties of the relative error of the estimators are derived and a numerical study given in terms of the M/G/1 queue where n is replaced by an independent geometric r.v. N. The conclusion is that the new estimators compare extremely favourable with previous ones. In particular, the conditional Monte Carlo estimator is the first heavy-tailed example of an estimator with bounded relative error. Further improvements are obtained in the random N case, by incorporating control variates and stratification techniques into the new estimation procedures.

Key words: Bounded relative error, complexity, conditional Monte Carlo, control variate, logaritmic efficiency, M/G/1 queue, Pollaczeck-Khinchine formula, rare event, regular variation, stratification, 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

◇Partially supported by MaPhySto — A Network in Mathematical Physics and Stochastics, funded by the Danish National Research Foundation

1 Introduction

This paper is concerned with the evaluation of $z(u) = \mathbb{P}(S_n > u)$ by simulation, where $S_n = Y_1 + \cdots + Y_n$ with the $Y_i > 0$ i.i.d. and heavy-tailed, in situations where u is large so that $z(u)$ is small. We will also consider the case where n is replaced by an independent integer-valued r.v. N. An example where this is of relevance is the steady-state waiting time of the M/G/1 queue, which according to the Pollaczeck-Khinchine formula has the same distribution as S_N for a certain choice of parameters (see [4] p. 237). In insurance risk $\mathbb{P}(S_N > u)$ is also a representation of the ruin probability with initial reserve u ([4] p. 399). A third example (from financial mathematics) is credit risk; here N could be the number of defaulted obligors in a period and Y_i the loss from the ith default [14]. In the last two examples, the interesting values of $z(u)$ may be of order 10^{-2}. In the first (arising from problems in telecommunications and data transmission where $z(u)$ could be a bit loss rate) the magnitude could go down to say 10^{-10}. As is well known [9, 15], the efficiency of crude Monte Carlo simulation greatly deterioates as $z(u)$ decreases, making the simulation a non-trivial problem requiring variance reduction ideas.

Let F denote the common distribution of the Y_i. Heavy tails means in wide terms that exponential moments fail to exist. However, most often a more narrow framework is used like F being subexponential [12, 20] or even just the special case of primary importance where F is regularly varying with a relatively small index α; this means that the tail $\overline{F}(x)$ is $L(x)/x^\alpha$ for some slowly varying function $L(\cdot)$. We will not go into the general subexponential framework here but be satisfied with treating the regularly varying case as well as what is maybe the secondmost important example: the heavy-tailed Weibull case where $\overline{F}(x) = \mathrm{e}^{-u^\beta}$ with $0 < \beta < 1$. The application relevance of such modeling assumptions has been vividly argued; see e.g. [12, 2].

We call a r.v. $Z = Z(u)$ an *estimator* for $z(u)$ if Z can be generated by simulation and $\mathbb{E}^* Z = z(u)$ where $\mathbb{P}^* = \mathbb{P}^{*,u}$ is the probability measure used in the simulation (the given distribution of $Y_1, Y_2, \ldots$ or an importance sampling distribution). We use the standard terms *bounded relative error* for $\mathbb{V}\mathrm{ar}(Z(u))/z(u)^2$ being bounded in u and *polynomial time* for $\mathbb{V}\mathrm{ar}(Z(u))/z(u)^{2-\epsilon}$ being bounded in u for any $\epsilon > 0$ (often also *logarithmic efficient* or just *efficient* is used [9, 15, 6, 16]). Similar terminology is used for the case of a random N. With light tails, the most established approach for simulation of $z(u)$ is the *exponential change of measure*, as determined by the saddlepoint method, see [4] pp. 373–376. As discussed there, this scheme can be seen as an implementation of the general principle in importance sampling to take the changed measure used for the simulation as

close as possible to the conditional distribution given the rare event. In the present case this means sampling $Y_1, \ldots, Y_n$ using an asymptotic description of their conditional distribution $\mathbb{P}^{n,u}$ given $S_n > u$. The traditional description in the subexponential setting states that one Y_i is larger than u and the rest are in some sense "typical", that is, unaffected by the conditioning (for precise statements in this direction, see Proposition 1.2 p. 252 in [3], and Lemma 6.6 p. 405 in [4]). However, as noted in [6], the most straightforward ideas of using this asymptotics as basis for importance sampling fail. The first polynomial time algorithm reported in the literature, [5], in fact uses a different idea, namely conditional Monte Carlo, invoking the order statistics $Y_{(1)} < \cdots < Y_{(n)}$ for the conditioning. The estimator is

$$\mathbb{P}(S_n > u \mid Y_{(1)}, \ldots, Y_{(n-1)}) \;=\; \frac{\overline{F}\big(Y_{(n-1)} \vee (u - S_{(n-1)})\big)}{\overline{F}(Y_{(n-1)})}, \tag{1}$$

where $S_{(n-1)} = Y_{(1)} + \cdots + Y_{(n-1)}$. Later, polynomial time importance sampling ideas were given in [6] and [16]. We only consider the (more efficient) ideas of [16], which are based upon the hazard rate $\Lambda(x) = -\log \overline{F}(x)$. A key ingredient (not the only one!) in the algorithms of [16] is *hazard rate twisting* which changes $\Lambda(x)$ to $\theta\Lambda(x)$ for some small θ. That is, the tail $\overline{F}$ is changed to $\overline{F}_\theta = \overline{F}^\theta$. The more refined *weighted delayed hazard rate twisting* simulates the Y_i from a distribution which is a mixture of F conditioned to $(0, x^*]$ and F_θ conditioned to (x^*, ∞), with the weights, θ and x^* chosen to depend appropriately on u.

