The Transform Likelihood Ratio Method for Rare Event Simulation with Heavy Tails

D.P. Kroese
Dept. of Mathematics, University of Queensland, Brisbane 4072, Australia.

R.Y. Rubinstein
Faculty of Industrial Engineering and Management, Technion, Haifa, Israel.

Abstract

We present a novel method, called the *transform likelihood ratio* (TLR) method, for estimation of rare event probabilities with *heavy-tailed* distributions. Via a simple transformation (change of variables) technique the TLR method reduces the original rare event probability estimation with heavy tail distributions to an equivalent one with *light* tail distribution, such as the uniform or exponential distribution. Once this transformation has been established we estimate the rare event probability via importance sampling, using the classical *exponential change of measure* or the *standard likelihood ratio* change of measure. In the latter case the importance sampling distribution is chosen from the *same parametric family* as the transformed distribution. We estimate the optimal parameter vector of the importance sampling distribution using the *cross-entropy* method. We prove the polynomial complexity of the TLR method for certain heavy-tailed models and demonstrate numerically its high efficiency for various heavy-tailed models previously thought to be intractable. We also show that the TLR method can be viewed as a universal tool in the sense that not only it provides a unified view for heavy-tailed simulation but also can be efficiently used in simulation with light-tailed distributions. We present extensive simulation results which support the efficiency of the TLR method.

Keywords. Cross-Entropy, Heavy Tail Distributions, Rare Events, Simulation, Importance Sampling, Likelihood Ratio

1 Introduction

The performance of modern systems, such as coherent reliability systems, inventory systems, insurance risk, storage systems, computer networks and telecommunications networks, is often characterized by probabilities of *rare events* and is frequently studied through simulation. However, estimation of rare event probabilities with crude Monte Carlo techniques requires a prohibitively large numbers of trials. Two methods, called *splitting/RESTART* and *importance sampling* (IS), have been extensively investigated by the simulation community in the last decade.

The basic idea of *splitting* proposed by Kahn and Harris [20] is to partition the state-space of the system into a series of nested subsets and to consider the rare event as the intersection of a nested sequence of events. When a given subset is entered by a sample trajectory during the simulation, numerous random re-trials are generated with the initial state for each re trial being the state of the system at the entry point. Thus, by doing so, the system trajectory is split into a number of new sub-trajectories, hence the name "splitting". A similar idea has been developed by Villen-Altamarino and Villen-Altamarino [31, 32] into a refined simulation technique under the name *RESTART* which has been extended by different authors [9, 10, 11, 12, 13, 16, 15, 17, 27, 28] to the multiple threshold case.

The main idea of IS [30, 14] is to make the occurrence of rare events more frequent by carrying out the simulation under a different probability distribution – the so-called *change of measure* (CM) – and to estimate the probability of interest via a corresponding *likelihood ratio* (LR) estimator. The aim is to select a CM that minimizes the variance of the LR estimator. It is well-known that, in theory, there exists a CM that yields a *zero-variance* LR estimator. However, in practice such an optimal CM cannot be computed since it depends on the underlying quantity/quantities being estimated.

Prominent among the CMs is the *exponential change of measure* (ECM). Here, instead of the original pdf $f(x)$, the simulation is carried out under an "exponentially twisted" pdf $f_\theta(x) = c\,\mathrm{e}^{\theta x} f(x)$, where θ is called the *twisting* or *tilting parameter* and c is a normalizing constant. ECM often yields efficient and sometimes "optimal" IS estimates, see for example Sadowsky [26] and Asmussen and Rubinstein [5], but is usually feasible only for relative simple models, see also [18, 21, 29].

An alternative approach to ECM is to use an IS pdf, say $f(\boldsymbol{x};\boldsymbol{v})$, which belongs to the *same parametric family* as the original distribution (also called the *nominal* distribution), say $f(\boldsymbol{x};\boldsymbol{u})$. We shall call such an approach the *standard likelihood ratio* (SLR) approach. Similar to ECM, the SLR approach typically does not lead to the optimal zero-variance estimator, but yields significant variance reduction, see for instance [24] and below. The advantage of such an approach is that (a) it can be applied to rather general static and dynamic models, and (b) the optimal reference parameter $\boldsymbol{v}^*$ of the IS density $f(\boldsymbol{x};\boldsymbol{v})$ can be derived with standard optimization techniques.

We show in this paper that the SLR approach is readily applicable to both light- *and* heavy-tailed distributions. Recall that a random variable X with distribution function F is said to have a *light-tail* distribution if

$$\mathbb{E}e^{sX} < \infty, \qquad \text{for some } s > 0 \, .$$

By Markov's inequality, we have $\mathbb{E}e^{sX} \geq \mathbb{E}e^{sX} I_{\{X>x\}} \geq e^{sx}\mathbb{P}(X > x)$, so that

$$\mathbb{P}(X > x) \leq e^{-sx} \, c \, , \quad x \geq 0 \, ,$$

for some constant c. In other words, X has a "tail" $\bar{F}(x) = 1 - F(x)$ which decays at an exponential rate or faster. Examples of the light-tailed distributions are the exponential, normal, geometric, Poisson and any distribution with bounded support. Also the Weibull distribution with *increasing failure rate*, that is $\bar{F}(x) = e^{-x^a}$ with $a \geq 1$, is a light-tail distribution.

When $\mathbb{E}e^{sX} = \infty$ for all $s > 0$, X is said to have a *heavy-tail* distribution. Examples of heavy-tail distributions are the log-normal, Rayleigh and the Weibull distribution with *decreasing failure rate*, that is $\bar{F}(x) = e^{-x^a}$, $a < 1$. Also any regularly varying distribution, that is $\bar{F}(x) = L(x)/x^\alpha$, with $L(tx)/L(x) \to 1$ as $x \to \infty$ for all $t > 0$, is heavy-tail. A typical example is the Pareto distribution, which has a tail $\bar{F}(x) = (1 + cx)^{-a}, x \geq 0, (a, c > 0)$. We write $X \sim \mathsf{Pareto}(a, c)$ to indicate that X has the above distribution.

A particularly important class of heavy-tailed distributions is that of the sub-exponential distributions. A distribution with cdf F on $(0, \infty)$ is said to be *sub-exponential* if, with $X_1, X_2, \ldots, X_n$ a random sample from F, we have

$$\lim_{\gamma \to \infty} \frac{\mathbb{P}(X_1 + \cdots + X_n > \gamma)}{\mathbb{P}(X_1 > \gamma)} = n \, , \tag{1}$$

for all n. Examples are the Pareto and log-normal distributions and the Weibull distribution with decreasing failure rate. See [8] for additional properties of this class of distributions.

Because by definition the exponential moments do not exist for heavy-tailed distributions, the exponential change of measure is intrinsically impossible for heavy-tailed distributions when a positive twisting parameter is required. So an alternative method must be used. Asmussen, Binswanger and Højgaard in their landmark paper [3] consider various estimators for rare events of the form $\{S_N > x\}$, where S_N is the random or deterministic sum of i.i.d. positive random variables with sub-exponential pdf, $f(x)$ say. Two asymptotic efficient estimators are given. The first one, based on Asmussen and Binswanger [2] uses conditional Monte Carlo [24] in combination with order statistics. The second estimator uses importance sampling, where the IS density, $h(x)$ say, consists of two parts: for small values of x, $g(x)$ is proportional to $f(x)$ and for large values of x, $g(x)$ is much larger than $f(x)$, decreasing slightly faster than $1/x$. Juneja and Shahabuddin [19] consider a similar problem as in [3] and their approach is to estimate $\{S_N > x\}$ via IS using a density $h(x)$ which is obtained from the original $f(x)$ by "twisting" the *hazard rate*. Several variations of this idea are considered. Note that all the above heavy tail methods have limited application since they deal basically only with the estimation of probabilities of the above events $\{S_N > x\}$.

The effectiveness of the SLR method to rare event simulation depends strongly on (a) the selection of a proper class of IS distributions $\{f(\cdot; \boldsymbol{v})\}$ and (b) an efficient method for determining the optimal reference parameter $\boldsymbol{v}^*$.

We address (b), in this paper by using the *cross-entropy* (CE) method to estimate the optimal reference parameter in any SLR procedure. The CE method was proposed in [22] an *adaptive* IS algorithm for rare events simulation, in which the reference parameter $\boldsymbol{v}^*$ is estimated by minimizing the sample variance of the SLR estimator. The proposed algorithm is called the *variance minimization* (VM) algorithm. In [23] this IS algorithm was further modified to minimize, instead of the sample variance, the sample Kullback-Leibler distance, or *cross-entropy* (CE) distance, between the theoretical zero-variance change of measure and the importance sampling distribution. The estimation method thus obtained is called the *simulated cross-entropy* or just the *cross-entropy* (CE) method.

