

(13 August 2018)

SEMIPARAMETRIC CROSS ENTROPY FOR RARE-EVENT SIMULATIONZ. I. BOTEV,^{*} *The University of New South Wales*A. RIDDER,^{**} *Vrije Universiteit*L. ROJAS-NANDAYAPA,^{***} *The University of Queensland***Abstract**

The Cross Entropy method is a well-known adaptive importance sampling method for rare-event probability estimation, which requires estimating an optimal importance sampling density within a parametric class. In this article we estimate an optimal importance sampling density within a wider semiparametric class of distributions. We show that this semiparametric version of the Cross Entropy method frequently yields efficient estimators. We illustrate the excellent practical performance of the method with numerical experiments and show that for the problems we consider it typically outperforms alternative schemes by orders of magnitude.

Keywords: light-tailed; regularly-varying; subexponential; rare-event probability; Cross Entropy method, Markov chain Monte Carlo

2010 Mathematics Subject Classification: Primary 65C05

Secondary 65C60;65C40

1. Introduction

In this article we consider the problem of estimating rare-event probabilities of the form

$$\ell = \mathbb{P}(S(\mathbf{X}) > \gamma), \quad \mathbf{X} = (X_1, \dots, X_d),$$

where $S(\mathbf{x}) = x_1 + \dots + x_d$ and X_1, \dots, X_d are (possibly dependent) random variables. We call these the jump variables. Such estimation problems arise in various contexts, see, for example, [1, 3, 9]. We describe an adaptive importance sampling algorithm, which can be viewed as

^{*} Postal address: School of Mathematics and Statistics, University of New South Wales, Sydney, NSW 2052 Australia

^{**} Postal address: School of Mathematics and Physics, The University of Queensland, Brisbane, QLD 4072, Australia

^{***} Postal address: Department Econometrics and Operations Research, Vrije Universiteit, 1081 HV, Amsterdam

the semiparametric version of the well-known Cross Entropy (CE) method for estimation of rare-event probabilities [15]. The main ingredients of the semiparametric CE method are as follows.

First, similar to [5, 6] we use a Markov Chain Monte Carlo (MCMC) algorithm to obtain random variables distributed according to the minimum variance importance sampling density. In our context the minimum variance importance sampling density is simply the density of the vector \mathbf{X} conditioned on the rare event $S(\mathbf{X}) > \gamma$. Second, with the MCMC sample at hand, we construct a conditional (or a Rao-Blackwell) estimator of each of the marginal densities of the minimum variance importance sampling density. Finally, we use the product of these (estimated) marginal densities as our importance sampling density in order to estimate ℓ . Under idealized conditions that ignore the error arising from the MCMC sampling, we show that the resulting estimator achieves either logarithmic or bounded relative error efficiencies. The strength of the method is not only that it outperforms the currently recommended estimation procedures for heavy-tailed probabilities, but that the exact same procedure is efficient in problems with light-tailed probabilities. For example, we show that unlike any existing procedures, the method is efficient in the Weibull case for all values of the tail index α , even in the light-tailed case with $\alpha > 1$.

Numerical experiments show that, despite the heuristic nature of the MCMC step, the estimator can in practice be frequently more reliable and efficient than tailor-made importance sampling schemes. In other words, an advantage of the methodology advocated here is that a single broadly-applicable heuristic algorithm provides satisfactory practical performance on a range of different estimation problems (both in light- and heavy-tailed cases) and frequently this performance is superior to estimation schemes that are specifically designed to a particular rare-event estimation problem.

The rest of the paper is organized as follows. In Section 2 we quickly review the parametric CE method and introduce its semiparametric version. This is followed by a number of examples with details about the practical implementation of the estimator. The examples aims to demonstrate the superior performance of the proposed algorithm compared to existing estimation algorithms on a number of prototypical examples. In Section 4 we provide theoretical analysis of the efficiency of a simple version of the estimator for light- and heavy- tailed random variables. Finally, Section 5 gives some concluding remarks.

2. Cross Entropy method

2.1. Parametric Cross Entropy method

In order to introduce the semiparametric version of the CE method, we briefly review the CE method itself. Let $f(\mathbf{x})$ be the joint density of the vector $\mathbf{X} = (X_1, \dots, X_d)$ and suppose that it is part of the parametric family

$$\mathcal{F} = \left\{ f(\cdot; \mathbf{v}) : \mathbb{R}^d \rightarrow \mathbb{R}_{>0} : \int f(\mathbf{x}; \mathbf{v}) d\mathbf{x} = 1; \mathbf{v} \in \mathcal{V} \right\}, \quad (1)$$

where $\mathcal{V} \subset \mathbb{R}^p$ is the feasible parameter set. The assumption is that $f(\mathbf{x}) \equiv f(\mathbf{x}; \mathbf{u}) \in \mathcal{F}$ for some $\mathbf{u} \in \mathcal{V}$. Then, the objective is to find a parameter $\mathbf{v} \in \mathcal{V}$ that yields a good importance sampling estimator of the form:

$$\widehat{\ell}_{\text{CE}} = \frac{1}{m} \sum_{i=1}^m \mathbb{I}\{S(\mathbf{Y}_i) > \gamma\} \frac{f(\mathbf{Y}_i; \mathbf{u})}{f(\mathbf{Y}_i; \mathbf{v})}, \quad \mathbf{Y}_1, \dots, \mathbf{Y}_m \stackrel{\text{iid}}{\sim} f(\mathbf{y}; \mathbf{v}). \quad (2)$$

In the CE method the best parameter $\mathbf{v}^* \in \mathcal{V}$ is the one which minimizes the cross entropy distance between $f(\cdot; \mathbf{v}) \in \mathcal{F}$ and the zero-variance importance sampling density

$$\pi(\mathbf{x}) = \frac{\mathbb{I}\{S(\mathbf{x}) > \gamma\} f(\mathbf{x})}{\mathbb{P}(S(\mathbf{X}) > \gamma)}.$$

In other words,

$$\mathbf{v}^* = \operatorname{argmin}_{\mathbf{v} \in \mathcal{V}} \int \pi(\mathbf{x}) \ln \left(\frac{\pi(\mathbf{x})}{f(\mathbf{x}; \mathbf{v})} \right) d\mathbf{x} = \operatorname{argmax}_{\mathbf{v} \in \mathcal{V}} \int \pi(\mathbf{x}) \ln f(\mathbf{x}; \mathbf{v}) d\mathbf{x}. \quad (3)$$

In practice the integral $\int \pi(\mathbf{x}) \ln \left(\frac{\pi(\mathbf{x})}{f(\mathbf{x}; \mathbf{v})} \right) d\mathbf{x}$ is estimated from a preliminary simulation so that we obtain the estimator of \mathbf{v}^* :

$$\widehat{\mathbf{v}}^* = \operatorname{argmax}_{\mathbf{v} \in \mathcal{V}} \sum_{i=1}^n \ln f(\mathbf{X}_i, \mathbf{v}), \quad (4)$$

where $\mathbf{X}_1, \dots, \mathbf{X}_n$ is an approximate sample from π obtained via Markov chain Monte Carlo (MCMC) sampling over the restricted set \mathcal{S}_γ , see [7] and Remark 1 below. In this way we use MCMC to learn about the optimal (in cross entropy sense) parameter \mathbf{v}^* . In many applications the parametric density $f(\cdot; \mathbf{v})$ is of product form: $f(\mathbf{x}; \mathbf{v}) = \prod_{i=1}^d f_i(x_i; v_i)$. For the special case where each $f_i(x_i; v_i)$ belongs to a one-parameter exponential family parametrized by the mean [18, Pages 69-70], the solution of (4) is given by the maximum-likelihood estimator of the mean vector:

$$\widehat{v}_i^* = \frac{1}{n} \sum_{j=1}^n X_{j,i}, \quad i = 1, \dots, d,$$

where $X_{j,i}$ is the i -th coordinate of the j -th sample \mathbf{X}_j . We thus use the importance sampling estimator (2) with $\mathbf{v} = \widehat{\mathbf{v}}^*$.

Remark 1. (*Generating $\mathbf{X}_1, \dots, \mathbf{X}_n$ via Gibbs sampling.*) In our discussion we assume that the conditional densities $\pi(x_i | \mathbf{x}_{-i})$ are available in closed form. We can thus use the following Gibbs sampling procedure to obtain $\mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{\text{approx}}{\sim} \pi$.

Algorithm 2.1. (Gibbs Sampler.)

Require: An initial state $\mathbf{X}_0 \sim f(\mathbf{x})$ and sample size n .

for $t = 0, \dots, n - 1$ **do**

 Set $\mathbf{Y} = \mathbf{X}_t$.

for $i = 1, \dots, d$ **do**

 Draw $Y_i \sim \pi(y_i | Y_1, \dots, Y_{i-1}, X_{t,i+1}, \dots, X_{t,d})$.

 Set $\mathbf{X}_{t+1} = \mathbf{Y}$.

2.2. Semiparametric Importance sampling

Recall that the original CE method aims to find the best importance sampling density $f(\cdot; \mathbf{v}^*) \in \mathcal{F}$ within the parametric family (1); namely by solving the parametric optimization program (3). In contrast, in the semiparametric CE method the objective is to find the optimal importance sampling density amongst a family of densities given by some common property. Again, the optimality criterion is to minimize the cross-entropy distance from the zero-variance density. Denote by \mathcal{G}_1 the set of all single-variate probability density functions; that is, $g(x) : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ is absolute continuous with $\int g(x) dx = 1$. Let \mathcal{G} be the family of product-form densities on \mathbb{R}^d :

$$\mathcal{G} = \left\{ g(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0} : g(\mathbf{x}) = \prod_{i=1}^d g_i(x_i); g_i \in \mathcal{G}_1, i = 1, \dots, d \right\}.$$

In this paper we consider \mathcal{G} as the target set of importance sampling densities. Hence, the objective is to solve the functional optimization program $\min_{g \in \mathcal{G}} \int \pi(\mathbf{x}) \ln \left(\frac{\pi(\mathbf{x})}{g(\mathbf{x})} \right) d\mathbf{x}$. This is equivalent to

$$g(\mathbf{x}) = \operatorname{argmin}_{g_1, \dots, g_d \in \mathcal{G}_1} \int \pi(\mathbf{x}) \ln \left(\frac{\pi(\mathbf{x})}{\prod_{i=1}^d g_i(x_i)} \right) d\mathbf{x} = \operatorname{argmax}_{g_1, \dots, g_d \in \mathcal{G}_1} \int \pi(\mathbf{x}) \ln \left(\prod_{i=1}^d g_i(x_i) \right) d\mathbf{x}. \quad (5)$$

Lemma 1. *Let $\pi_i(x_i)$ be the i -th marginal of the zero-variance density $\pi(\mathbf{x})$. Then the solution to the semiparametric CE program (5) is $g_i = \pi_i$ for all $i = 1, \dots, d$. In other words, the*

optimal importance sampling density within the space of all product-form densities is the one given by the product of the marginals of $\pi(\mathbf{x})$.