Despite the good computational and theoretical properties of these algorithms reported in [16], it appears intuitively unnatural that the importance sampling change of measure is i.i.d. Indeed, the above description of $\mathbb{P}^{n,u}$ is highly asymmetric, showing the particular role taken by one of the Y_i (the big one). The contribution of this paper is to present some changed algorithms which takes this into account, and which will turn out to yield efficiency improvements over existing algorithms which are very substantial. Our algorithms depart from the identity

$$\mathbb{P}(S_n > u) \;=\; n\,\mathbb{P}(S_n > u, M_n = Y_n),$$

where we use the notation $M_k = \max(Y_1, \ldots, Y_k)$, so that $Y_{(n)} = M_n$. If f is the density of F and f^* the density of an importance sampling distribution (we return to the choice of f^* later), we twist only the distribution of Y_n and arrive at the estimator

$$n\,\frac{f(Y_n)}{f^*(Y_n)}\,I(S_n > u, M_n = Y_n). \tag{2}$$

Using instead conditional Monte Carlo yields the estimator

$$n\,\mathbb{P}\big(S_n > u, M_n = Y_n \mid Y_1, \ldots, Y_{n-1}\big) \;=\; n\,\overline{F}\big(M_{n-1} \vee (u - S_{n-1})\big). \qquad (3)$$

The paper is organized as follows. In Section 2, we give a theoretical proof of the efficiency of our estimators for the Pareto case and present some numerical studies. Section 3 addresses the same issues for the Weibull case where it turns out that there is a certain critical value of β for (3) to be polynomial time. The empirical performance of (2) and in particular (3) is excellent. For example for an M/G/1 queue with Pareto service times of index $\alpha = 1.5$, Table 1 below shows that (3) reduces the variance by a factor 5–14 compared to [5] and 33-529 compared to [16]. The performance improvement for Weibull tails compared to [16] is similar, whereas the algorithm of [5] is not even polynomial in this case. In addition to this, we would like to point out the simplicity of (3), making the code much shorter and more transparent than for some of the other estimators we compare with.

Section 4 discusses combinations with other variance reduction ideas (control variates and stratification) in the case that $n = N$ is random as in the examples above.

2 The regularly varying case

In this section we assume $f(x) = L(x)/(1+x)^{1+\alpha}$, $x > 0$, with L slowly varying and $\alpha > 0$ (we don't need conditions like $\alpha \geq 1$ to exclude infinite mean, etc.). Then $\overline{F}(x) \sim L(x)/(1+x)^{\alpha}$ by Karamata's theorem ([13]). Here we use the notation $a(x) \sim b(x) \Leftrightarrow \lim_{x\to\infty} a(x)/b(x) = 1$.

We first consider the estimator (2). It remains to specify the importance sampling density $f^*(y)$. As in [16, 8], we take f^* to be regularly varying with index α^* of the form $b/\log u$, for simplicity just Pareto so that $f^*(x) = \alpha^*/(1+x)^{1+\alpha^*}$.

Theorem 2.1 (a) *With f^* as stated, the estimator* (2) *is polynomial time for any fixed n;* (b) *more precisely, if there exist u_0, z_0 such that $L(uz)/L(u)$ is either monotonically increasing or monotonically increasing in $u \geq u_0$ for all $z \geq z_0$, then an asymptotic upper bound for the squared relative error is*

$$c_1 n^2 \log u,$$

where c_1 is a constant independent of n. (c) *In the random N case the estimator is polynomial time provided $\mathbb{E}N^{3\alpha+3} < \infty$.*

[Some discussion of the condition of (b) is given in Remark 2.1 below].

In the proof, we need the density g^* proportional to $L(x)^2/(1+x)^{1+2\alpha-\alpha^*}$ (note the dependence on u via $\alpha^* = b/\log u$). That is, $g^* = f^2/f^*/c^*$ where

$$c^* \;=\; \int f^2/f^* \;\sim\; c_2 \log u \;\text{ where }\; c_2 \;=\; b^{-1}\int_0^\infty \frac{L(x)^2}{(1+x)^{2\alpha+1}}\,\mathrm{d}x,$$

Lemma 2.1 *Let Y^* have density g^*. Then* (a) *for any $\epsilon > 0$, there is a constant c_ϵ such that $\mathbb{P}(Y^* > u) \le\; c_\epsilon(1+u)^{-(2\alpha-\epsilon)}$ for all large u;* (b) *under the conditions of Theorem 2.1(b), $\mathbb{P}(Y^* > u) \sim\; c_3\,\overline{F}(u)^2\,\log u$.*

Proof. We have

$$\mathbb{P}(Y^* > u) \;\sim\; c_4\int_u^\infty \frac{L(z)^2}{(1+z)^{2\alpha-\alpha^*+1}}\,\mathrm{d}z \;\sim\; c_4\int_u^\infty \frac{L(z)^2}{z^{2\alpha-\alpha^*+1}}\,\mathrm{d}z\,.$$

From this (a) follows by noting that we can bound α^* by ϵ for u large enough and that $\int_u^\infty L(z)^2/z^{1+\beta}\,\mathrm{d}z \sim L(u)^2/u^\beta$ by Karamata's theorem. For (b), substitute $z = uy$ to get

$$\mathbb{P}(Y^* > u) \;\sim\; c_4\frac{L(u)^2}{u^{2\alpha-\alpha^*}}\int_1^\infty \frac{L(uy)^2/L(u)^2}{(1+y)^{2\alpha-\alpha^*+1}}\,\mathrm{d}y\,.$$

Here $u^{\alpha^*} \to b$ whereas $L(uy)/L(u) \to 1$ for all y. The convergence is in fact uniform on $[1, z_0]$ so that using monotone convergence on (z_0, ∞) ensures the existence of a limit of the integral. This completes the proof. □

Proof of Theorem 2.1. Let $A = \{x_1 + \cdots + x_n > u, x_n = \max_{k\le n} x_k\}$. The second moment of (2) is then

$$\begin{aligned} &n^2\int\ldots\int_A \frac{f^2(x_n)}{f^*(x_n)}\,\mathrm{d}x_n\prod_{i=1}^{n-1} f(x_i)\,\mathrm{d}x_i \\ &\quad=\; n^2c^*\int\ldots\int_A g^*(x_n)\,\mathrm{d}x_n\prod_{i=1}^{n-1} f(x_i)\,\mathrm{d}x_i \;=\; n^2\,c^*\,\mathbb{P}^{**}(M_n = Y_n, S_n > u), \end{aligned}$$

where $\mathbb{P}^{**}$ is the probability measure under which the Y_i are independent with Y_n having density g^* and the rest f.