We address (a), by presenting a novel method, called the *transform likelihood ratio* (TLR) method, for constructing efficient IS estimators that are applicable for both light- and heavy-tail distributions. The idea is to transform the random variables and to apply a change of measure to the distribution of the transformed random variables. This simple "change of variable" technique allows us to transform an original rare event probability with heavy tail distributions to an equivalent (auxiliary) one with an arbitrary tail distribution, such as the uniform or exponential distribution, and then we apply a change of measure to the new (auxiliary) distribution. We typically transform to light-tailed distributions, and then apply the ECM or the SLR method to obtain a convenient class of IS distributions. Recall that in the latter case, the IS distribution belongs to the *same parametric family* as the original auxiliary one. As mentioned before we shall use the CE method to estimate the optimal parameter vector of the (parametric) IS distribution.

The goal of this paper is to show that the SLR and TLR methods broaden substantially the application scope of rare event simulation, and to demonstrate their high efficiency numerically for various heavy-tailed models previously thought to be intractable. We also show that the TLR method can be viewed as an universal tool in the sense that it can be efficiently used in light-tailed simulation as well. In a forthcoming paper [4] the focus will be more on the *complexity* of the estimators. In particular we will prove the polynomial complexity of the TLR method for various sums of heavy-tailed random variables and explore in more detail the asymptotic optimality of various queueing models, when using the SLR or TLR method. In the appendix of the present paper we give a direct proof of polynomial complexity of the TLR method for the sum of $n = 2$ heavy tail Weibull random variables, and we conjecture that similar results hold for general n.

The theoretical framework in which one typically examines rare-event probability estimation is based on complexity theory according to which the IS estimators are classified either as *polynomial-time* or as *exponential-time*. It is shown in [5, 24] that for an (unbiased) IS estimator, $\widehat{\ell}(x)$ of $\ell(x)$, to be polynomial-time as a function of some x, it suffices that its *squared coefficient* *of variation* (SCV),

$$\kappa^2(x) = \frac{\mathrm{Var}(\widehat{\ell}(x))}{\ell^2(x)}. \tag{2}$$

or its *relative error*, $\kappa(x)$, be bounded in x by some polynomial function, $p(x)$. For such polynomial-time estimators, the required sample size to achieve a fixed relative error does not grow too fast as the event becomes rarer. Because polynomial complexity is not always easy to achieve or to prove, the weaker notion of *asymptotic optimality* is often used, meaning

$$\lim_{x\to\infty} \frac{\ln \mathbb{E}(\widehat{\ell})^2}{\ln \ell^2(x)} = 1 \ . \tag{3}$$

For a detailed discussion on complexity, see [5].

The remainder of this paper is organized as follows. In Section 2 we describe the main ideas behind the SLR method. Here we also present a general adaptive CE procedure for estimating the optimal reference parameters for the SLR method. It can be readily implemented if the underlying distributions have *finite support* or if they belong to a *natural exponential family*, since in those cases there are *analytical* solutions to those optimization problems. In Section 3 we present the TLR method and its application to heavy-tail distributions. We provide several enlightening examples on the standard SLR method and its TLR modification and demonstrate analytically how the latter can outperform the former. In Section 4 we illustrate that seemingly different implementations of SLR and TLR may in fact be completely equivalent. Section 5 deals with the estimation of tail probabilities for the waiting time in a GI/G/1 queue with heavy-tailed service time and/or inter-arrival time distributions. In Section 6 we demonstrate numerically the efficiency of the TLR method for fast estimation of rare events for various simulation models involving light and heavy tail distributions. In the Appendix we derive the asymptotic form of the minimal variance parameter for the TLR estimator for sum of two i.i.d. Weibull random variables with heavy tails, and prove polynomial complexity.

2 The SLR Method via Importance Sampling and Cross Entropy

In this section we discuss the main ideas behind the CE algorithm for rare event simulation following closely [7].

Let S be a real function taking values in some space $\mathcal{X}$, and let $\boldsymbol{X}$ be a random element in $\mathcal{X}$ with pdf $f(\cdot;\boldsymbol{u})$ in some parametric family $\mathcal{F} = \{f(\cdot;\boldsymbol{v}),\ \boldsymbol{v} \in \mathcal{V}\}$, with respect to a certain base measure μ. Typically, $\mathcal{X}$ is some subset of $\mathbb{R}^n$ and $\boldsymbol{X}$ is a random vector $(X_1,\dots,X_n)$. Suppose we are interested in the probability that $S(\boldsymbol{X})$ is greater than or equal to some real number γ – which we will refer to as *level* – under $f(\cdot;\boldsymbol{u})$. This probability can be expressed as

$$\ell = \mathbb{P}_{\boldsymbol{u}}(S(\boldsymbol{X}) \geq \gamma) = \mathbb{E}_{\boldsymbol{u}}\, I_{\{S(\boldsymbol{X})\geq\gamma\}} = \int I_{\{S(\boldsymbol{x})\geq\gamma\}}\, f(\boldsymbol{x};\boldsymbol{v})\, \mu(d\boldsymbol{x})\ ,$$

If this probability is very small, we call $\{S(\boldsymbol{X}) \geq \gamma\}$ a *rare event.*

The naive way to estimate ℓ is to use *crude Monte-Carlo* (CMC) simulation: Draw a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ from $f(\cdot\,; \boldsymbol{u})$; then

$$\frac{1}{N} \sum_{i=1}^{N} I_{\{S(\boldsymbol{X}_i) \geq \gamma\}}$$

is an unbiased estimator of ℓ. However this poses serious problems when $\{S(\boldsymbol{X}) \geq \gamma\}$ is a rare event since a large simulation effort is required in order to estimate ℓ accurately.

An alternative approach is based on importance sampling: take a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ from an *importance sampling* (IS) density g on $\mathcal{X}$, and estimate ℓ using the following unbiased estimator

$$\widehat{\ell} = \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\boldsymbol{X}_i) \geq \gamma\}} W(\boldsymbol{X}_i) \, , \tag{4}$$

where $W(\boldsymbol{X}) = f(\boldsymbol{X}; \boldsymbol{u})/g(\boldsymbol{X})$ is the *likelihood ratio* (LR). The estimator in (4) is called the *likelihood ratio estimator.*

It is well known [24] that the optimal way to estimate ℓ is to use the change of measure with density

$$g^*(\boldsymbol{x}) = \frac{I_{\{S(\boldsymbol{x}) \geq \gamma\}} f(\boldsymbol{x}; \boldsymbol{u})}{\ell} \, . \tag{5}$$

Namely, by using this change of measure we have in (4)

$$I_{\{S(\boldsymbol{X}_i) \geq \gamma\}} \frac{f(\boldsymbol{X}_i; \boldsymbol{u})}{g^*(\boldsymbol{X}_i)} = \ell,$$

for all i. Since ℓ is a constant, the estimator (4) has zero variance, and we need to produce only $N = 1$ sample.

But, of course, ℓ in (5) is unknown, and sampling from the optimal importance sampling density g^* is therefore problematic. Instead, consider the situation where the choice of IS densities g is restricted to the *same* parametric family $\mathcal{F}$; so g differs from the original density $f(\cdot; \boldsymbol{u})$ by a single parameter (vector) $\boldsymbol{v}$, which we will call the *reference parameter.* We will write the likelihood ratio in (4), with $g(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{v})$, as

$$W(\boldsymbol{X}; \boldsymbol{u}, \boldsymbol{v}) = \frac{f(\boldsymbol{X}; \boldsymbol{u})}{f(\boldsymbol{X}; \boldsymbol{v})} \, . \tag{6}$$

In this case the LR estimator $\widehat{\ell}$ in (4) becomes

$$\widehat{\ell} = \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\boldsymbol{X}_i) \geq \gamma\}} W(\boldsymbol{X}_i; \boldsymbol{u}, \boldsymbol{v}), \tag{7}$$

where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ is a random sample from $f(\cdot; \boldsymbol{v})$. We will call (7) the *SLR estimator*, in contrast to the (non-parametric) LR estimator (4). To find an optimal $\boldsymbol{v}$ in the SLR estimator $\widehat{\ell}$ one typically considers [24] the following variance minimization program

$$\min_{\boldsymbol{v}} \mathrm{Var}_{\boldsymbol{v}} I_{\{S(\boldsymbol{X})\geq\gamma\}}\, W(\boldsymbol{X};\boldsymbol{u},\boldsymbol{v}) \ . \tag{8}$$

Since under $f(\cdot;\boldsymbol{v})$ the expectation $\ell = \mathbb{E}_{\boldsymbol{v}} I_{\{S(\boldsymbol{X})\geq\gamma\}}\, W(\boldsymbol{X};\boldsymbol{u},\boldsymbol{v})$ is constant, the optimal solution of (8) coincides with that of

$$\min_{\boldsymbol{v}} V(\boldsymbol{v}) = \min_{\boldsymbol{v}} \mathbb{E}_{\boldsymbol{v}} I_{\{S(\boldsymbol{X})\geq\gamma\}}\, W^2(\boldsymbol{X};\boldsymbol{u},\boldsymbol{v}) \ . \tag{9}$$