The proof is given in the Appendix. In practice the marginal densities of π are not available (just like the exact \mathbf{v}^* in (3) is not available) and need to be estimated from simulation. Here we use the estimators

$$\widehat{\pi}_i(y_i) = \frac{1}{n} \sum_{k=1}^n \pi(y_i | \mathbf{X}_{k,-i}), \quad i = 1, \dots, d, \quad (6)$$

where

- $\mathbf{X}_1, \dots, \mathbf{X}_n$ is an approximate sample from π obtained via Gibbs sampling as in (4) (see also Remark 1);
- the vector $\mathbf{X}_{k,-i}$ is the same as \mathbf{X}_k except that the i -th component is removed;
- $\pi(x_i | \mathbf{X}_{k,-i})$ is the conditional density of x_i given all the other components of \mathbf{X}_k .

The estimator (6) is motivated by the simple identity:

$$\begin{aligned} \mathbb{E}_\pi[\widehat{\pi}_i(y)] &= \frac{1}{n} \sum_{k=1}^n \mathbb{E}_\pi[\pi(y | \mathbf{X}_{k,-i})] = \mathbb{E}_\pi[\pi(y | \mathbf{X}_{-i})] \\ &= \mathbb{E}_\pi[\pi(y | X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_d)] \\ &= \int \pi(y | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d) \pi(\mathbf{x}) \, d\mathbf{x} \\ &= \int \frac{\pi(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_d)}{\pi(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d)} \pi(\mathbf{x}) \, d\mathbf{x} \\ &= \int \frac{\pi(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_d)}{\pi(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d)} \, d\mathbf{x}_{-i} \times \int^{\overbrace{\pi(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d)}} \pi(x_1, \dots, x_d) \, dx_i \\ &= \int \pi(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_d) \, d\mathbf{x}_{-i} = \pi_i(y). \end{aligned}$$

We define the approximation to the optimal semiparametric CE solution by the product of marginal density estimators (6), that is,

$$\widehat{g}(\mathbf{y}) \stackrel{\text{def}}{=} \prod_{i=1}^d \widehat{\pi}_i(y_i). \quad (7)$$

Then we estimate ℓ by the importance sampling estimator

$$\widehat{\ell} = \frac{1}{m} \sum_{i=1}^m \mathbb{I}\{S(\mathbf{Y}_i) > \gamma\} \frac{f(\mathbf{Y}_i)}{\widehat{g}(\mathbf{Y}_i)}, \quad \mathbf{Y}_1, \dots, \mathbf{Y}_m \stackrel{\text{iid}}{\sim} \widehat{g}(\mathbf{y}), \quad (8)$$

Note that, conditional on $\mathbf{X}_1, \dots, \mathbf{X}_n$, each $\widehat{\pi}_i$ is an equally weighted mixture of n densities (with k -th component $\pi(y_i | \mathbf{X}_{k,-i})$) and hence sampling $Y_i \sim \widehat{\pi}_i(y_i)$ can be performed using the composition method [16][Page 53]. In other words, choose a component of the mixture at random by generating K uniformly from the set of integers $\{1, \dots, n\}$. Then, given $K = k$, sample Y_i from the k -th mixture component $Y \sim \pi(y_i | \mathbf{X}_{k,-i})$. Finally, deliver Y_i as a realization from $\widehat{\pi}(y_i)$ and (Y_1, \dots, Y_d) as a realization from $\widehat{g}(\mathbf{y})$.

Remark 2. (*Using exact conditional density.*) Note that once we have sampled Y_1, \dots, Y_{d-1} from $\widehat{\pi}_1, \dots, \widehat{\pi}_{d-1}$, respectively, we have the option of sampling the final Y_d from the exact conditional $\pi(y_d | Y_1, \dots, Y_{d-1})$, instead of from the d -th marginal $\widehat{\pi}_d$. This reduces the cross entropy distance to π even further and yields the alternative and typically more efficient estimator (8) with $\widehat{g}(\mathbf{y})$ redefined as

$$\widehat{g}(\mathbf{y}) \leftarrow \widehat{\pi}_1(y_1) \times \dots \times \widehat{\pi}_{d-1}(y_{d-1}) \times \pi(y_d | y_1, \dots, y_{d-1}).$$

3. Examples and Practical Implementation

In this section we consider the prototypical problem of estimating $\mathbb{P}(X_1 + \dots + X_d > \gamma)$, where the jumps X_1, X_2, \dots may or may not be dependent. In the case of independent jumps, the proposed importance sampling can yield practical performance surpassing that of well established alternative estimation procedures such as the Asmussen-Kroese (AK) estimator [2, 4]. This is in part due to the fact that our estimator incorporates the ingenious exchangeability and conditioning proposed in [2]. First, recall that the AK estimator in [2] based on one replication is given by

$$\widehat{\ell}_{\text{AK}} = d \overline{F}\left(\left(\gamma - \sum_{j=1}^{d-1} X_j\right) \vee \max_{j < d} X_j\right), \quad X_1, \dots, X_{d-1} \stackrel{\text{iid}}{\sim} F.$$

The motivation for the estimator is the identity $\ell = d \mathbb{P}(X_1 + \dots + X_d > \gamma, X_d = M_d) = d \mathbb{E} \overline{F}\left(\left(\gamma - \sum_{j=1}^{d-1} X_j\right) \vee \max_{j < d} X_j\right)$, where $x \vee y = \max\{x, y\}$ and $M_d \stackrel{\text{def}}{=} \max_{j \leq d} X_j$. This conditional estimator enjoys excellent practical performance for the problems we consider below. For further details we refer to [4, 13], where the authors prove that the estimator is a vanishing relative error one.

We obtain an estimator that outperforms $\widehat{\ell}_{\text{AK}}$ in terms of (estimated) relative time variance

by exploiting the decomposition proposed in [14] and the ex

$$\begin{aligned}
\ell &= \mathbb{P}(M_d > \gamma) + \mathbb{P}(S(\mathbf{X}) > \gamma, M_d < \gamma) \\
&= \mathbb{P}(M_d > \gamma) + d \mathbb{P}(S(\mathbf{X}) > \gamma, X_d = M_d < \gamma), \quad \text{by exchangeability of jumps} \\
&= 1 - \mathbb{P}(M_d < \gamma) + d \mathbb{P}(X_d = M_d < \gamma) \mathbb{P}(S(\mathbf{X}) > \gamma | X_d = M_d < \gamma) \\
&= \underbrace{1 - [F(\gamma)]^d}_{\text{dominant term}} + \underbrace{\mathbb{P}(M_d < \gamma) \widetilde{\mathbb{P}}(S(\mathbf{X}) > \gamma)}_{\text{residual probability}},
\end{aligned}$$

where the new probability measure $\widetilde{\mathbb{P}}(\cdot) = \mathbb{P}(\cdot | X_d = M_d < \gamma)$ with corresponding density

$$\widetilde{f}(\mathbf{x}) = f(\mathbf{x} | X_d = M_d < \gamma) = \frac{d f(\mathbf{x})}{[F(\gamma)]^d} \mathbb{I}\{M_d < \gamma, X_d = M_d\}.$$

Estimating the residual probability, we obtain the one replication estimator for ℓ as

$$\widehat{\ell} = 1 - [F(\gamma)]^d + \frac{\widetilde{f}(\mathbf{Y})}{\widetilde{g}(\mathbf{Y})} \mathbb{I}\{S(\mathbf{Y}) > \gamma\}, \quad \mathbf{Y} \sim \widetilde{g}(\mathbf{y}), \quad (9)$$

where $\widetilde{g}(\mathbf{y}) \stackrel{\text{def}}{=} \widehat{\pi}_1(y_1) \cdots \widehat{\pi}_{d-1}(y_{d-1}) \pi(y_d | y_1, \dots, y_{d-1})$ is the estimated importance sampling pdf described in Remark 2.

In the following examples we used the relative time variance product (RTVP) and the ratio of relative errors as a measure of efficiency:

$$\text{Ratio} \stackrel{\text{def}}{=} \frac{\widehat{\sigma}_{\text{AK}} / \widehat{\ell}_{\text{AK}}}{\widehat{\sigma} / \widehat{\ell}}, \quad \text{RTVP} \stackrel{\text{def}}{=} \text{Ratio}^2 \times \frac{\tau_{\text{AK}}}{\tau},$$

where $\widehat{\sigma}_{\text{AK}}$ and $\widehat{\sigma}$ are the sample standard deviations of $\widehat{\ell}_{\text{AK}}$ and $\widehat{\ell}$ (all based on m replications), respectively, and τ_{AK} and τ are the CPU times taken to compute the respective estimators. The quantity τ includes the CPU time needed for the preliminary MCMC simulations.

Example 1. (*Weibull case.*) Here we wish to estimate $\mathbb{P}(X_1 + \cdots + X_d > \gamma)$ and assume that each of the jumps X_i has density $\alpha x^{\alpha-1} e^{-x^\alpha}$ for $x > 0$ and $0 < \alpha < 1$. Hence, $\overline{F}(x) = e^{-x^\alpha}$. In comprehensive simulations studies the proposed estimator outperformed the Asmussen-Kroese (AK) estimator in terms of relative time variance for all values of the parameters α and γ . The improvement, however, was not uniform, see Table 1, where, for example for $\alpha = 0.1$, we can see savings from as little as 71 times to as large as approximately 6000. The general trend is for large gains for smaller γ and $\alpha > 0.6$ or $\alpha < 0.3$. The AK estimator was strongest in the range $\alpha \in [0.3, 0.6]$ with values for $\alpha \notin [0.3, 0.6]$ rendering it less efficient compared to (9).

Note that the AK estimator is much faster to evaluate than (9), but this speed is insufficient to offset the substantial gains in squared relative error (given by Ratio column).