Let $\beta \in (2/3, 1)$ and let $u' = u^\beta/(n-1)$, $u'' = u-(n-1)u' = u(1-u^{\beta-1})$. Then $Y_n > u''$ on the event $\{M_{n-1} \le u'\}$. It follows that

$$\begin{aligned} \mathbb{P}^{**}(S_n > u, M_n = Y_n) \;&\le\; \mathbb{P}^{**}\big(Y_n > u''\big) \;+\; \mathbb{P}^{**}\big(M_{n-1} > u',\, Y_n > u'\big) \\ &\le\; \mathbb{P}^{**}\big(Y_n > u''\big) \;+\; n\,\mathbb{P}^{**}(Y_n > u')\,\mathbb{P}^{**}(Y_1 > u'). \end{aligned}$$

Choose $\delta, \epsilon > 0$ such that $2\alpha + \delta < 2\alpha + \delta + \epsilon\beta < 3\alpha\beta$. Then using Lemma 2.1(a), the second term can be bounded by

$$c_5\, n \Big(\frac{n-1}{u^\beta}\Big)^{2\alpha-\epsilon} \Big(\frac{n-1}{u^\beta}\Big)^{\alpha-\epsilon} \;=\; c_5\, n\, \frac{(n-1)^{3\alpha-2\epsilon}}{u^{3\alpha\beta-2\epsilon\beta}} \;\le\; c_5\, n\, \frac{(n-1)^{3\alpha-2\epsilon}}{u^{2\alpha+\delta}}.$$

Again by Lemma 2.1(a), $\big(u(1-u^{\beta-1})\big)^{-(2\alpha-\epsilon)}$ is an upper bound for the first term. Since the asymptotics of this expression is $u^{-(2\alpha-\epsilon)}$, part (a) of the theorem follows.

Part (b) of the theorem follows now immediately by just invoking part (b) rather than part (a) of Lemma 2.1. For part (c), just condition upon $N = n$ and use dominated convergence to control the second term. □

Remark 2.1 The improvement over hazard rate twisting obtained when using (2) can be understood as follows. Assume for simplicity that F is Pareto with density $\alpha/(1+x)^{\alpha+1}$, $x > 0$, and that n is fixed. The second moment in hazard rate twisting is

$$\begin{aligned}&\int\cdots\int_{x_1+\cdots+x_n>u} \prod_{i=1}^n \frac{f^2(x_i)}{f^*(x_i)}\,\mathrm{d}x_i \\ &= \Big(\frac{\alpha^2}{\alpha^*\alpha_x}\Big)^n \int\cdots\int_{x_1+\cdots+x_n>u} \prod_{i=1}^n f_u(x_i)\,\mathrm{d}x_i \;=\; \Big(\frac{\alpha^2}{\alpha^*\alpha_u}\Big)^n \overline{F}_u^{*n}(x),\end{aligned}$$

letting F_u be the Pareto distribution with parameter $\alpha_u = 2\alpha - \alpha^*$.

Easy modifications of arguments in [13] pp. 278–279 show that $\overline{F}_u^{*n}(u) \sim n\,\mathrm{e}^b\, u^{-2\alpha}$, as is expected from $u^{\alpha^*} \to \mathrm{e}^b$. Thus the asymptotics of the second moment is

$$\mathrm{e}^b \Big(\frac{\alpha}{2b}\Big)^n n(\log u)^n \overline{F}(u)^2.$$

Thus, comparison with Theorem (2.1) shows that (2) improves the squared relative error by a factor of $(\log u)^{n-1}$.

□

An immediate corollary of Theorem 2.1 is that the estimator (3) is also polynomial time. To this end, just note that (3) is the conditional expectation of (2) given $Y_1, \ldots, Y_{n-1}$ (the choice of f^* is immaterial for this) and the standard fact that conditioning reduces variance. In fact, a stronger result holds:

Theorem 2.2 *The estimator* (3) *has bounded relative error, assuming in the case of a random N that*

$$\limsup_{u\to\infty} \frac{\mathbb{E}\big[L(u/2N)N^{2\alpha+4}\big]}{L_1(u)} \;<\; \infty, \tag{4}$$

for some function $L_1(u)$ *satisfying* $L_1(u) \sim L(u)$.

Remark 2.2 The condition (4) is in the examples we looked at equivalent to or only marginally stronger than $\mathbb{E}N^{2\alpha+4} < \infty$. To see this, note first that $L(u/2N)/L_1(u) \to 1$, as follows from $L(u)$ being slowly varying. Thus a sufficient condition for the l.h.s. of (4) being equal to $\mathbb{E}N^{2\alpha+4}$ is $L(ut)/L_1(u)$ being bounded or monotone in u for a fixed t (which is essentially the condition of Theorem 2.1). If e.g. $L(x) = (\log(a+x))^\beta$, we can let $L_1(x) = (\log x)^\beta$ and have

$$\frac{L(ut)}{L_1(u)} \;=\; \Big(1 + \frac{\log(t + a/u)}{\log u}\Big)^\beta,$$

which is decreasing for $\beta > 0$ and increasing for $\beta < 0$ (note that $\log(t + a/u)/\log u$ is decreasing, it being the ratio between a decreasing and an increasing function). □