The above optimization problem can still be difficult to solve, since the density with respect to which the expectation is computed depends on the decision variable $\boldsymbol{v}$. To overcome this obstacle, we rewrite (9) as

$$\min_{\boldsymbol{v}} V(\boldsymbol{v}) = \min_{\boldsymbol{v}} \mathbb{E}_{\boldsymbol{w}} I_{\{S(\boldsymbol{X})\geq\gamma\}}\, W(\boldsymbol{X};\boldsymbol{u},\boldsymbol{v})\, W(\boldsymbol{X};\boldsymbol{u},\boldsymbol{w}) \ . \tag{10}$$

Note that (10) is obtained from (9) by multiplying and dividing the integrand by $f(\boldsymbol{x};\boldsymbol{w})$ where $\boldsymbol{w}$ is an *arbitrary* reference parameter. Note also that in (9) and (10) the expectation is taken with respect to the densities $f(\cdot;\boldsymbol{v})$ and $f(\cdot;\boldsymbol{w})$, respectively. Moreover, $W(\boldsymbol{X};\boldsymbol{u},\boldsymbol{w}) = f(\boldsymbol{X};\boldsymbol{u})/f(\boldsymbol{X};\boldsymbol{w})$, and $\boldsymbol{X} \sim f(\boldsymbol{x};\boldsymbol{w})$. Note finally that for the particular case $\boldsymbol{w} = \boldsymbol{u}$ we obtain from (10)

$$\min_{\boldsymbol{v}} V(\boldsymbol{v}) = \min_{\boldsymbol{v}} \mathbb{E}_{\boldsymbol{u}} I_{\{S(\boldsymbol{X})\geq\gamma\}}\, W(\boldsymbol{X};\boldsymbol{u},\boldsymbol{v}) \ . \tag{11}$$

We shall call each of the equivalent problems (8) – (11), the *variance minimization* (VM) problem ; and we call the parameter vector ${}_*\boldsymbol{v}$, that minimizes the programs (8) – (11) the *optimal VM reference parameter vector*.

An alternative way to find a good reference parameter vector for $\widehat{\ell}$ is based on Kullback-Leibler cross-entropy method. According to the cross-entropy method one can choose the tilting parameter vector $\boldsymbol{v}$ such that the "distance" between g^* above and the density $f(\cdot;\boldsymbol{v})$ is minimal. The Kullback-Leibler distance between g and h is defined as:

$$\mathcal{D}(g,h) = \mathbb{E}_g \ln\frac{g(\boldsymbol{X})}{h(\boldsymbol{X})} = \int g(\boldsymbol{x}) \ln g(\boldsymbol{x})\, \mu(d\boldsymbol{x}) - \int g(\boldsymbol{x}) \ln h(\boldsymbol{x})\, \mu(d\boldsymbol{x}) \ . \tag{12}$$

So, minimizing the Kullback-Leibler distance between g^* in (5) and $f(\cdot;\boldsymbol{v})$ is equivalent to choosing $\boldsymbol{v}$ such that $-\int g^*(\boldsymbol{x}) \ln f(\boldsymbol{x};\boldsymbol{v})\, \mu(d\boldsymbol{x})$ is minimized, or equivalently, that $\int g^*(\boldsymbol{x}) \ln f(\boldsymbol{x};\boldsymbol{v})\, \mu(d\boldsymbol{x})$ is maximized. Formally we write

$$\max_{\boldsymbol{v}} D(\boldsymbol{v}) = \max_{\boldsymbol{v}} \int g^*(\boldsymbol{x}) \ln f(\boldsymbol{x};\boldsymbol{v})\, \mu(d\boldsymbol{x}) \ . \tag{13}$$

Substituting g^* from (5) into (13) we obtain the following optimization program

$$\begin{aligned} \max_{\boldsymbol{v}} D(\boldsymbol{v}) &= \max_{\boldsymbol{v}} \int \frac{I_{\{S(\boldsymbol{x})\geq\gamma\}} f(\boldsymbol{x};\boldsymbol{u})}{\ell} \ln f(\boldsymbol{x};\boldsymbol{v})\, \mu(d\boldsymbol{x}) \\ &= \max_{\boldsymbol{v}} \mathbb{E}_{\boldsymbol{u}}\, I_{\{S(\boldsymbol{X})\geq\gamma\}} \ln f(\boldsymbol{X};\boldsymbol{v}) \ . \end{aligned} \tag{14}$$

Using again importance sampling, with a change of measure $f(\cdot;\boldsymbol{w})$, we can rewrite (14) as

$$\max_{\boldsymbol{v}} D(\boldsymbol{v}) = \max_{\boldsymbol{v}} \mathbb{E}_{\boldsymbol{w}}\, I_{\{S(\boldsymbol{X})\geq\gamma\}}\, W(\boldsymbol{X};\boldsymbol{u},\boldsymbol{w}) \ln f(\boldsymbol{X};\boldsymbol{v}), \tag{15}$$

for *any* tilting parameter $\boldsymbol{w}$. The optimal solution of (15) can be written as

$$\boldsymbol{v}^* = \operatorname*{argmax}_{\boldsymbol{v}} \mathbb{E}_{\boldsymbol{w}}\, I_{\{S(\boldsymbol{X})\geq\gamma\}}\, W(\boldsymbol{X};\boldsymbol{u},\boldsymbol{w}) \ln f(\boldsymbol{X};\boldsymbol{v}). \tag{16}$$

We may *estimate* $\boldsymbol{v}^*$ by solving the following stochastic program (also called *stochastic counterpart* of (15))

$$\max_{\boldsymbol{v}} \widehat{D}(\boldsymbol{v}) = \max_{\boldsymbol{v}} \frac{1}{N}\sum_{i=1}^{N} I_{\{S(\boldsymbol{X}_i)\geq\gamma\}}\, W(\boldsymbol{X}_i;\boldsymbol{u},\boldsymbol{w})\, \ln f(\boldsymbol{X}_i;\boldsymbol{v})\ , \tag{17}$$

where $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_N$ is a random sample from $f(\cdot;\boldsymbol{w})$. The solution of (17) may be readily obtained by solving (with respect to $\boldsymbol{v}$) the following system of equations:

$$\frac{1}{N}\sum_{i=1}^{N} I_{\{S(\boldsymbol{X}_i)\geq\gamma\}}\, W(\boldsymbol{X}_i;\boldsymbol{u},\boldsymbol{w})\, \nabla \ln f(\boldsymbol{X}_i;\boldsymbol{v}) = \boldsymbol{0}, \tag{18}$$

where the gradient is with respect to $\boldsymbol{v}$. This, of course, provided that the expectation and differentiation operators can be interchanged (see [25]) and the function $\widehat{D}$ in (17) is convex and differentiable with respect to $\boldsymbol{v}$. We note that for any fixed $\boldsymbol{x}$ the function

$$\boldsymbol{v} \mapsto \nabla \ln f(\boldsymbol{x};\boldsymbol{v}) \tag{19}$$

is the so-called *score function*. The random variable $\nabla \ln f(\boldsymbol{X};\boldsymbol{v})$ with $\boldsymbol{X} \sim f(\cdot;\boldsymbol{v})$ is called the *efficient score*.

The advantage of this approach is that the solution of (18) can often be calculated *analytically*. In particular, this happens if the distributions of the random variables has a discrete distribution or belong to a *natural exponential family* (NEF). For further details see [7]. It is shown in [7] that asymptotically in γ the optimal tilting parameter vectors obtained from VM and CE programs either coincide or differ very little. So, if not stated otherwise we shall use henceforth the CE program only.

Note that the CE program (17) is useful only in the case where the probability of the "target event" $\{S(\boldsymbol{X}) \geq \gamma\}$ is not too small, say $\ell \geq 10^{-5}$. In such cases, the above program might be useful in terms of determining iteratively a potentially more accurate estimator. In rare-event context, however (say, $\ell \leq 10^{-6}$), the program (17) is useless, since owing to the rarity of the events $\{S(\boldsymbol{X}_i) \geq \gamma\}$, the random variables $I_{\{S(\boldsymbol{X}_i)\geq\gamma\}}$, $i = 1,\ldots,N$ and the associated derivatives of $\widehat{D}(\boldsymbol{v})$, as given in the right-hand side of (18), vanish with high probability for reasonable sizes of N.

To overcome this difficulty, we describe now a *multi-level* algorithm. The idea is to introduce a sequence of reference parameters $\{\boldsymbol{v}_t,\, t \geq 0\}$ and a sequence of levels $\{\gamma_t,\ t \geq 1\}$, and iterate in both γ_t and $\boldsymbol{v}_t$ (see Algorithm 2.1 below).