TABLE 1: Comparison of importance sampling method with the AK estimator. Algorithmic parameters were chosen to be $n = 10^3, m = 10^6, d = 10$. The AK estimator is based on $m = 10^6$ replications.

| $\alpha = 0.1$ | | | | | $\alpha = 0.2$ | | | | |
|----------------|------------------|---------------|---------|------|----------------|------------------|------------|---------|------|
| γ | $\widehat{\ell}$ | Rel. Err. | Ratio | RTVP | γ | $\widehat{\ell}$ | Rel. Err. | Ratio | RTVP |
| 10^{10} | $4.54/10^4$ | $1.7/10^6$ | 13^2 | 71 | 10^4 | $1.97/10^2$ | $6.5/10^5$ | 3^2 | 3.7 |
| 10^{11} | $3.40/10^5$ | $4.1/10^7$ | 22^2 | 197 | 10^5 | $4.64/10^4$ | $1.8/10^5$ | 5.6^2 | 12 |
| 10^{12} | $1.30/10^6$ | $6.4/10^8$ | 72^2 | 2071 | 10^6 | $1.31/10^6$ | $3/10^6$ | 9.2^2 | 33 |
| 10^{13} | $2.16/10^8$ | $8/10^9$ | 59^2 | 1429 | 10^7 | $1.23/10^{10}$ | $4.3/10^7$ | 10^2 | 42 |
| 10^{15} | $1.84/10^{13}$ | $1.3/10^{10}$ | 125^2 | 5944 | 10^8 | $5.13/10^{17}$ | $6.5/10^8$ | 7^2 | 20 |

| $\alpha = 0.6$ | | | | | $\alpha = 0.9$ | | | | |
|----------------|------------------|------------|---------|------|----------------|------------------|-----------|---------|--------|
| γ | $\widehat{\ell}$ | Rel. Err. | Ratio | RTVP | γ | $\widehat{\ell}$ | Rel. Err. | Ratio | RTVP |
| 10^2 | $9.47/10^6$ | $2.6/10^4$ | 19^2 | 130 | 30 | $1.33/10^4$ | $9/10^4$ | 13^2 | 50 |
| 150 | $7.83/10^8$ | $1.5/10^4$ | 41^2 | 550 | 40 | $6.27/10^7$ | $9/10^4$ | 78^2 | 1758.7 |
| 200 | $1.34/10^9$ | $1.5/10^4$ | 63^2 | 1376 | 50 | $2.25/10^9$ | $1/10^3$ | 254^2 | 17746 |
| 500 | $1.83/10^{17}$ | $1.7/10^4$ | 5.5^2 | 11 | 60 | $7.01/10^{12}$ | $1/10^3$ | 556^2 | 87103 |
| 10^3 | $7.00/10^{27}$ | $9.5/10^5$ | 6^2 | 13 | 100 | $4.34/10^{22}$ | $1/10^3$ | 300^2 | 23768 |

Remark 3. (*Efficient evaluation of \widehat{g} .*) If we define, $c_k \stackrel{\text{def}}{=} (\gamma - \sum_{j \neq i} \mathbf{X}_{k,j})^+$, then (6) simplifies to

$$\widehat{\pi}_i(y_i) = \frac{1}{n} \sum_{k=1}^n \pi(y_i | \mathbf{X}_{k,-i}) = \frac{1}{n} \alpha y_i^{\alpha-1} e^{-y_i^\alpha} \sum_{k=1}^n \mathbb{I}\{y_i \geq c_k\} / e^{-c_k} = f(x_i) \frac{1}{n} \sum_{k=1}^n \mathbb{I}\{y_i \geq c_{(k)}\} \times e^{c_{(k)}},$$

where the term $\sum_{k=1}^n \mathbb{I}\{y_i \geq c_{(k)}\} \times e^{c_{(k)}}$ can be evaluated for an arbitrary y_i quickly by first computing and storing in memory the cumulative sums $\sum_{k=1}^i e^{c_{(k)}}$, $i = 1, \dots, n$ and then using table look-up methods with $\mathcal{O}(n)$ time complexity.

Example 2. (*Pareto case.*) Assume that the jumps X_i have Pareto density and distribution functions given by $f(x) = \alpha/x^{\alpha+1}$, $F(x) = 1 - 1/x^\alpha$, $x \geq 1$. The following table shows the results of a comparison with the AK estimator for different values of α and γ . Again, the efficiency gains with the proposed method can be of the order of 10^4 .

TABLE 2: Comparison of importance sampling with the AK estimator for Pareto case. Here $n = 10^3$, $m = 10^6$, $d = 10$.

| $\alpha = 0.5$ | | | | | $\alpha = 1$ | | | | |
|----------------|------------------|---------------|---------|-------|---------------|------------------|----------------|---------|------|
| $\gamma - d$ | $\widehat{\ell}$ | Rel. Err. | Ratio | RTVP | $\gamma - d$ | $\widehat{\ell}$ | Rel. Err. | Ratio | RTVP |
| 10^8 | $1.00/10^3$ | $5.6/10^7$ | 33^2 | 209 | 10^4 | $1.00/10^3$ | $5.1/10^6$ | 7^2 | 11 |
| 10^{10} | $1.00/10^4$ | $5.8/10^8$ | 107^2 | 3007 | 10^6 | $1.00/10^5$ | $1.0/10^7$ | 38^2 | 330 |
| 10^{11} | $3.16/10^5$ | $1.8/10^8$ | 176^2 | 6270 | 10^8 | $1.00/10^7$ | $1.4/10^9$ | 91^2 | 1711 |
| 10^{12} | $9.99/10^6$ | $5.92/10^9$ | 364^2 | 34271 | 10^{10} | $1.00/10^9$ | $2.61/10^{11}$ | 42^2 | 322 |
| 10^{15} | $3.16/10^7$ | $1.9/10^{10}$ | 584^2 | 82494 | 10^{13} | $1.00/10^{12}$ | $3/10^{14}$ | 24^2 | 123 |
| $\alpha = 5$ | | | | | $\alpha = 10$ | | | | |
| $\gamma - d$ | $\widehat{\ell}$ | Rel. Err. | Ratio | RTVP | $\gamma - d$ | $\widehat{\ell}$ | Rel. Err. | Ratio | RTVP |
| 10^1 | $2.58/10^4$ | $1.5/10^4$ | 10^2 | 66 | 5 | $1.75/10^6$ | $2.4/10^4$ | 30^2 | 609 |
| 10^2 | $1.06/10^9$ | $1.2/10^5$ | 4^2 | 11 | 10 | $1.09/10^9$ | $9.93/10^5$ | 6^2 | 22 |
| 10^3 | $1.00/10^{14}$ | $1.13/10^6$ | 4^2 | 11 | 10^2 | $1.00/10^{19}$ | $8.8/10^6$ | 4^2 | 13 |
| 10^4 | $1.00/10^{19}$ | $1/10^7$ | 4.4^2 | 11 | 500 | $1.02/10^{26}$ | $1.6/10^6$ | 5^2 | 11 |
| 10^5 | $1.00/10^{24}$ | $1.2/10^8$ | 4^2 | 11 | 1500 | $1.73/10^{31}$ | $5.5/10^7$ | 4.4^2 | 13 |

Example 3. (*Compound Sum.*) We are interested in estimating the tail probability of a compound sum of the form $\mathbb{P}(X_1 + \dots + X_R > \gamma)$, where the jumps X_i are iid with Weibull distribution with parameter $0 < \alpha < 1$, and (without loss of generality) $R \sim \text{Geom}(\varrho)$ is a geometric random variable with pdf $\varrho(1-\varrho)^{r-1}$, $r = 1, 2, \dots$. We have $\mathbb{P}(S_R > \gamma) = \mathbb{P}(X_1 + \dots + X_R > \gamma) =$

$$\begin{aligned} \varrho \sum_{r=1}^{\infty} (1-\varrho)^{r-1} \mathbb{P}(S_r > \gamma) &= \varrho \sum_{r=1}^{\infty} (1-\varrho)^{r-1} \mathbb{P}(M_r > \gamma) + \varrho \sum_{r=2}^{\infty} (1-\varrho)^{r-1} \mathbb{P}(M_r < \gamma, S_r > \gamma) \\ &= \underbrace{\frac{\overline{F}(\gamma)}{\overline{F}(\gamma) + \varrho F(\gamma)}}_{\text{dominant term}} + \frac{\varrho(1-\varrho)(F(\gamma))^2}{\overline{F}(\gamma) + \varrho F(\gamma)} \underbrace{\widetilde{\mathbb{P}}(S_R > \gamma)}_{\text{residual probability}}, \end{aligned}$$

where under the new probability measure $\widetilde{\mathbb{P}}$ we have $(R-1) \sim \text{Geom}(\overline{F}(\gamma) + \varrho F(\gamma))$ with pdf $\widetilde{\mathbb{P}}(R=r) = f_R(r)$, $r = 2, 3, \dots$ and $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} f(x)$ with pdf given by the truncated Weibull density $f(x) = \alpha x^{\alpha-1} e^{-x^\alpha} / (1 - e^{-\gamma^\alpha})$, $0 < x < \gamma$. Hence, we can again apply our importance sampling estimator to estimate the residual probability $\widetilde{\mathbb{P}}(S_R > \gamma)$. The minimum variance pdf for the estimation of the residual is

$$\pi(\mathbf{y}, r) \propto f_R(r) \prod_{j=1}^r f(y_j) \mathbb{I}\{S_r > \gamma\},$$

which can be easily sampled from using the Gibbs sampler in Algorithm 2.1 by noting that

$$\pi(r|\mathbf{Y}) \propto f_R(r) \mathbb{I}\{r \geq r^*(\mathbf{Y})\}, \quad r^*(\mathbf{Y}) \stackrel{\text{def}}{=} \min\{r : Y_1 + \dots + Y_r > \gamma\}.$$

Table 3 gives the results of a number of numerical experiments. The results of our proposed method are significantly better in all cases, except $\alpha = 0.2$ with $1/\varrho \in \{50, 100\}$. In the latter case, the variance reduction achieved by the proposed method is not sufficient to offset the computational cost of simulating compound sums of expected length of $1/\varrho$. Note that for $\alpha \geq 0.5$, the proposed method can be thousands of times more efficient. Our proposed method is also more efficient than the recently proposed improved Asmussen-Kroese estimator [12][Table 2]. For example, based on the reported variances and computing time in [12], in terms of RTVP our estimator is from 8.5 to 45 times more efficient. We must note, however, that the results given in Table 2 of [12] appear to be incorrect. For example, for $\varrho = 0.15$, $\alpha = 0.75$, $\gamma = 63.361$ Table 2 reports the estimate 5.23×10^{-4} with relative error of 0.4%. In contrast, we obtained the estimate 5.38×10^{-4} with relative error 0.03%, which we verified with a Crude Monte Carlo simulation using 10^9 repetitions.