Proof of Theorem 2.2. We split the second moment of (3) (with $n+1$ instead of n for notational convenience, and omitting the factor $(n+1)^2$) into the three parts

$$v_1 \;=\; \mathbb{E}\Big[\overline{F}\big(M_n \vee (u - S_n)\big)^2;\; M_n > u/2\Big] \tag{5}$$

$$v_2 \;=\; \mathbb{E}\Big[\overline{F}\big(M_n \vee (u - S_n)\big)^2;\; M_n \le u/2, Y_{(n-1)} \le \epsilon u\Big] \tag{6}$$

$$v_3 \;=\; \mathbb{E}\Big[\overline{F}\big(M_n \vee (u - S_n)\big)^2;\; M_n \le u/2, Y_{(n-1)} > \epsilon u\Big] \tag{7}$$

where $\epsilon = 1/(2n)$. Here $v_1 \le \overline{F}(u/2)^2 \sim 2^\alpha \overline{F}(u)^2$. If $M_n \le u/2, Y_{(n-1)} \le \epsilon u$, we have $S_n \le u(1 - 1/2n)$ and hence

$$v_2 \;\le\; \overline{F}(u/2n)^2 \;\le\; \frac{L(u)^2}{u^{2\alpha}} \frac{L(u/2n)^2 (2n)^{2\alpha}}{L(u)^2} \;\sim\; \overline{F}(u)^2 (2n)^{2\alpha}$$

Finally,

$$\begin{aligned} v_3 \;&\le\; \mathbb{P}(Y_{(n-1)} > \epsilon u) \;\le\; n^2 \overline{F}(\epsilon u)^2 \;\le\; n^2 \frac{L(u)^2}{u^{2\alpha}} \frac{L(u/2n)^2 (2n)^{2\alpha}}{L(u)^2} \\ &\sim\; 4^\alpha n^{2+2\alpha} \overline{F}(u)^2. \end{aligned}$$

Putting these estimates together completes the proof. For the case of a random N, just condition upon $N = n + 1$ and invoke (4) (remember the omitted $(n+1)^2$ factor). □

Theorem 2.2 appears to give the first example of an estimator with bounded relative error in a heavy-tailed setting. However, asymptotic efficiency does not guarantee efficiency for a given set of parameters (a good

example of this is the importance sampling algorithm of [6]!). Nevertheless, complexity studies give a guideline to whether to proceed with an estimator or not. Encouraged by Theorems 2.1 and 2.2, we performed a comparison of different estimators for the Pareto case $\overline{F}(x) = (1+x)^{-\alpha}$ and the M/G/1 queue where N is geometric with $\mathbb{P}(N = n) = (1-\rho)\rho^n$, $n = 0, 1, 2, \ldots$ and $f(y) = \mathbb{P}(U > y)/\mathbb{E}U$ where U is a generic service time. Thus F plays the role of the integrated tail distribution so that the service time distribution itself is Pareto but with index $\alpha_U = \alpha + 1$ instead of α. The most important range for α_U is often argued to be the interval $(1, 2)$ (finite mean but infinite variance) so we took $\alpha_U = 3/2$, corresponding to $\alpha = 1/2$ and supplemented this with the larger value $\alpha = 3/2$ corresponding to $\alpha_U = 5/2$. Three different traffic intensities $\rho = 0.25, 0.5, 0.75$ were considered whereas for u we considered the four values which the standard approximation

$$\mathbb{P}(W > u) \;\sim\; \frac{\rho}{1-\rho}\overline{F}(u) \tag{8}$$

is 10^{-k}, with $k = 2, 5, 8, 11$. In the implementation we used

$$\mathbb{P}(W > u) \;=\; \mathbb{P}(Y_1 + \cdots + Y_N > u) \;=\; \rho\,\mathbb{P}(Y_1 + \cdots + Y_{N^*} > u),$$

where $\mathbb{P}(N^* = n) = (1-\rho)\rho^{n-1}$, $n = 1, 2, \ldots$, and simulated using N^*. The algorithms were replicated $R = 10^7$ times and Tables 1–2 give the corresponding relative error defined as the halfwidth of the 95% confidence interval divided by the point estimate.

In the list of algorithms, (1) and (3) are self-explanatory. CE means simple importance sampling (hazard rate twisting), simulating using $\alpha^* = n/\log u$ as suggested by the cross-entropy argument in [8]. JS is the weighted delayed hazard rate twisting of Juneja & Shahabuddin [16], implemented using the parameters suggested there. The asymmetric importance sampling (2) was implemented for two different f^*, such that $(2)_{\mathrm{CE}}$ means a Pareto f^* with $\alpha^* = 1/\log u$ as suggested by a cross-entropy argument similar to [8] (we omit the details of the derivation) and $(2)_{\mathrm{JS}}$ corresponds to the same f^* as was used in algorithm JS.

The point estimates (not given here) showed excellent agreement with the approximation (8) in the present Pareto case, but we note that regularly varying cases where (8) is quite inaccurate (and hence simulation is a realistic alternative) have been reported in, e.g., [1, 17].

The findings of Tables 1–2 are that the conditional Monte Carlo algorithms (1) and (3) performs better than any of the importance sampling algorithms, with (3) representing a substantial improvement of (1). The weighted delayed hazard rate twisting in algorithm JS is substantially more

ρ	k	(1)	CE	$(2)_{\mathrm{CE}}$	(3)	JS	$(2)_{\mathrm{JS}}$
0.25	2	0.071	0.221	0.152	0.032	0.185	0.224
	5	0.105	0.436	0.260	0.031	0.421	0.506
	8	0.122	0.783	0.335	0.031	0.582	0.703
	11	0.115	1.856	0.397	0.031	0.713	0.859
0.5	2	0.111	1.475	0.192	0.045	0.253	0.402
	5	0.144	1.442	0.301	0.044	0.515	0.812
	8	0.146	8.180	0.380	0.044	0.702	1.098
	11	0.153	3.808	0.445	0.044	0.855	1.337
0.75	2	0.141	3.136	0.232	0.054	0.314	0.744
	5	0.205	7.433	0.341	0.054	0.591	1.381
	8	0.188	7.128	0.423	0.054	0.795	1.888
	11	0.180	17.787	0.494	0.054	0.960	2.272

Table 1: Results for the Pareto case, $\alpha = 0.5$.