We initialize by choosing a not very small ρ, say $\rho = 10^{-2}$ and by defining $\boldsymbol{v}_0 = \boldsymbol{u}$. Next, we let γ_1 ($\gamma_1 < \gamma$) be such that, under the original density $f(\boldsymbol{x};\boldsymbol{u})$, the probability $\ell_1 = \mathbb{E}_{\boldsymbol{u}} I_{\{S(\boldsymbol{X}) \geq \gamma_1\}}$ is at least ρ. We then let $\boldsymbol{v}_1$ be the optimal CE reference parameter for estimating ℓ_1, and repeat the last two steps iteratively with the goal of estimating the pair $\{\ell, \boldsymbol{v}^*\}$. In other words, each iteration of the algorithm consists of two main *phases*.

In the first phase γ_t is updated, in the second $\boldsymbol{v}_t$ is updated. Specifically, starting with $\boldsymbol{v}_0 = \boldsymbol{u}$ we obtain the subsequent γ_t and $\boldsymbol{v}_t$ as follows:

1. **Adaptive updating of** γ_t. For a fixed $\boldsymbol{v}_{t-1}$, let γ_t be a $(1-\rho)$-*quantile* of $S(\boldsymbol{X})$ under $\boldsymbol{v}_{t-1}$. That is, γ_t satisfies

$$\mathbb{P}_{\boldsymbol{v}_{t-1}}(S(\boldsymbol{X}) \geq \gamma_t) \;\; \geq \;\; \rho, \tag{20}$$
$$\mathbb{P}_{\boldsymbol{v}_{t-1}}(S(\boldsymbol{X}) \leq \gamma_t) \;\; \geq \;\; 1-\rho, \tag{21}$$

   where $\boldsymbol{X} \sim f(\cdot;\boldsymbol{v}_{t-1})$.

   A simple estimator $\widehat{\gamma}_t$ of γ_t can be obtained by drawing a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ from $f(\cdot;\boldsymbol{v}_{t-1})$, calculating the performances $S(\boldsymbol{X}_i)$ for all i, ordering them from smallest to biggest: $S_{(1)} \leq \ldots \leq S_{(N)}$ and finally, evaluating the $(1-\rho)$ sample quantile as

$$\widehat{\gamma}_t = S_{(\lceil (1-\rho)N \rceil)}. \tag{22}$$

   Note that $S_{(j)}$ is called the j-th *order-statistic* of the sequence $S(\boldsymbol{X}_1)$, $\ldots, S(\boldsymbol{X}_N)$. Note also that $\widehat{\gamma}_t$ is chosen such that the event $\{S(\boldsymbol{X}) \geq \widehat{\gamma}_t\}$ is not too rare (it has a probability of around ρ), and therefore updating the reference parameter via a procedure such as (22) is not void of meaning.

2. **Adaptive updating of** $\boldsymbol{v}_t$. For fixed γ_t and $\boldsymbol{v}_{t-1}$, derive $\boldsymbol{v}_t$ from the solution of the following CE program

$$\max_{\boldsymbol{v}} D(\boldsymbol{v}) = \max_{\boldsymbol{v}} \mathbb{E}_{\boldsymbol{v}_{t-1}} I_{\{S(\boldsymbol{X}) \geq \gamma_t\}} W(\boldsymbol{x};\boldsymbol{u},\boldsymbol{v}_{t-1}) \ln f(\boldsymbol{X};\boldsymbol{v}) \,. \tag{23}$$

   The stochastic counterpart of (23) is as follows: for fixed $\widehat{\gamma}_t$ and $\widehat{\boldsymbol{v}}_{t-1}$, derive $\widehat{\boldsymbol{v}}_t$ from the solution of following program

$$\max_{\boldsymbol{v}} \widehat{D}(\boldsymbol{v}) = \max_{\boldsymbol{v}} \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\boldsymbol{X}_i) \geq \widehat{\gamma}_t\}} W(\boldsymbol{X}_i;\boldsymbol{u},\widehat{\boldsymbol{v}}_{t-1}) \, \ln f(\boldsymbol{X}_i;\boldsymbol{v}) \,. \tag{24}$$

Thus, at the first iteration, starting with $\widehat{\boldsymbol{v}}_0 = \boldsymbol{u}$, to get a good estimate for $\widehat{\boldsymbol{v}}_1$, the target event is artificially made less rare by (temporarily) using a level $\widehat{\gamma}_1$ which is chosen smaller than γ. The value for $\widehat{\boldsymbol{v}}_1$ obtained in this way will (hopefully) make the event $\{S(\boldsymbol{X}) \geq \gamma\}$ less rare in the next iteration, so in the next iteration a value $\widehat{\gamma}_2$ can be used which is closer to γ itself. The algorithm terminates when at some iteration $t = T$ a level is reached which is at least γ and thus the original value of γ can be used without getting too few samples.

As mentioned before, the optimal solutions of (23) and (24) can often be obtained *analytically*, in particular when $f(\boldsymbol{x}; \boldsymbol{v})$ belongs to a NEF.

The above rationale results in the following algorithm (see [7]):

Algorithm 2.1 (Main CE Algorithm for Rare Event Simulation)

1. *Define* $\widehat{\boldsymbol{v}}_0 = \boldsymbol{u}$. *Set* $t = 1$ *(iteration = level counter).*

2. *Generate a sample* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ *from the density* $f(\cdot; \boldsymbol{v}_{t-1})$ *and compute the sample* $(1-\rho)$*-quantile* $\widehat{\gamma}_t$ *according to* (22), *provided* $\widehat{\gamma}_t$ *is less than* γ. *Otherwise set* $\widehat{\gamma}_t = \gamma$.

3. *Use the* **same** *sample* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ *to solve the stochastic program* (24). *Denote the solution by* $\widehat{\boldsymbol{v}}_t$.

4. *If* $\widehat{\gamma}_t < \gamma$, *set* $t = t + 1$ *and reiterate from step 2. Else proceed with step 5.*

5. *Estimate the rare-event probability* ℓ *using the SLR estimate*

$$\widehat{\ell} = \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\boldsymbol{X}_i) \geq \gamma\}} \, W(\boldsymbol{X}_i; \boldsymbol{u}, \widehat{\boldsymbol{v}}_T), \tag{25}$$

where T *denotes the final number of iterations (= number of levels used).*

Remark 2.1 In typical applications the sample size N in step 2 can be chosen much smaller than the final sample size in step 5. When we need to distinguish between the two sample sizes, in particular when reporting numerical experiments, we will use the notation N and N_1 for step 2 and 5, respectively.

Remark 2.2 To obtain a more accurate estimate of $\boldsymbol{v}^*$ it is sometimes useful, especially when the sample size is relatively small, to repeat steps 2–4 for a number of additional iterations after level γ has been reached.

We shall call Algorithm 2.1 the CE algorithm with the standard likelihood ratio (SLR). The convergence of Algorithm 2.1 is given in [7].

Example 2.1 ((Natural) Exponential Family) Let $\boldsymbol{X}$ be a random vector with density $f(\cdot; \boldsymbol{\eta})$, where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)'$ is an m-dimensional parameter column vector. $\boldsymbol{X}$ is said to belong to an *m-parameter exponential family* if there exist real-valued functions $t_i(\boldsymbol{x})$ and $h(\boldsymbol{x}) > 0$ and a (normalizing) function $c(\boldsymbol{\eta}) > 0$, such that

$$f(\boldsymbol{x}; \boldsymbol{\eta}) = c(\boldsymbol{\eta}) \, \mathrm{e}^{\boldsymbol{\eta} \cdot \boldsymbol{t}(\boldsymbol{x})} \, h(\boldsymbol{x}) \,, \tag{26}$$

where $\boldsymbol{t}(\boldsymbol{x}) = (t_1(\boldsymbol{y}), \dots, t_m(\boldsymbol{y}))'$ and $\boldsymbol{\eta} \cdot \boldsymbol{t}(\boldsymbol{x})$ denotes the inner product. The corresponding score function (19) is given by

$$\nabla \ln f(\boldsymbol{x}; \boldsymbol{\eta}) = \frac{\nabla c(\boldsymbol{\eta})}{c(\boldsymbol{\eta})} + \boldsymbol{t}(\boldsymbol{x}) \ ,$$

so that the solution to the CE program (23) (with $\boldsymbol{\theta}$ instead of $\boldsymbol{u}$, and $\boldsymbol{\eta}$ instead of $\boldsymbol{v}$) follows from

$$\mathbb{E}_{\boldsymbol{\eta}_{t-1}} I_{\{S(\boldsymbol{X}) \geq \gamma_t\}} W(\boldsymbol{X}; \boldsymbol{u}, \boldsymbol{\eta}_{t-1}) \left\{ \frac{\nabla c(\boldsymbol{\eta})}{c(\boldsymbol{\eta})} + \boldsymbol{t}(\boldsymbol{X}) \right\} = \boldsymbol{0}, \tag{27}$$

where the likelihood ratio is given by

$$W(\boldsymbol{X}; \boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{c(\boldsymbol{\theta})}{c(\boldsymbol{\eta})} \, \mathrm{e}^{(\boldsymbol{\theta} - \boldsymbol{\eta}) \cdot \boldsymbol{t}(\boldsymbol{X})} \ .$$