TABLE 3: Compound Weibull sum with expected number of jumps $1/\varrho$. Here $n = 10^4$, $m = 10^6$.

| $\alpha = 0.2$ with $\gamma = 10^6$ fixed | | | | | $\alpha = 0.5$ with $\gamma = 500$ fixed | | | | |
|--|------------------|------------|----------|--------|---|------------------|------------|----------|----------|
| $1/\varrho$ | $\widehat{\ell}$ | Rel. Err. | Ratio | RTVP | $1/\varrho$ | $\widehat{\ell}$ | Rel. Err. | Ratio | RTVP |
| 5 | $6.56/10^7$ | $1.4/10^5$ | 3.6^2 | 9.6 | 3 | $7.34/10^{10}$ | $7.3/10^4$ | 4^2 | 16 |
| 10 | $1.31/10^6$ | $3.1/10^5$ | 2.8^2 | 3.5 | 5 | $1.60/10^9$ | $1/10^3$ | 4.1^2 | 12 |
| 20 | $2.65/10^6$ | $5.1/10^5$ | 2.2^2 | 1.2 | 10 | $1.17/10^8$ | $1.7/10^3$ | 47^2 | 445 |
| 50 | $6.81/10^6$ | $1.7/10^4$ | 1.4^2 | 0.03 | 20 | $1.24/10^5$ | $7.2/10^4$ | 246^2 | 7300 |
| 100 | $1.42/10^5$ | $1.7/10^4$ | 2^2 | 0.04 | 50 | $7.9/10^3$ | $2.1/10^4$ | 58^2 | 110 |
| $\alpha = 0.8$ with $\gamma = 30/\varrho$ depending on ϱ | | | | | $\alpha = 0.95$ with $\gamma = 30/\varrho$ depending on ϱ | | | | |
| $1/\varrho$ | $\widehat{\ell}$ | Rel. Err. | Ratio | RTVP | $1/\varrho$ | $\widehat{\ell}$ | Rel. Err. | Ratio | RTVP |
| 3 | $6.29/10^{11}$ | $1.2/10^3$ | 330^2 | 46000 | 5 | $2.61/10^{13}$ | $4.8/10^4$ | 10^6 | $> 10^5$ |
| 5 | $1.65/10^{11}$ | $6.4/10^4$ | 930^2 | 200000 | 10 | $2.18/10^{13}$ | $3/10^4$ | $> 10^6$ | $> 10^5$ |
| 10 | $6.94/10^{12}$ | $3.8/10^4$ | 2561^2 | 780000 | 20 | $2.00/10^{13}$ | $2.2/10^4$ | $> 10^6$ | 40000 |
| 20 | $4.64/10^{12}$ | $2.7/10^4$ | 3636^2 | 34000 | 50 | $1.91/10^{13}$ | $1.9/10^4$ | $> 10^6$ | $> 10^5$ |
| 50 | $3.68/10^{12}$ | $2.1/10^4$ | 1485^2 | 27000 | 100 | $1.88/10^{13}$ | $1.7/10^4$ | $> 10^6$ | $> 10^5$ |

4. Robustness Properties of Semiparametric Cross Entropy Estimator

In this section we study the robustness properties of the estimator (8) when $\gamma \rightarrow \infty$ in some simplified prototypical settings. Clearly, then $\ell = \ell(\gamma) = \mathbb{P}(S(\mathbf{X}) > \gamma) \rightarrow 0$. We are interested in the behavior of the standard error of the estimator in this regime, specifically, relative to its mean ℓ . Since we take a finite constant sample size, it suffices to analyze the robustness of the single-run estimator of ℓ :

$$Z = Z(\gamma) = \mathbb{I}\{S(\mathbf{X}) > \gamma\} \frac{f(\mathbf{X})}{g(\mathbf{X})}, \quad (10)$$

where $\mathbf{X} \sim g(\mathbf{x}) = \prod_{i=1}^d g_i(x_i) = \prod_{i=1}^d \pi_i(x_i)$. For our analysis we assume that the importance sampling density g is available. In practice we estimate g via \widehat{g} from MCMC simulation as we discussed in Section 2.2. In this respect, our analysis is similar in spirit to the one conducted for the parametric Cross Entropy method [8]. The estimator has bounded relative error if $\limsup_{\gamma \rightarrow \infty} \sqrt{\text{Var}(Z)}/\ell < \infty$, which is equivalent to having bounded relative second moment [17]:

$$\limsup_{\gamma \rightarrow \infty} \frac{\mathbb{E}Z^2}{\ell^2} < \infty.$$

Assumption 1. *In this section we assume that the jump variables X_1, \dots, X_d are positive continuous, and that they are independent and identically distributed random variables with right-unbounded support.*

We denote by $F(x)$ the cdf of a jump X_i with associated pdf $f_1(x)$. Let $\overline{F}(x) = 1 - F(x)$ be the tail cdf, F^{*d} be the d -fold convolution of F , with $\overline{F}^{*d} = 1 - F^{*d}$. Note that the rare-event probability of interest is $\ell = \mathbb{P}(X_1 + \dots + X_d > \gamma) = \overline{F}^{*d}(\gamma)$. Furthermore, the i -th marginal π_i of the zero-variance pdf can be rewritten as

$$\begin{aligned} \pi_i(x_i) &= \int_{\mathbb{R}_{>0}^{d-1}} \pi(\mathbf{x}) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_d \\ &= \int_{\mathbb{R}_{>0}^{d-1}} \frac{\mathbb{I}\{S(\mathbf{x}) > \gamma\} f(\mathbf{x})}{\ell} dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_d \\ &= \int_{\mathbb{R}_{>0}^{d-1}} \frac{\mathbb{I}\{S(\mathbf{x}) > \gamma\} \prod_{j=1}^d f_1(x_j)}{\ell} dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_d \\ &= \frac{f_1(x_i)}{\ell} \int_{\mathbb{R}_{>0}^{d-1}} \mathbb{I}\{x_1 + \dots + x_d > \gamma\} \prod_{j \neq i} f_1(x_j) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_d \\ &= \frac{f_1(x_i)}{\ell} \mathbb{P}(X_1 + \dots + X_{i-1} + X_{i+1} + \dots + X_d > \gamma - x_i) = \frac{f_1(x_i) \overline{F}^{*(d-1)}(\gamma - x_i)}{\overline{F}^{*d}(\gamma)}. \end{aligned}$$

Note that for $x_i > \gamma$ we clearly have $\overline{F^{*(d-1)}}(\gamma - x_i) = 1$, and thus $\pi_i(x_i) = f_1(x_i)/\ell$. Hence, the single-run estimator Z can be written as

$$Z = \mathbb{I}\{S(\mathbf{X}) > \gamma\} \frac{f(\mathbf{X})}{g(\mathbf{X})} = \mathbb{I}\{S(\mathbf{X}) > \gamma\} \prod_{i=1}^d \frac{f_1(X_i)}{\pi_i(X_i)} = \mathbb{I}\{S(\mathbf{X}) > \gamma\} \prod_{i=1}^d \frac{\overline{F^{*d}}(\gamma)}{\overline{F^{*(d-1)}}(\gamma - X_i)} \quad (11)$$

Finally, using $\mathbb{E}Z^2 = \mathbb{E}_g Z^2 = \mathbb{E}_g Z f(\mathbf{X})/g(\mathbf{X}) = \mathbb{E}_f Z$, we get for the second moment of estimator Z :

$$\mathbb{E}Z^2 = \mathbb{E}_f \mathbb{I}\{S(\mathbf{X}) > \gamma\} \prod_{i=1}^d \frac{\overline{F^{*d}}(\gamma)}{\overline{F^{*(d-1)}}(\gamma - X_i)}. \quad (12)$$

Proposition 1. *Suppose that the jumps X_1, \dots, X_d are i.i.d. with a light-tailed or a subexponential Weibull or Pareto distribution. Then, the semiparametric importance sampling estimator (10) is at least logarithmically efficient as $\gamma \rightarrow \infty$.*

In the subsequent sections we prove this result by considering the heavy- and light-tailed cases separately.

4.1. Heavy-tailed case

In this section we assume that all jumps X_i are drawn from a subexponential distribution F satisfying (for all integer d)

$$\lim_{\gamma \uparrow \infty} \frac{\overline{F^{*d}}(\gamma)}{\overline{F}(\gamma)} = d. \quad (13)$$

In the sequel we shall frequently use the trivial property

$$\overline{F^{*d}}(x) \geq \overline{F}(x), \quad x \geq 0. \quad (14)$$

Furthermore, we shall need Kesten's bound Lemma 1.3.5(c) in [9], which states that for every $\varepsilon > 0$ there exists a constant c_1 such that for all $d \geq 2$

$$\overline{F^{*d}}(x) \leq c_1(1 + \varepsilon)^d \overline{F}(x), \quad x \geq 0. \quad (15)$$

Denoting the maximum $M_d = \max_{i \leq d} X_i$, we can decompose the relative second moment as follows:

$$\frac{\mathbb{E}Z^2}{\ell^2} = \frac{\mathbb{E}\mathbb{I}\{M_d > \gamma\} Z^2}{\ell^2} + \frac{\mathbb{E}\mathbb{I}\{M_d \leq \gamma\} Z^2}{\ell^2}. \quad (16)$$

In Lemma 2 we shall prove that the first term is bounded as $\gamma \rightarrow \infty$. Concerning the second term, we examine its behavior for various common probability models in the next two sections.

Lemma 2.

$$\limsup_{\gamma \rightarrow \infty} \frac{\mathbb{E}\mathbb{I}\{M_d > \gamma\} Z^2}{\ell^2} < \infty.$$

Proof. Since $\mathbb{I}\{S(\mathbf{x}) > \gamma\} \leq 1$, we use (12) to find

$$\mathbb{E}\mathbb{I}\{M_d > \gamma\} Z^2 \leq \mathbb{E}_f \mathbb{I}\{M_d > \gamma\} \prod_{i=1}^d \frac{\overline{F^{*d}}(\gamma)}{F^{*(d-1)}(\gamma - X_i)} \quad (17)$$

Then observe that, if $M_d > \gamma$, there exists at least one jump $X_j > \gamma$, and, hence, that there is at least one j for which $\overline{F^{*(d-1)}}(\gamma - X_j) = 1$. For all other jumps it holds trivially $\overline{F^{*(d-1)}}(\gamma - X_i) \geq \overline{F^{*(d-1)}}(\gamma)$, thus it follows that (17) is bounded from above by

$$\begin{aligned} \frac{\mathbb{E}_f \mathbb{I}\{M_d > \gamma\} \prod_{i=1}^d \overline{F^{*d}}(\gamma)}{\prod_{i \neq j}^d \overline{F^{*(d-1)}}(\gamma - X_i)} &\leq \mathbb{E}_f \mathbb{I}\{M_d > \gamma\} \frac{\prod_{i=1}^d \overline{F^{*d}}(\gamma)}{\prod_{i \neq j}^d \overline{F^{*(d-1)}}(\gamma)} \\ &= \mathbb{P}_f(M_d > \gamma) \frac{(\overline{F^{*d}}(\gamma))^d}{(\overline{F^{*(d-1)}}(\gamma))^{d-1}} \\ &\leq \frac{(\overline{F^{*d}}(\gamma))^{d+1}}{(\overline{F^{*(d-1)}}(\gamma))^{d-1}} = (\overline{F^{*d}}(\gamma))^2 \left(\frac{\overline{F^{*d}}(\gamma)}{\overline{F^{*(d-1)}}(\gamma)} \right)^{d-1}, \end{aligned}$$

where the last inequality follows from $\mathbb{P}_f(M_d > \gamma) \leq \mathbb{P}_f(S(\mathbf{X}) > \gamma) = \overline{F^{*d}}(\gamma)$. Now we use the bounds (14) and (15) for

$$\left(\frac{\overline{F^{*d}}(\gamma)}{\overline{F^{*(d-1)}}(\gamma)} \right)^{d-1} \leq \left(\frac{\overline{F^{*d}}(\gamma)}{\overline{F}(\gamma)} \right)^{d-1} \leq c_1^{d-1} (1 + \varepsilon)^{d(d-1)}.$$

Collecting all bounds we obtain

$$\begin{aligned} \frac{\mathbb{E}\mathbb{I}\{M_d > \gamma\} Z^2}{\ell^2} &= \frac{1}{(\overline{F^{*d}}(\gamma))^2} \mathbb{E}_f \mathbb{I}\{M_d > \gamma\} \prod_{i=1}^d \frac{\overline{F^{*d}}(\gamma)}{F^{*(d-1)}(\gamma - X_i)} \\ &\leq c_1^{d-1} (1 + \varepsilon)^{d(d-1)} < \infty. \end{aligned} \quad (18)$$

□

Since we have bounded relative error for the first term in (16), then we can at most have bounded relative error for estimator (10). For example, if the second term in (16) vanishes or is bounded, then (10) has bounded relative error.