ρ	k	(1)	CE	$(2)_{\mathrm{CE}}$	(3)	JS	$(2)_{\mathrm{JS}}$
0.25	2	0.100	0.281	0.169	0.051	0.200	0.255
	5	0.150	0.435	0.260	0.031	0.420	0.518
	8	0.124	1.105	0.335	0.031	0.583	0.702
	11	0.102	1.311	0.396	0.031	0.715	0.861
0.5	2	0.161	1.954	0.234	0.077	0.282	0.492
	5	0.201	1.474	0.302	0.044	0.514	0.835
	8	0.152	2.562	0.381	0.044	0.703	1.113
	11	0.149	4.620	0.446	0.044	0.854	1.340
0.75	2	0.212	10.846	0.333	0.114	0.361	0.933
	5	0.201	6.025	0.342	0.054	0.589	1.469
	8	0.189	12.101	0.422	0.054	0.795	1.848
	11	0.231	26.358	0.492	0.054	0.961	2.298

Table 2: Results for the Pareto case, $\alpha = 1.5$.

efficient than the naive twisting of α in the CE algorithm, but when combined with asymmetric importance sampling, the CE algorithm is improved very substantially, the JS one not. An intuitive argument why this is the case may be that the role of the delay in the JS algorithm (represented by the change point x^*) is to ensure a sufficient number of Y_i that are not large even if $Y_1 + \cdots + Y_N$ is so.

3 The Weibull case

In this section, we assume that F is *Weibull-like* (with a terminology from [17]), meaning that the density $f(x)$ is asymptotically of the form $cx^{\gamma}\mathrm{e}^{-x^{\beta}}$

where $0 < \beta < 1$. The tail then satisfies $\overline{F}(x) \sim cx^{1+\gamma-\beta}\mathrm{e}^{-x^\beta}/\beta$.

We will only give a theoretical study of the efficiency properties of the estimator (3) and not of (2). This is motivated from the findings (both theoretical and empirical) in the Pareto case and the numerical studies below.

Theorem 3.1 *Assume* $\beta < \overline{\beta} = \log(3/2)/\log 2 = 0.585$, *i.e.* $2^{1+\beta} < 3$. *Then the estimator* (3) *is polynomial time for any fixed* n.

Remark 3.1 The occurence of a critical value of β is not uncommon for the Weibull distribution. See, e.g., [7] where some of the results take a different form for $1/2 < \beta < 1$, $1/3 < \beta < 1/2$ and so on. Critically of $\beta < 1/2$ for a certain result to hold true has been found in many later cases and is often referred to as *square-root insensitivity*, see e.g. [10] and references there. We have not seen the present critical value 0.585 show up before, but it is in fact maximal for our result, as can easily be seen by reverting the proof below.

Theorem 3.1 is obviously weaker than the corresponding Theorem 2.1 for the Pareto case which contains in addition bounded relative error and validity for the random N case. The numerical examples presented below strongly suggest that both of these extensions hold true for the Weibull case as well but we have no proof of this. □

In Fig. 1, the two areas shaded in different tones form together the support of the distribution of (M_n, S_n). Note that $u - y > x$ in the dark shaded area and that $x > u - y$ in the light shaded area. For the proof, we divide for each u the support into the $2n-1$ regions $0, 1', \ldots, (n-1)', 1'', \ldots, (n-1)''$, shown in Fig. 1 (e.g., k' is the triangle with border lines $u = x + y$, $y = kx$ and $y = (k+1)x$). We will denote by $\gamma_1, \gamma_2, \ldots$ certain powers of x or u whose particular values are unimportant (note that it suffices to bound the relative error by $u^{\gamma_k}\mathrm{e}^{-2u^\beta}$) and similarly $c_1, c_2, \ldots$ denote constants.

Lemma 3.1 *Let* $(x, y) \in k' \cup k''$ *for some* $k = 1, \ldots, n-1$. *Then the conditional density* $g(\cdot|x)$ *of* S_n *given* $M_n = Y_n = x$ *satisfies*

$$g(y|x) \;\le\; c_1 x^{\gamma_1} \exp\left\{-(k-1)x^\beta - (y-kx)^\beta\right\}, \quad y \ge x.$$

Proof. We have $f(x) \le c_2(1+x)^\gamma \mathrm{e}^{-x^\beta}\psi(x)$ where

$$c_3 = \sup_{y \ge a} \psi(y) < \infty, \quad \psi(y) \le c_4 f(y),\ y \le a, \tag{9}$$

for some $a > 0$. Letting S be the compact simplex $\subset \mathbb{R}^{n-1}$ specified by the constraints

$$0 \le x_1 \le x, \ \ldots, \ 0 \le x_{n-1} \le x, \quad x_1 + \cdots + x_{n-1} = y - x$$

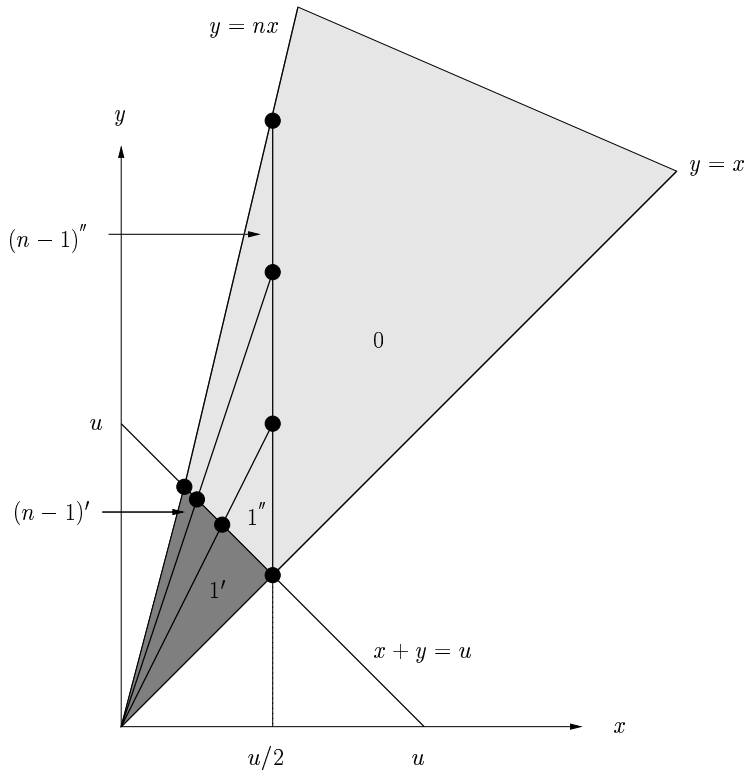

Figure 1: The support of the distribution of (M_n, S_m) is divided into $2n-1$ regions.