Equation (27) can often be solved analytically. It is interesting to note that second moment of each term $I(\boldsymbol{X})W(\boldsymbol{X}) = I_{\{S(\boldsymbol{X}) \geq \gamma\}} W(\boldsymbol{X}; \boldsymbol{\theta}, \boldsymbol{\eta})$ of the SLR estimator (7) can be expressed (see for example (5.3.33) of [24]) as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\eta}} (I(\boldsymbol{X}) W(\boldsymbol{X}))^2 &= \mathbb{E}_{\boldsymbol{\theta}} I(\boldsymbol{X}) W(\boldsymbol{X}) \\
&= \int I(\boldsymbol{x}) \frac{c(\boldsymbol{\theta})}{c(\boldsymbol{\eta})} \, \mathrm{e}^{(\boldsymbol{\theta} - \boldsymbol{\eta}) \cdot \boldsymbol{t}(\boldsymbol{X})} \, c(\boldsymbol{\theta}) \mathrm{e}^{\boldsymbol{\theta} \cdot \boldsymbol{t}(\boldsymbol{x})} \, h(\boldsymbol{x}) \, \mu(d\boldsymbol{x}) \\
&= \frac{c^2(\boldsymbol{\theta})}{c(\boldsymbol{\eta})} \int I(\boldsymbol{x}) \mathrm{e}^{(2\boldsymbol{\theta} - \boldsymbol{\eta}) \cdot \boldsymbol{t}(\boldsymbol{X})} \, h(\boldsymbol{x}) \, \mu(d\boldsymbol{x}) \\
&= \frac{c^2(\boldsymbol{\theta})}{c(\boldsymbol{\eta}) c(2\boldsymbol{\theta} - \boldsymbol{\eta})} \mathbb{E}_{2\boldsymbol{\theta} - \boldsymbol{\eta}} I(\boldsymbol{X}) \\
&= \mathbb{E}_{\boldsymbol{\eta}} W^2 \, \mathbb{E}_{2\boldsymbol{\theta} - \boldsymbol{\eta}} I(\boldsymbol{X}) \ .
\end{aligned} \tag{28}$$

Now let us turn to an important special one-dimensional case. Specifically, let X be a random variable from an exponential family (26) with $t(x) = x$. X is said to belong to a *natural exponential family* (NEF) that is *parameterized by the mean* if the density of X belongs a class $\{f(x; v)\}$ with

$$f(x; v) = \mathrm{e}^{x\eta(v) - \zeta(\eta(v))} \, h(x) \ ,$$

where v is the mean (expectation) corresponding to $f(\cdot; v)$. Note that if $h(x)$ is a pdf, then ζ is the corresponding cumulant function:

$$\zeta(s) = \ln \int \mathrm{e}^{sx} h(x),$$

and $f(\cdot; v)$ is obtained from h by an exponential change of measure with twisting parameter $\eta(v)$. Let $X \sim f(x; u)$ for some nominal reference parameter u. Then [7], the maximizer v^* of (23) is given by

$$v^* \ = \ \frac{\mathbb{E}_u I_{\{S(X) \geq \gamma\}} X}{\mathbb{E}_u \, I_{\{S(X) \geq \gamma\}}} = \frac{\mathbb{E}_w W(X; u, w) \, I_{\{S(X) \geq \gamma\}} X}{\mathbb{E}_w I_{\{S(X) \geq \gamma\}} W(X; u, w)} \ , \tag{29}$$

for any reference parameter w.

The estimator $\widehat{v}$ of v^* in (29) can be obtained analytically from the solution of the stochastic program (23), that is,

$$\widehat{v} = \frac{\sum_{i=1}^N I_{\{S(X_i)\geq\gamma\}} W(X_i;u,w) X_i}{\sum_{i=1}^N I_{\{S(X_i)\geq\gamma\}} W(X_i;u,w)} \tag{30}$$

where $X_1, \ldots, X_N$ is a random sample from the density $f(\cdot;w)$.

A similar explicit formula can be found for the case where $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a vector of independent random variables such that each component X_j belongs to a NEF parameterized by the mean. In particular, if $\boldsymbol{u} = (u_1, \ldots, u_n)$ is the nominal reference parameter, then for each $j = 1, \ldots, n$ the density of X_j is given by

$$f_j(x;u_j) = \mathrm{e}^{x\eta(u_j)-\zeta(\eta(u_j))}\, h_j(x).$$

It is not difficult to see that under independence assumption the problem (23) becomes "separable", that is, it reduces to n subproblems. Thus, the optimal reference parameter vector $\boldsymbol{v}^* = (v_1^*, \ldots, v_n^*)$ is given by

$$v_j^* \;=\; \frac{\mathbb{E}_{\boldsymbol{u}} I_{\{S(\boldsymbol{X})\geq\gamma\}}\, X_j}{\mathbb{E}_{\boldsymbol{u}}\, I_{\{S(\boldsymbol{X})\geq\gamma\}}} = \frac{\mathbb{E}_{\boldsymbol{w}} I_{\{S(\boldsymbol{X})\geq\gamma\}}\, W(\boldsymbol{X};\boldsymbol{u},\boldsymbol{w})\, X_j}{\mathbb{E}_{\boldsymbol{w}} I_{\{S(\boldsymbol{X})\geq\gamma\}}\, W(\boldsymbol{X};\boldsymbol{u},\boldsymbol{w})}\,. \tag{31}$$

Moreover, we can estimate the jth component of $\boldsymbol{v}^*$ as

$$\widehat{v}_j \;=\; \frac{\sum_{i=1}^N I_{\{S(\boldsymbol{X})\geq\gamma\}} W(\boldsymbol{X}_i;\boldsymbol{u},\boldsymbol{w}) X_{ij}}{\sum_{i=1}^N I_{\{S(\boldsymbol{X})\geq\gamma\}} W(\boldsymbol{X}_i;\boldsymbol{u},\boldsymbol{w})}, \tag{32}$$

where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ is a random sample from the density $f(\cdot;\boldsymbol{w})$, and X_{ij} is the jth component of $\boldsymbol{X}_i$.

2.1 Examples

For better insight we present now two examples with both light and heavy tails while using Algorithm 2.1 with the standard likelihood ratio. Although the quantities of interest can be computed analytically, we present them to illustrate the Algorithm 2.1. It is important to realize that in both examples we obtain the optimal reference parameter for the SLR estimator via the cross-entropy optimization, via explicit formulas such as (29). On the other hand, in order to study the complexity properties of the SLR estimator we derive the SCV of the estimator via formulas of type (28) for exponential families.

Example 2.2 Suppose we are interested in estimating $\ell = \ell(\gamma) = \mathbb{P}(S(\boldsymbol{X}) \geq \gamma)$, where

$$S(\boldsymbol{X}) = \min(X_1, \ldots, X_n) \tag{33}$$

and the random variables $X_1, \ldots, X_n$ are exponentially identically distributed with mean u; thus each X_i has density $f(\cdot; u) = u^{-1} \exp(-xu^{-1}), x \geq 0$. Obviously,

$$\ell = \prod_{i=1}^{n} \mathbb{P}(X_i \geq \gamma) = \mathrm{e}^{-n\gamma u^{-1}} . \tag{34}$$

For large γ, the squared coefficient of variation (SCV) of the crude Monte Carlo (CMC) estimator (see (2)) is

$$\kappa^2(\gamma) \approx \frac{1}{N} \mathrm{e}^{n\gamma u^{-1}} .$$

Hence the CMC estimator has *exponential* complexity in γ. It is easy to verify from (28) that for i.i.d. and exponentially distributed random variables X_i, we have that

$$\mathbb{E}_u I_{\{S(\boldsymbol{X}) \geq \gamma\}} W(\boldsymbol{X}; u, v) = \left(\frac{v^2}{u(2v-u)} \right)^n \mathbb{E}_{\frac{uv}{2v-u}} I_{\{S(\boldsymbol{X}) \geq \gamma\}} . \tag{35}$$