4.1.1. *Weibull distribution* As in Example 1, here we assume that each of the jumps X_i have density $\alpha x^{\alpha-1} e^{-x^\alpha}$ for $0 < \alpha < 1$. The purpose is to analyze the second term in (16).

Lemma 3.

$$\limsup_{\gamma \rightarrow \infty} \frac{\mathbb{E}\mathbb{I}\{M_d < \gamma\} Z^2}{\ell^2} = 0.$$

Proof. Denote $S_d = S(\mathbf{X})$. Using (12) and $\ell = \overline{F^{*d}}(\gamma)$, we get

$$\frac{\mathbb{E}\mathbb{I}\{M_d < \gamma\} Z^2}{\ell^2} = \mathbb{E}_f \mathbb{I}\{M_d < \gamma, S_d > \gamma\} \frac{\prod_{i=1}^{d-2} \overline{F^{*d}}(\gamma)}{\prod_{i=1}^d \overline{F^{*(d-1)}}(\gamma - X_i)}.$$

From the bounds (14) and (15), we obtain that this expression can be bounded above by

$$\begin{aligned} & \mathbb{E}_f \mathbb{I}\{M_d < \gamma, S_d > \gamma\} \frac{\prod_{i=1}^{d-2} c_1 (1 + \varepsilon)^d \overline{F}(\gamma)}{\prod_{i=1}^d \overline{F}(\gamma - X_i)} \\ &= c_2 \mathbb{E}_f \mathbb{I}\{M_d < \gamma, S_d > \gamma\} \exp\left(- (d-2)\gamma^\alpha + \sum_{i=1}^d (\gamma - X_i)^\alpha\right). \end{aligned}$$

We now consider the following integral over the region $\{\mathbf{x} : 0 < x_i < \gamma, \sum_i x_i > \gamma\}$:

$$\begin{aligned} & \mathbb{E}_f \mathbb{I}\{M_d < \gamma, S_d > \gamma\} \exp\left(- (d-2)\gamma^\alpha + \sum_{i=1}^d (\gamma - X_i)^\alpha\right) \\ &= \alpha^d \int \dots \int \left(\prod_{i=1}^d x_i^{\alpha-1} \right) \exp\left(- (d-2)\gamma^\alpha + \sum_{i=1}^d ((\gamma - x_i)^\alpha - x_i^\alpha)\right) d\mathbf{x} \end{aligned}$$

After the change of variable $u_i = x_i/\gamma$ for all i we obtain that this integral is a Laplace-type integral:

$$\alpha^d \gamma^{d\alpha} \underbrace{\int \dots \int_{\mathcal{D}} h(\mathbf{u}) e^{-\gamma^\alpha \phi(\mathbf{u})} d\mathbf{u}}_{\text{Laplace-type}},$$

where:

$$\begin{aligned} \mathcal{D} &\stackrel{\text{def}}{=} \left\{ \mathbf{u} : 0 < u_i < 1, \sum_i u_i > 1 \right\} \\ h(\mathbf{u}) &\stackrel{\text{def}}{=} \prod_{i=1}^d u_i^{\alpha-1} \\ \phi(\mathbf{u}) &\stackrel{\text{def}}{=} d - 2 + \sum_{i=1}^d (u_i^\alpha - (1 - u_i)^\alpha) \end{aligned}$$

We now note the following properties of the Laplace integral. First, if $\bar{\mathcal{D}}$ denotes the closure of the open set \mathcal{D} , the function $\phi(\mathbf{u})$ attains its unique global minimum within the bounded domain $\bar{\mathcal{D}} \subseteq \mathbb{R}^d$ on the boundary at $\mathbf{u}^* = (1/d, \dots, 1/d)$. This can be seen either by applying the Lagrange constraint optimization method or more simply by noting that $u^\alpha - (1 - u)^\alpha$ is

monotonically increasing and $\phi(\mathbf{u})$ is invariant to permutations of the components of \mathbf{u} . The minimum

$$\phi(\mathbf{u}^*) = d - 2 + d^{1-\alpha} - d^{1-\alpha}(d-1)^\alpha,$$

as a function of d is such that for $d > 2$ we have the strict inequality $\phi(\mathbf{u}) \geq \phi(\mathbf{u}^*) > 0$ for all $\mathbf{u} \in \tilde{\mathcal{D}}$, see Figure 1. The point \mathbf{u}^* is not a critical point, because $\frac{\partial \phi}{\partial u_i}(\mathbf{u}) = \alpha(u_i^{\alpha-1} + (1-u_i)^{\alpha-1}) > 0$ for all i and $\mathbf{u} \in \mathcal{D}$.

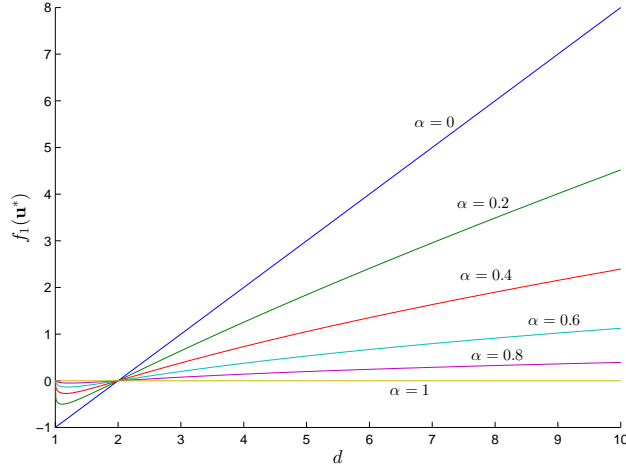


FIGURE 1: The behavior of the function $d - 2 + d^{1-\alpha} - d^{1-\alpha}(d-1)^\alpha$ for different values of the parameter α .

Second, the function $h : \mathbb{R}^d \rightarrow \mathbb{R}$ is continuous and the Hessian of the surface $p(u_1, \dots, u_{d-1}) = \phi(u_1, u_2, \dots, u_{d-1}, 1 - u_1 - u_2 - \dots - u_{d-1})$ is

$$\frac{\partial^2 p}{\partial u_i \partial u_j} = \alpha(\alpha - 1) \times \begin{cases} (1 - \sum_{k < d} u_k)^{\alpha-2} - (\sum_{k < d} u_k)^{\alpha-2} & i \neq j \\ u_i^{\alpha-2} - (1 - u_i)^{\alpha-2} + (1 - \sum_{k < d} u_k)^{\alpha-2} - (\sum_{k < d} u_k)^{\alpha-2} & i = j \end{cases},$$

which when evaluated at \mathbf{u}^* yields the nondegenerate Hessian matrix

$$\alpha(\alpha - 1) \left(d^{2-\alpha} - (1 - 1/d)^{\alpha-2} \right) \times \begin{pmatrix} 2 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 1 & \cdots & 1 \\ 1 & 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 2 \end{pmatrix}.$$

As a result of all these conditions we have the Laplace-type asymptotic expansion [20, Page

500] at a boundary point, which is not a critical point:

$$\int \dots \int_{\mathcal{D}} h(\mathbf{u}) e^{-\gamma^\alpha \phi(\mathbf{u})} d\mathbf{u} = \mathcal{O}(\gamma^{-\alpha(d+1)/2} \times e^{-\gamma^\alpha \phi(\mathbf{u}^*)}),$$

where the constant $\phi(\mathbf{u}^*) > 0$. It follows that

$$\begin{aligned} \frac{\mathbb{E}\mathbb{I}\{M_d < \gamma\} Z^2}{\ell^2} &\leq c_2 \alpha^d \gamma^{d\alpha} \int \dots \int_{\mathcal{D}} h(\mathbf{u}) e^{-\gamma^\alpha \phi(\mathbf{u})} d\mathbf{u} \\ &= \mathcal{O}(\gamma^{\alpha(d-1)/2} \times e^{-\gamma^\alpha \phi(\mathbf{u}^*)}) = \mathcal{O}(e^{\alpha(d-1)/2 \ln \gamma - \gamma^\alpha \phi(\mathbf{u}^*)}) \rightarrow 0 \quad (\gamma \rightarrow \infty). \end{aligned}$$

Hence the second term in (16) vanishes as $\gamma \rightarrow \infty$. \square

4.1.2. *Sum of Pareto random variables.* As in Example 2, we assume that X_i 's are independent and Pareto distributed random variables on $[1, \infty)$ with common parameter $\alpha > 0$. The main result is the logarithmic efficiency of the second term of (16).

Proposition 2. *For all $\varepsilon > 0$*

$$\limsup_{\gamma \rightarrow \infty} \frac{\mathbb{E}[Z^2; M_d \leq \gamma]}{\ell^{2-\varepsilon}} = 0.$$

Proof. The proof will be the result of a number of lemmas. First, similarly as in Lemma 3 we utilize expression (12) for rewriting the second moment as a product, and then we apply (14) and (15) to bound the factors. The result is that it is enough to prove that

$$\limsup_{\gamma \rightarrow \infty} \frac{1}{\ell^{2-\varepsilon}} \mathbb{E}_f \mathbb{I}\{M_d \leq \gamma, S_d > \gamma\} \prod_{i=1}^d \frac{\bar{F}(\gamma)}{\bar{F}(\gamma - X_i)} = 0. \quad (19)$$

Our approach is to consider a larger set containing $\{M_d \leq \gamma, S_d > \gamma\}$. For that purpose we define the we define the quantities

$$H_n(\gamma) := \mathbb{E}_f \left[\prod_{k=1}^n \frac{\bar{F}(\gamma)}{\bar{F}(\gamma - X_k)}; B_n \right], \quad n \geq 2.$$

where

$$B_n = \{S_{n-1} \leq \gamma, S_n > \gamma, M_n \leq \gamma\}, \quad n = 2, 3, \dots$$

Observe that $\{M_d \leq \gamma, S_d > \gamma\} \subset \bigcup_{n=2}^d B_n$. Further to this, observing that $\bar{F}(\gamma)/\bar{F}(\gamma - x) \leq 1$ for all $x \geq 1$, we can set

$$\prod_{k=1}^d \frac{\bar{F}(\gamma)}{\bar{F}(\gamma - x_k)} \leq \prod_{k=1}^n \frac{\bar{F}(\gamma)}{\bar{F}(\gamma - x_k)}, \quad n \leq d.$$