and letting μ be Lebesgue measure on S, we have

$$\begin{aligned} g(y|x) &= \int_S f(x_1)\dots f(x_{n-1})\,\mathrm{d}\mu(x_1,\dots,x_{n-1}) \\ &\le \left[c_2(1+x)^{\gamma_2}\right]^{n-1} \sup_S \exp\left\{-x_1^\beta - \dots - x_{n-1}^\beta\right\} \\ &\qquad \times \int_S \psi(x_1)\dots\psi(x_{n-1})\,\mathrm{d}\mu(x_1,\dots,x_{n-1}) \\ &\le c_5(1+x)^{\gamma_1} \sup_S \exp\left\{-x_1^\beta - \dots - x_{n-1}^\beta\right\}, \end{aligned}$$

where in the last step we used (9) to bound the μ-integral over an region of the form

$$S \cap \{x_1 \le a, \dots, x_k \le a, x_{k+1} > a, \dots, X_{n-1} > a\}$$

by $c_4^k[c_3(x-a)]^{n-k-1}$. Being concave, $x_1^\beta + \dots + x_{n-1}^\beta$ attains it minimum on S at an extremal point which is easily seen to mean that at a minimum

point $k-1$ of the x_i must equal x, one $y-kx$ and the rest 0. Thus the sup may be bounded by $\exp\{-(k-1)x^\beta-(y-kx)^\beta\}$. □

Proof of Theorem 3.1. Let $h(x,y)$ denote the joint density of M_n and S_n so that the second moment of the estimator is

$$\iint k(x,y)\,\mathrm{d}x\,\mathrm{d}y \quad \text{where} \quad k(x,y) \;=\; \exp\Big\{-2\big(x\vee(u-y)\big)^\beta\Big\}\,h(x,y).$$

Consider the partition in Fig. 1 of the support of h (the shaded area) into the regions $0, 1', \ldots, (n-1)', 1'', \ldots, (n-1)''$. We will carry out the proof by showing that the contributions $I_0, I_{1'}, \ldots$ to the integral from the separate regions grow no faster than $c_m u^{\gamma_m}\mathrm{e}^{-2u^\beta}$. This is simple for region 0, whereas for the remaining ones we will use similar ideas as in the proof of Lemma 3.1 based upon concavity and extremal points (the corner points marked by bullets in Fig. 1; note that the volumes grow at rate at most u^2).
0: In region 0, $x > u-y$ and $x > u/2$. So, recalling the standard fact that the density of M_n is

$$n\,f(x)F(x)^{n-1} \le nf(x), \tag{10}$$

we get

$$I_0 \;\le\; \mathrm{e}^{-2(u/2)^\beta}\mathbb{P}(M_n \ge u/2) \;\le\; c_6 u^{\gamma_3}\mathrm{e}^{-3(u/2)^\beta} \;\le\; c_6 u^{\gamma_3}\mathrm{e}^{-2u^\beta}.$$

$\mathbf{k}'$: Here $u-y \ge x$ and by Lemma 3.1,

$$h(x,y) \;\le\; c_7\,x^{\gamma_4}\exp\left\{-kx^\beta-(y-kx)^\beta\right\}.$$

Thus (note that $x \le u$ when $x \in k'$),

$$I_{k'} \;\le\; c_7 u^{\gamma_4}\iint_{k'}\exp\left\{-kx^\beta-(y-kx)^\beta-2(u-y)^\beta\right\}\,\mathrm{d}x\,\mathrm{d}y.$$

Since the area of k' grows like u^2, it suffices to check that the value of minus the exponent is at least $2u^\beta$ at each of the three boundary points. This is clear for $(0,0)$. The two other boundary points (at the line $x+y=u$) are $\big(u/(k+1), ku/(k+1)\big)$ and $\big(u/(k+2), (k+1)u/(k+2)\big)$ where minus the exponent is $(k+2)u^\beta/(k+1)^\beta$, resp. $(k+3)u^\beta/(k+2)^\beta$ which both are $> 2u^\beta$ (the function $(x+2)/(x+1)^\beta$ is $3/2^\beta > 2$ at $x=1$ and increasing, as is easily seen by checking that the log derivative is positive for $\beta < \overline{\beta}$, $x \ge 1$).

$\mathbf{k}''$: Here $u-y \le x$ so that as in the $\mathbf{k}'$ argument

$$I_{k''} \;\le\; c_8 u^{\gamma_5}\iint_{k''} nx^n\exp\left\{-(k+2)x^\beta-(y-kx)^\beta\right\}\,\mathrm{d}x\,\mathrm{d}y$$

We must again check that the value of minus the exponent is at least $2u^\beta$ at each of the four boundary points. The two boundary points at the line $u = x+y$ and the values to be checked are the same as for k' so these behave as should be. The two on the line $x = u/2$ are $\big(u/2, (k+1)u/2\big)$ and $\big(u/2, ku/2\big)$ where minus the exponent is at least $(k+2)x^\beta = (k+2)u^\beta/2^\beta \geq 2u^\beta$. $\square$

Tables 3–5 contain a similar geometric sum numerical study as for the Pareto case, considering a (standard) Weibull F with tail e^{-x^β} and three different values 0.25, 0.50, 0.75 of β (we omitted (1) since the algorithm is not polynomial time in the Weibull case). The choice of F and $\beta = 0.5$ is as in [16] and as check, we also reconstructed Table 2 of [16] and obtained very similar point estimates and confidence bands (in fact, algorithm JS came out slightly better than in [16] which is probably due to the N^* instead of N issue). We do not include the table here. Note, however, that e^{-x^β} is not the tail of an integrated tail distribution (the derivative is not monotone) and therefore there is no direct M/G/1 interpretation.