It follows that the variance of the estimator (25) is

$$\begin{aligned}
\mathrm{Var}(\widehat{\ell}) &= \frac{1}{N} \left\{ \mathbb{E}_u I_{\{\min(X_1,\ldots,X_n) \geq \gamma\}} W(\boldsymbol{X}; u, v) - \ell^2 \right\} \\
&= \frac{1}{N} \left\{ \left(\frac{v^2}{u(2v-u)} \right)^n \mathbb{E}_{\frac{uv}{2v-u}} I_{\{\min(X_1,\ldots,X_n) \geq \gamma\}} - \ell^2 \right\} \\
&= \frac{1}{N} \left\{ \left(\frac{v^2}{u(2v-u)} \right)^n \mathrm{e}^{-n\gamma(2u^{-1}-v^{-1})} - \ell^2 \right\} \\
&= \frac{1}{N} \left\{ \left(\frac{v^2 \mathrm{e}^{\gamma v^{-1})}}{u(2v-u)} \right)^n \ell^2 - \ell^2 \right\} .
\end{aligned}$$

Consequently, the SCV of $\widehat{\ell}$ is given by

$$\kappa^2(v, \gamma) = \frac{1}{N} \left\{ \left(\frac{v^2 \mathrm{e}^{\gamma v^{-1})}}{u(2v-u)} \right)^n - 1 \right\}$$

Since the exponential distribution belongs to a NEF which is parameterized by the mean, we can apply formula (29) directly and obtain that the optimal reference parameter is given by

$$v^* = u + \gamma .$$

For large $\gamma \gg u$ we have that $v^* \approx \gamma$, and the SCV becomes

$$\kappa^2(\gamma) \approx \frac{1}{N} \gamma^n \mathrm{e}^n (2u)^{-n}, \tag{36}$$

where N is the sample size. That is, for large γ, the SCV $\kappa^2(\gamma)$ of the CMC and of the SLR estimators (with the CE optimal parameter $v^* \approx \gamma$) increase in γ exponentially and polynomially, respectively. In other words, the CMC and the SLR estimators can be viewed as *exponential* and *polynomial* ones.

Example 2.3 (Heavy tails) As mentioned earlier, unlike the ECM the SLR estimate (25) is not limited to light-tail distributions but can also be applied to *heavy*-tail distributions. To illustrate this, we generalize Example 2.2 for $n = 1$ to the Weibull case. Specifically, consider the estimation of $\ell = \mathbb{P}(X \geq \gamma)$ with $X \sim \mathsf{Weib}(a, u^{-1})$, that is, X has density

$$f(x; u) = a\, u^{-1} \left(u^{-1} x\right)^{a-1} \mathrm{e}^{-\left(u^{-1} x\right)^{a}}, \quad x > 0\,. \tag{37}$$

To estimate ℓ via the CE method we shall use the family of distributions $\{\mathsf{Weib}(a,\ v^{-1}), v > 0\}$, where a is kept fixed. Note that for $a = 1$ we have the exponential class of distributions.

Using the CE approach, we find the optimal CE reference parameter by solving

$$\max_v D(v) = \max_v \int_\gamma^\infty f(x; u) \ln f(x; v)\, dx,$$

or, equivalently, by solving

$$\int_\gamma^\infty f(x; u) \frac{d}{dv} \ln f(x; v)\, dx = 0\,. \tag{38}$$

Substituting (37) into (38) yields the following simple expression for the optimal CE reference parameter v^*:

$$v^* = (u^a + \gamma^a)^{1/a}\,. \tag{39}$$

This is true for *any* $a > 0$. Note that $\{\mathsf{Weib}(a, v^{-1}), v > 0\}$ is an exponential family of the form (26), with $t(x) = x^a$, $\eta = -v^{-a}$, $c(\eta) = -\eta$ and $h(x) = a x^{a-1}$. So we can obtain (39) also via (27) as the solution to

$$\mathbb{E}_\theta I_{\{X \geq \gamma\}} \left\{ \frac{1}{\eta} + X^a \right\} = 0\,, \tag{40}$$

with $\theta = -u^{-a}$.

Similar to Example 2.2 and (28) the variance of the SLR estimator $\widehat{\ell}$ for any reference parameter $\boldsymbol{v}$ is found (after some algebra) to be

$$\begin{aligned}
\mathrm{Var}(\widehat{\ell}) &= \frac{1}{N} \left\{ \mathbb{E}_u I_{\{X \geq \gamma\}} W(X; u, v) - \ell^2 \right\} \\
&= \frac{1}{N} \left\{ \frac{\mathrm{e}^{(\gamma/v)^a}}{(u/v)^a [2 - (u/v)^a]}\, \ell^2 - \ell^2 \right\}\,,
\end{aligned}$$

where we have used the fact that $\ell = \mathrm{e}^{-(\gamma/u)^a}$. If we substitute v above with v^* and divide by ℓ^2, we find that the SCV κ^2 of $\widehat{\ell}$ is given by

$$\frac{1}{N} \left\{ \frac{\exp\left(\frac{(\gamma/u)^a}{1 + (\gamma/u)^a} \right) \left(1 + (\gamma/u)^a\right)^2}{2(\gamma/u)^a + 1} \right\}\,.$$

It follows that for large γ/u

$$\kappa^2 \approx \frac{1}{N}\,\frac{\mathrm{e}}{2}\left(\frac{\gamma}{u}\right)^a \; .$$

In other words, the SLR estimator $\widehat{\ell}$ has *polynomial* complexity in γ, for any $a > 0$, including the heavy-tail case $0 < a < 1$. It is a common misunderstanding that IS only works for light-tail distributions. In this example we saw that polynomial complexity can be easily obtained by using the CE method. But we can do even better! In Section 3 we will see how with the TLR method we can in fact achieve an SLR estimator with *bounded relative error*, meaning that the κ^2 is bounded by c/N for some constant c which does not depend on γ.

Remark 2.1 Consider (24). Assume that the X_i's are independent and $X_i \sim \mathsf{Weib}(a_i, u_i^{-1})$, $i = 1, \ldots, n$. It is readily seen that for fixed a_i, $i = 1, \ldots, n$, program (24) can be solved analytically, and the components of $\widehat{\boldsymbol{v}} = (\widehat{v}_1, \ldots, \widehat{v}_n)$ in Weibull pdf can be updated as

$$\widehat{v}_{t,j} = \left(\frac{\sum_{k=1}^{N} I_{\{S(\boldsymbol{X}_k) \geq \widehat{\gamma}_t\}} W(\boldsymbol{X}_k; u, \widehat{v}_{t-1,j}) \boldsymbol{X}_k^a}{\sum_{k=1}^{N} I_{\{S(\boldsymbol{X}_k) \geq \widehat{\gamma}_t\}} W(\boldsymbol{X}_k; u, \widehat{v}_{t-1,j})} \right)^{1/a} . \tag{41}$$

A different parameterization of the Weibull distribution gives an even simpler formula. Namely, if we use the change of measure

$$X_n \sim \mathsf{Weib}(a, u^{-1/a}) \longrightarrow \mathsf{Weib}(a, v^{-1/a}) \; , \quad v \geq u \; ,$$

thus,

$$f(x; v) = a\, v^{-1}\, x^{a-1}\, \mathrm{e}^{-v^{-1} x^a} \; .$$

Then the v-parameters are updated as

$$\widehat{v}_{t,j} = \frac{\sum_{k=1}^{N} I_{\{S(\boldsymbol{X}_k) \geq \widehat{\gamma}_t\}} W(\boldsymbol{X}_k; u, \widehat{v}_{t-1,j}) \boldsymbol{X}_k^a}{\sum_{k=1}^{N} I_{\{S(\boldsymbol{X}_k) \geq \widehat{\gamma}_t\}} W(\boldsymbol{X}_k; u, \widehat{v}_{t-1,j})} . \tag{42}$$

Remark 2.2 (Two-parameter update) For the Weibull distribution it is not difficult to formulate a two-parameter updating procedure in which both scale and shape parameter are updated. Specifically, consider the change of measure

$$X_i \sim \mathsf{Weib}(a_i, u_i^{-1/a_i}) \longrightarrow \mathsf{Weib}(b_i, v_i^{-1/b_i}), \quad v_i > 0, \; b_i > 0 \; .$$

The updating formula for the v_i is given in (42), but an analytic updating of the parameter vector $\boldsymbol{b} = (b_1, \ldots, b_n)$ is not available from (23). However, the gradient of $\widehat{D}(\boldsymbol{b}, \boldsymbol{v})$ with respect to $\boldsymbol{b}$ can be easily obtained from the gradient $\nabla_{\boldsymbol{b}} \ln f(\boldsymbol{X}; \boldsymbol{b}, \boldsymbol{v})$. It is readily seen that the ith component of $\nabla_{\boldsymbol{b}} \ln f(\boldsymbol{X}; \boldsymbol{b}, \boldsymbol{v})$

for the random vector $\boldsymbol{X}$ with independent components $X_i \sim \mathsf{Weib}(b_i, \widehat{v}_i^{-1/b_i})$, $i = 1, \dots, n$ equals

$$b_i^{-1} + \ln X_i - \frac{X_i^{b_i}}{\widehat{v}_i} \ln X_i. \tag{43}$$

Consequently, the i-th component of $\boldsymbol{b}$ can be obtained from the numerical solution of the following nonlinear equation

$$\frac{1}{N} \sum_{k=1}^{N} I_k W_k (b_i^{-1} + \ln X_{ki} - \frac{X_{ki}^{b_i}}{\widehat{v}_i} \ln X_{ki}) = 0. \tag{44}$$

Substituting $\widehat{v}_i$ from (42), into (44) we obtain

$$b_i^{-1} + \frac{\sum_{k=1}^{N} I_k W_k \ln X_{ki}}{\sum_{k=1}^{N} I_k W_k} - \frac{\sum_{k=1}^{N} I_k W_k X_{ki}^{b_i} \ln X_{ki}}{\sum_{k=1}^{N} I_k W_k X_{ki}^{b_i}} = 0. \tag{45}$$

One might solve (45) using the bisection method, say.