In this way we arrive at the following inequality

$$\begin{aligned}
& \mathbb{E}_f \mathbb{I}\{M_d \leq \gamma, S_d > \gamma\} \prod_{i=1}^d \frac{\bar{F}(\gamma)}{\bar{F}(\gamma - X_i)} \\
& \leq \sum_{n=2}^d \mathbb{E}_f \left[\prod_{i=1}^d \frac{\bar{F}(\gamma)}{\bar{F}(\gamma - X_i)} ; B_n \right] \\
& \leq \sum_{n=2}^d \mathbb{E}_f \left[\prod_{i=1}^n \frac{\bar{F}(\gamma)}{\bar{F}(\gamma - X_i)} ; B_n \right] = \sum_{n=2}^d H_n(\gamma).
\end{aligned} \tag{20}$$

Now, the quantities H_n in the sum above can be written in integral form as

$$\begin{aligned}
H_n(\gamma) &= \int_{B_n} \left(\prod_{k=1}^n \frac{\bar{F}(\gamma)}{\bar{F}(\gamma - x_k)} \right) \left(\prod_{k=1}^n f(x_k) \right) dx_n dx_{n-1} \dots dx_2 dx_1 \\
&= \int_1^{\gamma-(n-2)} \int_1^{\gamma-x_1-(n-3)} \dots \int_1^{\gamma-x_1-\dots-x_{n-2}} \int_{(y-x_1-\dots-x_{n-1}) \vee 1}^{\gamma} \prod_{k=1}^n \left(\frac{\gamma - x_k}{\gamma} \right)^\alpha \frac{\alpha}{x_k^{\alpha+1}} dx_n dx_{n-1} \dots dx_2 dx_1.
\end{aligned}$$

Further, the change of variable $y_k = x_k/\gamma$ yields

$$\frac{\alpha^n}{\gamma^{n\alpha}} \int_{\gamma^{-1}}^{1-(n-2)\gamma^{-1}} \int_{\gamma^{-1}}^{1-y_1-(n-3)\gamma^{-1}} \dots \int_{\gamma^{-1}}^{1-y_1-\dots-y_{n-2}} \int_{(1-y_1-\dots-y_{n-1}) \vee \gamma^{-1}}^1 \prod_{k=1}^n L(y_k) dy_n dy_{n-1} \dots dy_2 dy_1,$$

where

$$L(y) := (1-y)^\alpha y^{-(\alpha+1)}, \quad y \in (0, 1]. \tag{21}$$

In particular, it will be useful to write

$$H_n(\gamma) = \alpha^n \gamma^{-n\alpha} I_n(\gamma, 1), \tag{22}$$

where the function $I_n(\gamma, 1)$ is the multiple integral in the expression above. Moreover, $I_n(\gamma, \zeta)$ can be defined recursively for via

$$I_n(\gamma, \zeta) := \begin{cases} \int_{\zeta \vee \gamma^{-1}}^1 L(y) dy, & n = 1, \\ \int_{\gamma^{-1}}^{\zeta^{-(n-2)\gamma^{-1}}} L(y) I_{n-1}(\gamma, \zeta - y) dy, & n \geq 2. \end{cases} \tag{23}$$

Next we will prove that for $n = 2, 3, \dots$, it holds that

$$\limsup_{\gamma \rightarrow \infty} \frac{I_n(\gamma, 1)}{\gamma^{\alpha(n-2)} \ln \gamma} = 0. \tag{24}$$

Since both numerator and denominator of (24) have limit $+\infty$, we can apply L'Hopital. Lemma 5 in the appendix provides a recursive expression for the derivative of the functions $I_n(\gamma, \zeta)$:

$$\frac{\partial}{\partial \gamma} I_n(\gamma, \zeta) = nL(\gamma^{-1}) I_{n-1}(\gamma, \zeta - \gamma^{-1}) \gamma^{-2}, \quad n = 2, 3, \dots \tag{25}$$

Therefore, we obtain

$$\begin{aligned} \limsup_{\gamma \rightarrow \infty} \frac{I_n(\gamma, 1)}{\gamma^{\alpha(n-2)} \ln \gamma} &= \limsup_{\gamma \rightarrow \infty} \frac{\frac{d}{d\gamma} I_n(\gamma, 1)}{\frac{d}{d\gamma} \gamma^{\alpha(n-2)} \ln \gamma} \\ &= \limsup_{\gamma \rightarrow \infty} \frac{n L(\gamma^{-1}) I_{n-1}(\gamma, 1 - \gamma^{-1}) \gamma^{-2}}{(1 + \alpha(n-2) \ln \gamma) \gamma^{\alpha(n-2)-1}}, \quad n = 2, 3, \dots \end{aligned} \quad (26)$$

- $n = 2$. The expression in (26) becomes

$$\frac{2 L(\gamma^{-1}) I_1(\gamma, 1 - \gamma^{-1}) \gamma^{-2}}{\gamma^{-1}} = \frac{2 L(\gamma^{-1}) I_1(\gamma, 1 - \gamma^{-1})}{\gamma}.$$

Observe that

$$\begin{aligned} L(\gamma^{-1}) &= (1 - \gamma^{-1})^\alpha \gamma^{\alpha+1} = \mathcal{O}(\gamma^{\alpha+1}), \quad \gamma \rightarrow \infty; \\ L(1 - \gamma^{-1}) &= \gamma^{-\alpha} (1 - \gamma^{-1})^{-(\alpha+1)} = \mathcal{O}(\gamma^{-\alpha}), \quad \gamma \rightarrow \infty; \\ I_1(\gamma, 1 - \gamma^{-1}) &= \int_{1-\gamma^{-1}}^1 L(y) dy \leq \gamma^{-1} L(1 - \gamma^{-1}) = \mathcal{O}(\gamma^{-(\alpha+1)}), \quad \gamma \rightarrow \infty, \end{aligned}$$

where the inequality follows because the function $L(y)$ is decreasing on $(0, 1]$. Hence,

$$\limsup_{\gamma \rightarrow \infty} \frac{2 L(\gamma^{-1}) I_1(\gamma, 1 - \gamma^{-1})}{\gamma} = \limsup_{\gamma \rightarrow \infty} \frac{(\text{a constant}) \times \gamma^{\alpha+1} \gamma^{-(\alpha+1)}}{\gamma} = 0.$$

- $n \geq 2$. Assume (24) holds for n . Then reasoning as above and using Lemma 6 for equality (i), we get for $n + 1$

$$\begin{aligned} &\limsup_{\gamma \rightarrow \infty} \frac{I_{n+1}(\gamma, 1)}{\gamma^{\alpha(n-1)} \ln \gamma} \\ &= \limsup_{\gamma \rightarrow \infty} \frac{\frac{d}{d\gamma} I_{n+1}(\gamma, 1)}{\frac{d}{d\gamma} \gamma^{\alpha(n-1)} \ln \gamma} \\ &= \limsup_{\gamma \rightarrow \infty} \frac{(n+1) L(\gamma^{-1}) I_n(\gamma, 1 - \gamma^{-1}) \gamma^{-2}}{(1 + \alpha(n-1) \ln \gamma) \gamma^{\alpha(n-1)-1}} \\ &\stackrel{(i)}{=} \limsup_{\gamma \rightarrow \infty} \frac{(n+1) L(\gamma^{-1}) (I_n(\gamma, 1) + o(1)) \gamma^{-2}}{(1 + \alpha(n-1) \ln \gamma) \gamma^{\alpha(n-1)-1}} \\ &= \limsup_{\gamma \rightarrow \infty} \frac{(\text{a constant}) \times \gamma^{\alpha+1} I_n(\gamma, 1) \gamma^{-2} + o(1)}{(\text{a constant}) \times \gamma^{\alpha(n-1)-1} \ln \gamma} \\ &= \limsup_{\gamma \rightarrow \infty} (\text{a constant}) \times \frac{I_n(\gamma, 1) + o(1)}{\gamma^{\alpha(n-2)} \ln \gamma} = 0 \end{aligned}$$

Putting together these arguments we can complete the proof of the Proposition:

$$\begin{aligned}
& \limsup_{\gamma \rightarrow \infty} \frac{\mathbb{E}\mathbb{I}\{M_d \leq \gamma\} Z^2}{\ell^{2-\varepsilon}} \\
& \stackrel{(19)}{\leq} \limsup_{\gamma \rightarrow \infty} \frac{1}{\ell^{2-\varepsilon}} \mathbb{E}_f \mathbb{I}\{M_d \leq \gamma, S_d > \gamma\} \prod_{i=1}^d \frac{\overline{F}(\gamma)}{\overline{F}(\gamma - X_i)} \\
& \stackrel{(20)}{\leq} \limsup_{\gamma \rightarrow \infty} \frac{\sum_{n=2}^d H_n(\gamma)}{\ell^{2-\varepsilon}} \\
& \stackrel{(22)}{=} \limsup_{\gamma \rightarrow \infty} \sum_{n=2}^d \frac{\alpha^n I_n(\gamma)}{\gamma^{\alpha n} \ell^{2-\varepsilon}}
\end{aligned}$$

Now notice that

$$\ell = \overline{F^{*d}}(\gamma) \geq \overline{F}(\gamma) = \gamma^{-\alpha},$$

thus, for $\varepsilon < 1/\alpha$ (that is, $\varepsilon\alpha < 1$)

$$\ell^{2-\varepsilon} \geq \gamma^{-2\alpha} \gamma^{\alpha\varepsilon} \geq \gamma^{-2\alpha} \ln \gamma, \quad \gamma \rightarrow \infty.$$

Combining this with above, we get

$$\limsup_{\gamma \rightarrow \infty} \sum_{n=2}^d \frac{\alpha^n I_n(\gamma)}{\gamma^{\alpha n} \ell^{2-\varepsilon}} \leq \sum_{n=2}^d \alpha^n \limsup_{\gamma \rightarrow \infty} \frac{I_n(\gamma)}{\gamma^{\alpha(n-2)} \ln \gamma} = 0$$

4.2. Light-tailed case

In this section we consider the case where F belongs to a subfamily of light-tailed distributions as defined by Embrechts and Goldie [10]. We say that a distribution F belongs to the *Embrechts-Goldie* family of distributions indexed by the parameter $\theta \geq 0$ and denoted $\mathcal{L}(\theta)$, if

$$\lim_{\gamma \rightarrow \infty} \frac{\overline{F}(\gamma + x)}{\overline{F}(\gamma)} = e^{-\theta x}. \quad (27)$$

If θ is strictly larger than 0 then $\mathcal{L}(\theta)$ contains light-tailed distributions exclusively and is often referred as the *exponential class*. This is a very rich class of distributions that includes several well know light-tailed distributions such as the exponential, gamma and phase-type. In contrast, if $\theta = 0$, then $\mathcal{L}(0)$ corresponds to the class of *long-tailed distributions* which is a large subclass of heavy-tailed distributions. In this section we concentrate on the light-tailed case $\theta > 0$, but in order to derive our efficiency statements we draw some results for the class of the so called *long-tailed functions* (cf. [11, Definition 2.14]). More precisely, h is long-tailed if it is ultimately positive and

$$\lim_{\gamma \rightarrow \infty} \frac{h(\gamma + x)}{h(\gamma)} = 1, \quad \forall x. \quad (28)$$