For $\beta = 0.25$ and $\beta = 0.50$ the conclusions of Tables 3–5 are very much as for the Pareto case, whereas for $\beta = 0.75$ all algorithms show considerable performance degradation. This is maybe not surprising since the algorithms are specifically designed for heavy tails and at $\beta = 0.75$ we start approaching the border $\beta = 1$ to light tails.

ρ	k	CE	$(2)_{\mathrm{CE}}$	(3)	JS	$(2)_{\mathrm{JS}}$
0.25	2	0.163	0.154	0.035	0.186	0.228
	5	0.496	0.261	0.032	0.419	0.508
	8	1.278	0.335	0.031	0.583	0.702
	11	3.015	0.396	0.031	0.716	0.860
0.5	2	0.260	0.200	0.052	0.258	0.420
	5	1.719	0.303	0.045	0.513	0.821
	8	2.729	0.380	0.044	0.702	1.099
	11	13.187	0.446	0.044	0.854	1.336
0.75	2	1.961	0.256	0.071	0.331	0.823
	5	15.284	0.347	0.056	0.588	1.406
	8	57.216	0.424	0.054	0.795	1.867
	11	22.254	0.494	0.054	0.963	2.238

Table 3: Results for the Weibull case, $\beta = 0.25$.

ρ	k	CE	$(2)_{\mathrm{CE}}$	(3)	JS	$(2)_{\mathrm{JS}}$
0.25	2	0.158	0.171	0.054	0.201	0.255
	5	0.759	0.334	0.072	0.432	0.677
	8	1.015	0.378	0.056	0.590	0.859
	11	2.128	0.414	0.070	0.719	0.886
0.5	2	0.323	0.261	0.098	0.295	0.515
	5	2.694	0.614	0.185	0.650	1.415
	8	3.793	0.465	0.097	0.733	1.939
	11	8.194	0.502	0.101	0.876	1.486
0.75	2	2.311	0.388	0.153	0.315	0.863
	5	6.866	7.128	0.665	3.820	8.015
	8	15.580	1.319	0.597	1.101	3.077
	11	18.771	0.683	0.118	1.068	10.251

Table 4: Results for the Weibull case, $\beta = 0.5$.

ρ	k	CE	$(2)_{\mathrm{CE}}$	(3)	JS	$(2)_{\mathrm{JS}}$
0.25	2	0.159	0.198	0.082	0.222	0.290
	5	1.862	1.580	0.476	1.208	2.534
	8	3.499	6.673	1.296	2.652	7.588
	11	7.182	2.951	4.532	1.927	6.643
0.5	2	0.421	0.304	0.135	0.268	0.487
	5	127.183	2.815	1.1097	3.689	7.396
	8	30.289	10.451	8.043	19.517	25.075
	11	12.959	7.396	13.190	13.483	61.513
0.75	2	6.257	0.329	0.141	0.189	0.516
	5	14.562	2.088	0.713	1.326	4.067
	8	76.658	13.584	5.239	12.860	37.925
	11	92.182	86.575	43.958	70.867	193.427

Table 5: Results for the Weibull case, $\beta = 0.75$.

4 Combination with stratification and control variates

The subexponential asymptotics

$$\mathbb{P}(S_n > u) \ \sim \ n\overline{F}(u), \tag{11}$$

which is valid for a fixed n, indicates that a substantial part of the variability of the estimators in the random N tables may be due to the variability in a random N. Two ways to eliminate this are to use N as control variate (as suggested by the linearity of (11) in n) or to stratify N. Note that

both methods guarantee variance reduction. We refer, e.g., to [14] for the basic facts on these variance reduction techniques. The following Tables 6-10 contain numerical studies pertaining to this issue, where we use the same parameter values as in Tables 1-5. For the stratification (using proportional allocation, cf. [14]), we took 8 strata $N^* = 1, \ldots, 7$, $N^* > 7$ for $\rho = 0.25$ and 17 strata $N^* = 1, \ldots, 16$, $N^* > 16$ for $\rho = 0.50$ and $\rho = 0.75$. Again, the tables give the half-width of the confidence intervals and $\mathrm{JS}_{\mathrm{CV}}$, $\mathrm{JS}_{\mathrm{Str}}$ means the JS algorithm combined with control variates, resp. stratification, and similarly for the other algorithms. Of course, 0.000 just means $< 5 \cdot 10^{-4}$.

ρ	k	$(2)_{\mathrm{CV}}$	$(2)_{\mathrm{Str}}$	$(3)_{\mathrm{CV}}$	$(3)_{\mathrm{Str}}$	$\mathrm{JS}_{\mathrm{CV}}$	$\mathrm{JS}_{\mathrm{Str}}$
0.25	2	0.150	0.121	0.008	0.008	0.132	0.181
	5	0.258	0.210	0.000	0.000	0.303	0.419
	8	0.335	0.272	0.000	0.000	0.420	0.583
	11	0.396	0.323	0.000	0.000	0.513	0.713
0.5	2	0.192	0.179	0.009	0.009	0.222	0.248
	5	0.301	0.285	0.000	0.000	0.458	0.512
	8	0.381	0.362	0.000	0.000	0.625	0.700
	11	0.445	0.424	0.000	0.000	0.761	0.855
0.75	2	0.232	0.225	0.009	0.011	0.301	0.308
	5	0.341	0.334	0.000	0.005	0.572	0.588
	8	0.423	0.419	0.000	0.005	0.773	0.793
	11	0.491	0.487	0.000	0.005	0.938	0.961

Table 6: Variance reduction results for the Pareto case, $\alpha = 0.5$.