Remark 2.3 (Hazard rate twisting) It is interesting to note that hazard rate twisting [19] often amounts to SLR. In hazard rate twisting the change of measure for some distribution with pdf f (with support in $\mathbb{R}_+$) and tail distribution function $\bar{F}$ is such that the *hazard rate* (or failure rate) $\lambda(x) = f(x)/\bar{F}(x)$ is changed to $(1-\theta)\lambda(x)$, for some $0 \leq \theta < 1$. The pdf of the changed measure is now

$$f_\theta(x) = \lambda(x)(1-\theta)\,\mathrm{e}^{-(1-\theta)\Lambda(x)},$$

where $\Lambda(x) = \int_0^x \lambda(y)\,dy$. In particular, for the $\mathsf{Weib}(a, u^{-1})$ distribution we have $\lambda(x) = au^{-1}(u^{-1}x)^a$ and $\Lambda(x) = (u^{-1}x)^a$, so that

$$f_\theta(x) = (1-\theta)au^{-1}(u^{-1}x)^{a-1}\mathrm{e}^{-(1-\theta)(u^{-1}x)^a},$$

which corresponds to the SLR change of measure $\mathsf{Weib}(a, u^{-1}) \longrightarrow \mathsf{Weib}(a, v^{-1})$, with $v^{-1} = (1-\theta)^{1/a}u^{-1}$. Similarly, for the $\mathsf{Pareto}(a, u^{-1})$ distribution, with $\bar{F}(x) = (1 + x/u)^{-(a+1)}$, we have $\lambda(x) = au^{-1}(1 + u^{-1}x)^{-1}$ and $\Lambda(x) = a\ln(1 + u^{-1}x)$, so that

$$f_\theta(x) = (1-\theta)au^{-1}(1 + u^{-1}x)^{-((1-\theta)a+1)},$$

so that hazard rate twisting with parameter θ corresponds to the SLR change of measure $\mathsf{Pareto}(a, u^{-1}) \longrightarrow \mathsf{Pareto}(b, u^{-1})$ with $b = (1-\theta)a$. Note that in the Weibull case the the scale parameter u^{-1} is changed whereas in the second case the shape parameter a is changed.

3 The TLR Method

In this section we present the *transform likelihood ratio* (TLR) method as a simple, convenient and unifying way of constructing efficient IS estimators that are applicable for both light- and heavy-tailed distributions.

Let $\boldsymbol{X}$ be a random vector. Suppose we wish to estimate

$$\ell = \mathbb{E} I_{\{S(\boldsymbol{X}) \geq \gamma\}} \ .$$

The TLR method comprises two steps. The first is a simple *change of variable* step. That is, we write $\boldsymbol{X}$ as a function of another random vector $\boldsymbol{Z}$, for example

$$\boldsymbol{X} = H(\boldsymbol{Z}) \ . \tag{46}$$

If we define

$$\widetilde{S}(\boldsymbol{Z}) = S(H(\boldsymbol{Z})),$$

then

$$\ell = \mathbb{E} I_{\{\widetilde{S}(\boldsymbol{Z}) \geq \gamma\}} \ .$$

Suppose $\boldsymbol{Z}$ has density $h(\cdot;\boldsymbol{\theta})$ in some class of densities $\{h(\cdot;\boldsymbol{\eta})\}$. Then we can seek to estimate ℓ efficiently via IS using either the SLR method (staying in the same parametric class) or ECM. The parameter updating can again be done via the CE method. In particular, when using the SLR method we obtain in analogy to (25) the estimator

$$\widehat{\ell} = \frac{1}{N} \sum_{i=1}^{N} I_{\{\widetilde{S}(\boldsymbol{Z}_i;\boldsymbol{\theta}) \geq \gamma\}} \widetilde{W}(\boldsymbol{Z}_i;\boldsymbol{\theta},\boldsymbol{\eta}), \tag{47}$$

where

$$\widetilde{W}(\boldsymbol{Z}_i;\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{h(\boldsymbol{Z}_i;\boldsymbol{\theta})}{h(\boldsymbol{Z}_i;\boldsymbol{\eta})}$$

and $\boldsymbol{Z}_i \sim h(\boldsymbol{z};\boldsymbol{\eta})$. We shall call the SLR estimate (47) based on the transformation (46), the *transform LR* (TLR) estimate.

To find the optimal parameter vector $\boldsymbol{\eta}^*$ of the TLR estimator (47) we can solve in analogy to (23) the following CE program

$$\max_{\boldsymbol{\eta}} D(\boldsymbol{\eta}) = \max_{\boldsymbol{\eta}} \mathbb{E}_{\boldsymbol{\eta}_{t-1}} I_{\{\widetilde{S}(\boldsymbol{Z};\boldsymbol{\eta}_{t-1}) \geq \gamma_t\}} \widetilde{W}(\boldsymbol{Z};\boldsymbol{\theta},\boldsymbol{\eta}_{t-1}) \ln h(\boldsymbol{Z};\boldsymbol{\eta}) \tag{48}$$

and similarly for the stochastic counterpart of (48). For example, $h(\boldsymbol{z};\boldsymbol{\theta})$ might be any light tail NEF pdf, (and thus, the optimal reference parameter vector $\boldsymbol{\eta}^*$ could be obtained analytically from the stochastic version (counterpart) of (48)), or $h(\boldsymbol{z};\boldsymbol{\theta})$ might be a truncated version of the original pdf $f(\boldsymbol{x})$, denoted as $\widetilde{f}(\boldsymbol{x};c)$, where the truncation parameter c could be controllable as well.

It is crucial to understand that in contrast to the SLR estimate (25), its TLR counterpart (47) involves an additional stage, namely it uses the transformation stage (46). As result, the TLR estimate (47) presents a *three-stage* procedure rather then on a *two-stage* one (see (25)). Note that the three-stages of TLR are associated with

1. Transformation from the original pdf f to an auxiliary one h.

2. Updating the parameter vector $\boldsymbol{\eta}$ (at each iteration of Algorithm 2.1) using the stochastic counterpart of (48).

3. Estimating ℓ according to (47) with $\boldsymbol{\eta}$ replaced by $\widehat{\boldsymbol{\eta}}^*$, which presents the solution obtained from Algorithm 2.1 at stage two.

the transformation stage (46) an exponential pdf is used.

Example 3.1 (Inverse Transform Likelihood Ratio)
Consider the single-dimensional case. According to the *inverse transform* (IT) method a random variable $X \sim F(x)$ can be written as

$$X = F^{-1}(Z), \tag{49}$$

where $Z \sim \mathsf{U}(0,1)$ and F^{-1} is the inverse of the cdf F.

Let $h(\cdot;\nu)$ be another density on $(0,1)$ dominating the uniform density, and parameterized by some reference parameter ν. An example is the $\mathsf{Beta}(\nu,1)$-distribution, with density

$$h(z;\nu) = \nu\, z^{\nu-1}, \quad z \in (0,1)\ ,$$

with $\nu > 0$ or the $\mathsf{Beta}(1,\nu)$-distribution, with density

$$h(z;\nu) = \nu\,(1-z)^{\nu-1}, \quad z \in (0,1)\ .$$

The TLR estimator is given by

$$\widehat{\ell} = N^{-1} \sum_{i=1}^{N} I_{\{\widetilde{S}(Z_i) \geq \gamma\}} \widetilde{W}(Z_i;\nu)\ , \tag{50}$$

where $Z_1, \dots, Z_N$ is a random sample from $h(\cdot;\nu)$ and

$$\widetilde{W}(Z;\nu) = \frac{1}{h(Z;\nu)} \tag{51}$$

is the LR. We call (50) the *inverse transform - likelihood ratio* (ITLR) estimator [22].