Obviously, if $F \in \mathcal{L}(0)$, then the tail probability \bar{F} is long tailed. Important properties for the exponential class ($\theta > 0$) are

- $\mathcal{L}(\theta)$ is closed under convolutions [10, Theorem 3]. That is, if $F \in \mathcal{L}(\theta)$, then the d -fold convolution $F^{*d} \in \mathcal{L}(\theta)$.
- Define for $\alpha > 0$ the distribution $G(x) = 1 - (\bar{F}(x))^\alpha$. One can easily check that $G \in \mathcal{L}(\alpha\theta)$ whenever $F \in \mathcal{L}(\theta)$.
- The tail probability can be decomposed into the product of an exponential and a long tailed function

$$\bar{F}(\gamma) = e^{-\theta\gamma} h(\gamma). \quad (29)$$

Decomposition (29) will be useful for proving efficiency of the proposed estimator, but it is also interesting on its own. To verify it we define $h(\gamma) := \bar{F}(\gamma) e^{\theta\gamma}$. Since $F \in \mathcal{L}(\theta)$ it follows that

$$\lim_{\gamma \rightarrow \infty} \frac{h(\gamma+x)}{h(\gamma)} = \lim_{\gamma \rightarrow \infty} \frac{h(\gamma+x) e^{-\theta(\gamma+x)}}{h(\gamma) e^{-\theta(\gamma+x)}} = \lim_{\gamma \rightarrow \infty} \frac{\bar{F}(\gamma+x)}{\bar{F}(\gamma) e^{-\theta x}} = 1.$$

The next property states that the asymptotic decay of a long-tailed function is slower than the exponential rate [11, Lemma 2.17]. More precisely, if h is long tailed, then

$$\lim_{\gamma \rightarrow \infty} \frac{h(\gamma)}{e^{-\varepsilon\gamma}} = \infty, \quad \forall \varepsilon > 0. \quad (30)$$

These properties will be employed to construct an asymptotic upper bound for the semi-parametric estimator. In particular, the following Lemma shows that the ratio of two tail convolutions of the same distribution in $\mathcal{L}(\theta)$ cannot increase/decrease faster than at exponential rate.

Lemma 4. *Let $F \in \mathcal{L}(\theta)$, $\theta > 0$, and $d_1, d_2 \in \mathbb{N}$. Then $\overline{F^{*d_1}}(\gamma) / \overline{F^{*d_2}}(\gamma) = o(e^{\varepsilon\gamma})$, $\forall \varepsilon > 0$.*

Proof. Since $\mathcal{L}(\theta)$ is closed by convolution, then $F^{*d_1}, F^{*d_2} \in \mathcal{L}(\theta)$ and their tail distributions have decompositions as in (29) for some long tailed functions h_1 and h_2 . Therefore

$$\frac{\overline{F^{*d_1}}(\gamma)}{\overline{F^{*d_2}}(\gamma)} = \frac{h_1(\gamma)e^{-\theta\gamma}}{h_2(\gamma)e^{-\theta\gamma}} = \frac{h_1(\gamma)}{h_2(\gamma)}.$$

We first argue that both $h_1(\cdot)/h_2(\cdot)$ and its reciprocal function are long-tailed. This is so, because they are ultimately positive, and

$$\frac{h_1(\gamma+x)/h_2(\gamma+x)}{h_1(\gamma)/h_2(\gamma)} = \frac{h_1(\gamma+x)}{h_1(\gamma)} \times \frac{h_2(\gamma)}{h_2(\gamma+x)} \rightarrow 1.$$

The reciprocal function goes similarly. Thus, $h_2(\cdot)/h_1(\cdot)$ satisfies condition (30), which says

$$\lim_{\gamma \rightarrow \infty} \frac{h_2(\gamma)/h_1(\gamma)}{e^{-\varepsilon\gamma}} = \infty.$$

Clearly, this is equivalent to

$$\lim_{\gamma \rightarrow \infty} \frac{h_1(\gamma)/h_2(\gamma)}{e^{\varepsilon\gamma}} = 0.$$

□

We also have the following.

Assumption A: Let h be a long-tailed function such that $h(x) > 0$ for all $x \geq 0$. Then $G(\gamma) := \sup\{h(\gamma)/h(x) : 0 \leq x \leq \gamma\} = o(e^{\varepsilon\gamma})$ for all $\varepsilon > 0$.

Proposition 3. (Logarithmic efficiency of $\widehat{\ell}$.) *If Assumption A holds, the estimator $Z = \mathbb{I}\{S(\mathbf{X}) > \gamma\} \frac{f(\mathbf{X})}{g(\mathbf{X})}$ satisfies*

$$\lim_{\gamma \uparrow \infty} \frac{\mathbb{E}Z^2}{\ell^{2-\varepsilon}(\gamma)} = 0, \quad \forall \varepsilon > 0.$$

Proof. Recall

$$\mathbb{E}Z^2 = \mathbb{E}_f \mathbb{I}\{S(\mathbf{X}) > \gamma\} \prod_{i=1}^d \frac{\overline{F^{*d}}(\gamma)}{F^{*(d-1)}(\gamma - X_i)}.$$

We write

$$\prod_{i=1}^d \frac{\overline{F^{*d}}(\gamma)}{F^{*(d-1)}(\gamma - X_i)} = H(\gamma) \prod_{i=1}^d \frac{\overline{F^{*(d-1)}}(\gamma)}{F^{*(d-1)}(\gamma - X_i)},$$

where $H(\gamma) := \left[\overline{F^{*d}}(\gamma) / \overline{F^{*(d-1)}}(\gamma) \right]^d$. Since $F^{*(d-1)} \in \mathcal{L}(\theta)$ we can use the decomposition (29) to write $\overline{F^{*(d-1)}}(\gamma) = h(\gamma)e^{-\theta\gamma}$ for some $h(\cdot)$ long tailed function. Hence, we obtain the following bound

$$\prod_{i=1}^d \frac{\overline{F^{*(d-1)}}(\gamma)}{F^{*(d-1)}(\gamma - X_i)} = \prod_{i=1}^d \frac{h(\gamma)}{h(\gamma - X_i)} \frac{e^{-\theta\gamma}}{e^{-\theta(\gamma - X_i)}} \leq \left(\sup_{0 \leq x \leq \gamma} \frac{h(\gamma)}{h(\gamma - x)} \right)^d \prod_{i=1}^d e^{-\theta X_i} = (G(\gamma))^d e^{-\theta S(\mathbf{X})}$$

where $G(\gamma) := \sup_{0 \leq x \leq \gamma} \{h(\gamma)/h(\gamma - x)\}$. Using these we obtain

$$\frac{\mathbb{E}Z^2}{\ell^{2-\varepsilon}(\gamma)} \leq \frac{H(\gamma)G^d(\gamma)}{\ell^{2-\varepsilon}(\gamma)} \mathbb{E}_f \mathbb{I}\{S(\mathbf{X}) > \gamma\} e^{-\theta S(\mathbf{X})},$$

where $\theta > 0$. Hence,

$$\mathbb{E}_f \mathbb{I}\{S(\mathbf{X}) > \gamma\} e^{-\theta S(\mathbf{X})} \leq e^{-\theta\gamma} \mathbb{P}_f(S(\mathbf{X}) > \gamma) = e^{-\theta\gamma} \ell.$$

Thus we get

$$\frac{\mathbb{E}Z^2}{\ell^{2-\varepsilon}(\gamma)} \leq \frac{H(\gamma)G^d(\gamma)e^{-\theta\gamma}}{\ell^{1-\varepsilon}(\gamma)}.$$

Applying the properties of the exponential class we can write

$$\ell^{1-\varepsilon} = (\overline{F^{*d}}(\gamma))^{1-\varepsilon} = e^{-\theta(1-\varepsilon)\gamma} h_d(\gamma)$$

for some long tailed function h_d . In consequence,

$$\begin{aligned} \limsup_{\gamma \rightarrow \infty} \frac{\mathbb{E} Z^2}{\ell^{2-\varepsilon}(\gamma)} &\leq \limsup_{\gamma \rightarrow \infty} \frac{H(\gamma)G^d(\gamma)e^{-\theta\gamma}}{\ell^{1-\varepsilon}(\gamma)} \\ &= \limsup_{\gamma \rightarrow \infty} \frac{H(\gamma)G^d(\gamma)e^{-\theta\gamma}}{h_d(\gamma)e^{-(1-\varepsilon)\theta\gamma}} = \limsup_{\gamma \rightarrow \infty} \frac{H(\gamma)G^d(\gamma)}{h_d(\gamma)} e^{-\varepsilon\theta\gamma}. \end{aligned}$$

Now, property (29) and Lemma 4 and Lemma 4.2 imply that none of the functions H , G , h_d^{-1} and their product cannot increase at exponential rate, namely $H(\gamma)G^d(\gamma)/h_d(\gamma) = o(e^{\theta\varepsilon\gamma})$. Hence, the last limit is 0. \square

5. Conclusions

In this paper we have described a procedure for implementing an optimal cross-entropy importance sampling density for the purpose of estimating a rare-event probability, indexed by the rarity parameter γ . The goal is to estimate the optimal importance sampling density for a finite γ within the class of all densities in product form. This optimal importance sampling density is typically not available analytically and this is why in practical simulations we estimate it via MCMC simulation from the minimum variance pdf. The numerical examples suggest that the resulting estimator can yield significantly better efficiency compared to many currently recommended estimators. The same procedure is efficient in both light- and heavy-tailed cases. This is especially relevant for probabilities involving the Weibull distribution with tail index $\alpha < 1$, but close to unity. This setting yields behavior intermediate between the typical heavy- and light-tailed behavior expected of rare-events. As a result, while existing procedures are inefficient or fail completely, our method estimates reliably Weibull probabilities for any values of α , including $\alpha > 1$.

The practical implementation of the proposed method depends on a preliminary MCMC step, which is a powerful, but poorly understood heuristic that needs further investigation. In this article we have established the efficiency of the method in the light- and heavy-tailed case, but have done so by ignoring any errors arising from the preliminary MCMC step. Future work will need to address the impact of the MCMC approximation on the quality of the estimator.

A good starting point for such an analysis might be to consider the probabilistic relative error efficiency concept introduced in [19].