ρ	k	$(2)_{\mathrm{CV}}$	$(2)_{\mathrm{Str}}$	$(3)_{\mathrm{CV}}$	$(3)_{\mathrm{Str}}$	$\mathrm{JS}_{\mathrm{CV}}$	$\mathrm{JS}_{\mathrm{Str}}$
0.25	2	0.169	0.143	0.025	0.024	0.159	0.193
	5	0.259	0.211	0.001	0.001	0.302	0.418
	8	0.334	0.272	0.000	0.000	0.419	0.582
	11	0.396	0.323	0.000	0.000	0.514	0.712
0.5	2	0.234	0.220	0.043	0.038	0.259	0.273
	5	0.302	0.286	0.001	0.001	0.457	0.512
	8	0.380	0.362	0.000	0.000	0.626	0.702
	11	0.445	0.425	0.000	0.000	0.760	0.851
0.75	2	0.329	0.315	0.074	0.069	0.344	0.348
	5	0.342	0.336	0.002	0.006	0.573	0.587
	8	0.423	0.418	0.000	0.005	0.772	0.797
	11	0.492	0.486	0.000	0.005	0.937	0.963

Table 7: Variance reduction results for the Pareto case, $\alpha = 1.5$.

ρ	k	$(2)_{\mathrm{CV}}$	$(2)_{\mathrm{Str}}$	$(3)_{\mathrm{CV}}$	$(3)_{\mathrm{Str}}$	$\mathrm{JS}_{\mathrm{CV}}$	$\mathrm{JS}_{\mathrm{Str}}$
0.25	2	0.153	0.124	0.011	0.011	0.135	0.182
	5	0.260	0.212	0.005	0.005	0.302	0.418
	8	0.335	0.273	0.002	0.001	0.419	0.582
	11	0.396	0.322	0.000	0.000	0.511	0.714
0.5	2	0.200	0.188	0.018	0.018	0.230	0.253
	5	0.303	0.289	0.007	0.007	0.457	0.511
	8	0.381	0.361	0.002	0.002	0.625	0.698
	11	0.446	0.424	0.001	0.001	0.762	0.853
0.75	2	0.257	0.248	0.031	0.032	0.318	0.323
	5	0.347	0.341	0.011	0.012	0.571	0.584
	8	0.424	0.419	0.003	0.006	0.771	0.790
	11	0.494	0.488	0.001	0.006	0.939	0.964

Table 8: Variance reduction results for the Weibull case, $\beta = 0.25$.

ρ	k	$(2)_{\mathrm{CV}}$	$(2)_{\mathrm{Str}}$	$(3)_{\mathrm{CV}}$	$(3)_{\mathrm{Str}}$	$\mathrm{JS}_{\mathrm{CV}}$	$\mathrm{JS}_{\mathrm{Str}}$
0.25	2	0.170	0.146	0.029	0.028	0.162	0.195
	5	0.344	0.318	0.054	0.057	0.339	0.431
	8	0.412	0.326	0.063	0.046	0.446	0.587
	11	0.411	0.350	0.020	0.025	0.534	0.718
0.5	2	0.261	0.247	0.062	0.058	0.272	0.283
	5	0.789	0.565	0.140	0.140	0.620	0.607
	8	0.635	0.585	0.090	0.126	0.678	0.733
	11	0.503	0.486	0.044	0.032	0.802	0.874
0.75	2	0.373	0.367	0.109	0.101	0.290	0.295
	5	2.266	2.269	0.638	0.788	2.846	2.241
	8	1.063	0.948	0.241	0.259	1.063	1.101
	11	0.736	0.662	0.074	0.077	1.059	1.072

Table 9: Variance reduction results for the Weibull case, $\beta = 0.5$.

ρ	k	$(2)_{\mathrm{CV}}$	$(2)_{\mathrm{Str}}$	$(3)_{\mathrm{CV}}$	$(3)_{\mathrm{Str}}$	$\mathrm{JS}_{\mathrm{CV}}$	$\mathrm{JS}_{\mathrm{Str}}$
0.25	2	0.197	0.176	0.049	0.047	0.190	0.209
	5	1.492	1.469	0.471	0.418	1.864	2.263
	8	4.205	4.374	1.726	1.581	1.893	3.403
	11	2.303	1.778	2.099	1.864	3.474	1.725
0.5	2	0.290	0.283	0.091	0.085	0.240	0.247
	5	3.050	3.249	1.105	0.956	3.619	3.578
	8	14.864	27.744	11.819	15.355	17.372	34.213
	11	17.251	9.784	11.914	20.099	28.343	10.382
0.75	2	0.266	0.311	0.101	0.096	0.156	0.159
	5	2.020	2.009	0.659	0.594	1.268	1.306
	8	15.617	16.337	4.926	4.950	13.290	12.996
	11	46.885	81.367	32.731	41.419	174.065	74.014

Table 10: Variance reduction results for the Weibull case, $\beta = 0.75$.

The conclusions to be drawn is that the control variate method and stratification perform rather much the same. The variance reduction is by far the largest for algorithm (3) which we take as indication that the algorithm estimates $\mathbb{P}(S_n > u)$ very accurately for a fixed n and that the variability of the M/G/1 estimators is largely due to the variability in N. In contrast, for the other estimators the variability in the estimates of $\mathbb{P}(Y_1 + \cdots + Y_n > u)$ is non-negligible compared to the variation in N.

References

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

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

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

[4] S. Asmussen (2003) *Applied Probability and Queues* (2nd ed.). 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, C. Klüppelberg & K. Sigman (1998) Sampling at subexponential times, with queueing applications. *Stoch. Proc. Appl.* **79**, 265–286.

[8] S. Asmussen, D.P. Kroese & R.Y. Rubinstein (2004) Heavy tails, importance sampling and cross-entropy. *Stochastic Models* (pending revision).

[9] 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.

[10] A. Baltrunas, D.J. Daley & C. Klüppelberg (2002) Tail behaviour of the busy period in the GI/G/1 queue with subexponential service times. Submitted.

[11] 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 2002 Winter Simulation Conference, San Diego*, pp. 468–476.

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

[13] W. Feller (1971) *An Introduction to Probability Theory and Its Applications* **II** (2nd ed.). Wiley.

[14] P. Glasserman (2004) *Monte Carlo Methods for Financial Engineering*. Springer-Verlag.

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

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

[17] T. Mikosch & A.V. Nagaev (1998). Large deviations of heavy-tailed sums with applications in insurance. *Extremes*, **1** (1), 81–110.

[18] R.Y. Rubinstein & D.P. Kroese (2004) *The Cross-Entropy Method. A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer-Verlag, New York. To appear.

[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.