6. Appendix

6.1. Proofs. Section 2.2

Proof of Lemma 1. First note that for any single-variate function h :

$$\begin{aligned} \int_{\mathbb{R}^d} h(x_1) \pi(\mathbf{x}) \, d\mathbf{x} &= \int_{\mathbb{R}} h(x_1) \left(\int_{\mathbb{R}^{d-1}} \pi(x_1, x_2, \dots, x_d) \, dx_2 \cdots dx_d \right) dx_1 \\ &= \int h(x_1) \pi_1(x_1) \, dx_1. \end{aligned}$$

Next, using the properties of the cross-entropy distance we have that

$$\pi_1 = \operatorname{argmin}_{g_1 \in \mathcal{G}_1} \int \pi_1(x_1) \ln \left(\frac{\pi_1(x_1)}{g_1(x_1)} \right) dx_1 = \operatorname{argmax}_{g_1 \in \mathcal{G}_1} \int \pi_1(x_1) \ln g_1(x_1) \, dx_1.$$

Applying these two observations for any $i = 1, \dots, d$ gives

$$\begin{aligned} \operatorname{argmax}_{g_1, \dots, g_d \in \mathcal{G}_1} \int \pi(\mathbf{x}) \ln \left(\prod_{i=1}^d g_i(x_i) \right) d\mathbf{x} \\ &= \operatorname{argmax}_{g_1, \dots, g_d \in \mathcal{G}_1} \sum_{i=1}^d \int \pi(\mathbf{x}) \ln g_i(x_i) \, d\mathbf{x} \\ &= \operatorname{argmax}_{g_1, \dots, g_d \in \mathcal{G}_1} \sum_{i=1}^d \int \pi_i(x_i) \ln g_i(x_i) \, dx_i = \sum_{i=1}^d \operatorname{argmax}_{g_i \in \mathcal{G}_1} \int \pi_i(x_i) \ln g_i(x_i) \, dx_i, \end{aligned}$$

from where we obtain the solution $g_i = \pi_i$ for all $i = 1, \dots, d$. \square

6.2. Proofs. Section 4.1

Lemma 5. Assume $\zeta \geq n\gamma^{-1}$. Then

$$\frac{\partial}{\partial \gamma} I_n(\gamma, \zeta) = n L(\gamma^{-1}) I_{n-1}(\gamma, \zeta - \gamma^{-1}) \gamma^{-2}, \quad n = 2, 3, \dots \quad (31)$$

Proof. The proof is by induction. Recall the recursive introduction of the I_n functions:

$$\begin{aligned} I_1(\gamma, \zeta) &= \int_{\zeta \vee \gamma^{-1}}^1 L(y) \, dy; \\ I_n(\gamma, \zeta) &= \int_{\gamma^{-1}}^{\zeta - (n-2)\gamma^{-1}} L(y) I_{n-1}(\gamma, \zeta - y) \, dy, \quad n = 2, 3, \dots \end{aligned}$$

First consider

$$\begin{aligned} \frac{\partial}{\partial \gamma} I_2(\gamma, \zeta) &= \frac{\partial}{\partial \gamma} \int_{\gamma^{-1}}^{\zeta - \gamma^{-1}} L(y) I_1(\gamma, \zeta - y) \, dy + \frac{\partial}{\partial \gamma} \int_{\zeta - \gamma^{-1}}^{\zeta} L(y) \, dy I_1(\gamma, \gamma^{-1}) \\ &= \left[L(\zeta - \gamma^{-1}) I_1(\gamma, \gamma^{-1}) - L(\gamma^{-1}) I_1(\gamma, \zeta - \gamma^{-1}) - L(\zeta - \gamma^{-1}) I_1(\gamma, \gamma^{-1}) - I_1(\gamma, \zeta - \gamma^{-1}) L(\gamma^{-1}) \right] \frac{d}{d\gamma} \gamma^{-1} \\ &= 2 L(\gamma^{-1}) I_1(\gamma, \zeta - \gamma^{-1}) \gamma^{-2}. \end{aligned}$$

Next, assume that (31) holds for n . Then

$$\begin{aligned}
\frac{\partial}{\partial \gamma} I_{n+1}(\gamma, \zeta) &= \frac{\partial}{\partial \gamma} \int_{\gamma^{-1}}^{\zeta^{-(n-1)\gamma^{-1}}} L(y) I_n(\gamma, \zeta - y) dy \\
&= L(\zeta - (n-1)\gamma^{-1}) I_n(\gamma, (n-1)\gamma^{-1}) \frac{d}{d\gamma} (\zeta - (n-1)\gamma^{-1}) - L(\gamma^{-1}) I_n(\gamma, \zeta - \gamma^{-1}) \frac{d}{d\gamma} \gamma^{-1} \\
&\quad + \int_{\gamma^{-1}}^{\zeta^{-(n-1)\gamma^{-1}}} L(y) \frac{\partial}{\partial \gamma} I_n(\gamma, \zeta - y) dy \\
&= 0 + L(\gamma^{-1}) I_n(\gamma, \zeta - \gamma^{-1}) \gamma^{-2} + \int_{\gamma^{-1}}^{\zeta^{-(n-1)\gamma^{-1}}} L(y) n L(\gamma^{-1}) I_{n-1}(\gamma, \zeta - \gamma^{-1} - y) \gamma^{-2} dy \\
&= L(\gamma^{-1}) I_n(\gamma, \zeta - \gamma^{-1}) \gamma^{-2} + n L(\gamma^{-1}) \int_{\gamma^{-1}}^{\zeta - \gamma^{-1} - (n-2)\gamma^{-1}} L(y) I_{n-1}(\gamma, \zeta - \gamma^{-1} - y) dy \gamma^{-2} \\
&= L(\gamma^{-1}) I_n(\gamma, \zeta - \gamma^{-1}) \gamma^{-2} + n L(\gamma^{-1}) I_n(\gamma, \zeta - \gamma^{-1}) \gamma^{-2} \\
&= (n+1) L(\gamma^{-1}) I_n(\gamma, \zeta - \gamma^{-1}) \gamma^{-2}
\end{aligned}$$

□

Lemma 6. For $n = 1, 2, \dots$:

$$I_n(\gamma, \zeta - \gamma^{-1}) = I_n(\gamma, \zeta) + o(1), \quad \gamma \rightarrow \infty. \quad (32)$$

Proof. Apply induction and the recursive definition of I_n functions.

- $n = 1$.

$$\begin{aligned}
I_1(\gamma, \zeta - \gamma^{-1}) &= \int_{\zeta - \gamma^{-1}}^{\zeta} L(y) dy \\
&= I_1(\gamma, \zeta) + \int_{\zeta - \gamma^{-1}}^{\zeta} L(y) dy \\
&= I_1(\gamma, \zeta) + \gamma^{-1} L(\eta),
\end{aligned}$$

for some $\eta \in (\zeta - \gamma^{-1}, \zeta)$ (mean value theorem). Clearly, the second term is $o(1)$ for $\gamma \rightarrow \infty$.

- $n \geq 1$. Assume (32) holds. Then

$$\begin{aligned}
I_{n+1}(\gamma, \zeta - \gamma^{-1}) &= \int_{\gamma^{-1}}^{\zeta - n\gamma^{-1}} L(y) I_n(\gamma, \zeta - \gamma^{-1} - y) dy \\
&= \int_{\gamma^{-1}}^{\zeta^{-(n-1)\gamma^{-1}}} L(y) (I_n(\gamma, \zeta - y) + o(1)) dy - \int_{\zeta - n\gamma^{-1}}^{\zeta^{-(n-1)\gamma^{-1}}} L(y) I_n(\gamma, \zeta - \gamma^{-1} - y) dy \\
&= I_{n+1}(\gamma, \zeta) + o(1) \int_{\gamma^{-1}}^{\zeta^{-(n-1)\gamma^{-1}}} L(y) dy - \gamma^{-1} L(\eta) I_n(\gamma, \zeta - \gamma^{-1} - \eta) \\
&= I_{n+1}(\gamma, \zeta) + o(1), \quad \gamma \rightarrow \infty.
\end{aligned}$$

□

References

- [1] S. Asmussen and P. W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer-Verlag, New York, 2007.
- [2] S. Asmussen and D. P. Kroese. Improved algorithms for rare event simulation with heavy tails. *Advances in Applied Probability*, 38(2), 2006.
- [3] S. Asmussen, D. P. Kroese, and R. Y. Rubinstein. Heavy tails, importance sampling and cross-entropy. *Stochastic Models*, 21(1), 2005.
- [4] Søren Asmussen and Dominik Kortschak. On error rates in rare event simulation with heavy tails. In *Proceedings of the Winter Simulation Conference*, page 38. Winter Simulation Conference, 2012.
- [5] Z. I. Botev and D. P. Kroese. Efficient Monte Carlo simulation via the generalized splitting method. *Statistics and Computing*, 2010. DOI:10.1007/s11222-010-9201-4.
- [6] Zdravko I. Botev, Pierre L'Ecuyer, and Bruno Tuffin. Markov chain importance sampling with applications to rare event probability estimation. *Statistics and Computing*, 23(2):271–285, 2013.
- [7] J. C. C. Chan and D. P. Kroese. Improved cross-entropy method for estimation. *manuscript*, 2011.
- [8] Joshua C.C. Chan, Peter W. Glynn, and Dirk P. Kroese. A comparison of cross-entropy and variance minimization strategies. *Journal of Applied Probability*, 48:183–194, 2011.
- [9] P. Embrechts, C. Kluppelberg, and T. Mikosch. *Modelling extremal events*. Springer-Verlag, Berlin, 1997.
- [10] Paul Embrechts and Charles Goldie. *On closure and factorisation properties of subexponential and related distributions*. Cambridge Univ Press, 1980.
- [11] Sergey Foss, Dmitry Korshunov, and Stan Zachary. *An introduction to heavy-tailed and subexponential distributions*. Springer, 2011.
- [12] Samim Ghamami and Sheldon Ross. Improving the Asmussen–Kroese-type simulation estimators. *Journal of Applied Probability*, 49(4):1188–1193, 2012.

- [13] Jürgen Hartinger and Dominik Kortschak. On the efficiency of the Asmussen–Kroese estimator and its application to stop-loss transforms. *Blätter der DGVMF*, 30(2):363–377, 2009.
- [14] S. Juneja. Estimating tail probabilities of heavy tailed distributions with asymptotically zero relative error. *Queueing Systems*, 57(2-3):115–127, 2007.
- [15] D. P. Kroese, R. Y. Rubinstein, and P. W. Glynn. The cross-entropy method for estimation. In V. Govindaraju and C.R. Rao, editors, *Handbook of Statistics, Volume 31: Machine Learning*. North Holland, 2011.
- [16] D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo methods*. John Wiley & Sons, New York, 2011.
- [17] P. L’Ecuyer, J. Blanchet, B. Tuffin, and P. W. Glynn. Asymptotic robustness of estimators in rare-event simulation. *ACM Transactions on Modeling and Computer Simulation*, 20(1), 2010. Article 6.
- [18] R. Y. Rubinstein and D. P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer-Verlag, New York, 2004.
- [19] Bruno Tuffin and Ad Ridder. Probabilistic bounded relative error for rare event simulation learning techniques. In *Proceedings of the Winter Simulation Conference*, page 40. Winter Simulation Conference, 2012.
- [20] Roderick Wong. *Asymptotic approximation of integrals*, volume 34. Society for Industrial and Applied Mathematics, 2001.