Consider next the multivariate case where the components of $\boldsymbol{X} = (X_1, \dots, X_n)$ are independent and $X_i \sim F(\cdot;u_i)$ for a fixed parameter vector $\boldsymbol{u} = (u_1, \dots, u_n)$. In analogy with the univariate case we wish to estimate, for some performance function S,

$$\ell = \mathbb{E} I_{\{S(\boldsymbol{X}) \geq \gamma\}} = \mathbb{E} I_{\{\widetilde{S}(\boldsymbol{Z}) \geq \gamma\}}\ ,$$

where $\widetilde{S}(\boldsymbol{Z}) = S(F^{-1}(Z_1;u_1), \dots, F^{-1}(Z_n;u_n))$, $\boldsymbol{Z} = (Z_1, \dots, Z_n)$, and Z_j, $j = 1, \dots, n$ are i.i.d. and uniformly distributed on (0,1).

Let $h(\cdot;\boldsymbol{\nu})$ be another density on $(0,1)^n$ dominating the uniform density, and parameterized by some reference parameter vector $\boldsymbol{\nu}$. For example, we could choose h such that the Z_i's are independent with a $\mathsf{Beta}(1, \nu_i)$-distribution, in which case

$$h(\boldsymbol{z}; \boldsymbol{\nu}) = \prod_{i=1}^{n} \nu_i \, (1 - z_i)^{\nu_i - 1}, \quad \boldsymbol{z} \in (0,1)^n, \tag{52}$$

with $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$. As in the univariate case we have the ITLR estimator

$$\widehat{\ell} = N^{-1} \sum_{i=1}^{N} I_{\{\widetilde{S}(\boldsymbol{Z}_i) \geq \gamma\}} \widetilde{W}(\boldsymbol{Z}_i; \boldsymbol{\nu}) \ , \tag{53}$$

respectively, where $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ is a random sample from $h(\cdot; \boldsymbol{\nu})$ and

$$\widetilde{W}(\boldsymbol{Z}; \boldsymbol{\nu}) = \frac{1}{h(\boldsymbol{Z}; \boldsymbol{\nu})} \ . \tag{54}$$

Note that Algorithm 2.1 remains the same for the ITLR approach, provided the CE programs (23) and (24) are replaced by

$$\max_{\boldsymbol{\nu}} D(\boldsymbol{\nu}) = \max_{\boldsymbol{\nu}} \mathbb{E}_{\boldsymbol{\nu}_{t-1}} I_{\{\widetilde{S}(\boldsymbol{Z}) \geq \gamma_t\}} \widetilde{W}(\boldsymbol{Z}; \boldsymbol{\nu}_{t-1}) \ln h(\boldsymbol{Z}; \boldsymbol{\nu}), \tag{55}$$

and

$$\max_{\boldsymbol{\nu}} \widehat{D}(\boldsymbol{\nu}) = \max_{\boldsymbol{\nu}} \frac{1}{N} \sum_{i=1}^{N} I_{\{\widetilde{S}(\boldsymbol{Z}_i) \geq \widehat{\gamma}_t\}} \widetilde{W}(\boldsymbol{Z}_i; \widehat{\boldsymbol{\nu}}_{t-1}) \ln h(\boldsymbol{Z}_i; \boldsymbol{\nu}), \tag{56}$$

respectively, where $\boldsymbol{Z}_i \sim h(\cdot; \widehat{\boldsymbol{\nu}}_{t-1})$.

In particular, for the case (52) where the Z_i's are independent and $Z_i \sim \mathsf{Beta}(1, \nu_i)$, $i = 1, \ldots, n$ (56) can be solved analytically, and it is not difficult to see that the components of $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$ are updated as

$$\widehat{\nu}_{t,j} = - \frac{\displaystyle\sum_{i=1}^{N} I_{\{\widetilde{S}(\boldsymbol{Z}_i) \geq \widehat{\gamma}_t\}} \widetilde{W}(\boldsymbol{Z}_i; \widehat{\boldsymbol{\nu}}_{t-1})}{\displaystyle\sum_{i=1}^{N} I_{\{\widetilde{S}(\boldsymbol{Z}_i) \geq \widehat{\gamma}_t\}} \widetilde{W}(\boldsymbol{Z}_i; \widehat{\boldsymbol{\nu}}_{t-1}) \ln(1 - Z_{ij})} \ , \tag{57}$$

where Z_{ij} is the j-th component of $\boldsymbol{Z}_i$.

The following example shows that (I)TLR can lead to a more efficient estimator than the SLR method.

Example 3.2 (Example 2.2 continued) Suppose, as in Example 2.2, that we are interested in estimating $\ell = \mathbb{P}(S(\boldsymbol{X}) \geq \gamma)$, where

$$S(\boldsymbol{X}) = \min(X_1, \ldots, X_n), \quad X_1, \ldots, X_n \ \sim \ \mathsf{Exp}(u^{-1}) \ . \tag{58}$$

In this case we can write

$$X_i = -u \, \ln(1 - Z_i) \ , \ i = 1, \ldots, n \ , \tag{59}$$

where $Z_i \sim \mathsf{U}(0,1),\ i = 1, \dots, n$ and $Z_1, \dots, Z_n$ independent. We have

$$\widetilde{S}(\boldsymbol{Z}) = \min_i(-u \ln(1 - Z_i)) = -u \ln(1 - \min_i Z_i) ,$$

so that

$$\ell = \mathbb{P}(\widetilde{S}(\boldsymbol{Z}) \geq \gamma) = \mathbb{P}(\min_i Z_i \geq 1 - \eta),$$

with $\eta = \mathrm{e}^{-\gamma u^{-1}}$.

Let $h(\boldsymbol{z}; \nu) = \prod_{i=1}^n \nu z_i^{\nu-1},\ \nu > 0$ be the dominating density on $\mathsf{U}^n(0,1)$ for $\boldsymbol{Z}$. Note that (by symmetry) we choose all component pdfs the *same*, this in contrast to (52). To find the optimal ν we need to solve the CE program (55), which for this case reduces to

$$\max_{\nu>0} D(\nu) = \max_{\nu>0} \mathbb{E} I_{\{\widetilde{S}(\boldsymbol{Z}) \geq 1-\eta\}} \sum_{i=1}^n (\ln \nu + (\nu - 1)\ \ln Z_i)$$

Equating the gradient with respect to ν to 0 gives

$$\begin{aligned}\nu^* &= -\frac{n \mathbb{E} I_{\{\widetilde{S}(\boldsymbol{Z}) \geq 1-\eta\}}}{\mathbb{E} I_{\{\widetilde{S}(\boldsymbol{Z}) \geq 1-\eta\}} \ln Z_i} \\ &= -\frac{n \eta^n}{n \eta^{n-1} \int_{\eta-1}^1 \ln z \, dz} = \frac{\eta}{\ln(1-\eta)(1-\eta) + \eta} .\end{aligned}$$

It follows that for small η we have

$$\nu^* \approx \frac{2}{\eta} . \tag{60}$$

To find the asymptotic SCV κ^2 we need to find fist the variance of the ITLR estimator $\widehat{\ell}$. Let $V(\nu)$ be the second moment of $I_{\{\widetilde{S}(\boldsymbol{Z}) \geq \gamma\}} \widetilde{W}(\boldsymbol{Z}; 1, \nu)$. We have

$$\begin{aligned} V(\nu) &= \mathbb{E}_\nu \left\{ \left(\prod_{i=1}^n I_{\{Z_i \geq 1-\eta\}} \right)^2 \cdot \left(\frac{1}{h(\boldsymbol{Z}; \nu)} \right)^2 \right\} \\ &= \left(\mathbb{E}_\nu \left\{ I_{\{Z \geq 1-\eta\}} \left(\nu Z^{\nu-1} \right)^{-2} \right\} \right)^n \\ &= \left(\frac{1}{\nu} \int_{1-\eta}^{1} z^{1-\nu} \mathrm{d}z \right)^n \\ &= \left(\frac{(1 - (1-\eta)^{2-\nu})}{\nu(2-\nu)} \right)^n . \end{aligned} \tag{61}$$

From (61) and (60) we have for small η

$$V(\nu^*) \approx \left\{ \frac{\eta 2^{-1}}{2 - 2\eta^{-1}} [1 - (1-\eta)^{2-2\eta^{-1}}] \right\}^n .$$

So that, for small η

$$V(\nu^*) \approx \frac{(e^2-1)^n}{4n}\eta^{2n} .$$

For ℓ we have

$$\ell = \prod_{i=1}^{n} \mathbb{P}(Z_i \geq 1-\eta) = \left(\int\limits_{1-\eta}^{1} 1 dz \right)^n = \eta^n ,$$

Finally,

$$N \times \kappa^2 = \frac{V(\nu^*)}{\ell^2} - 1 \approx \left(\frac{e^2-1}{4} \right)^n - 1. \qquad (62)$$

Note that κ^2 in (62) does not depend on η and therefore neither on γ. Consequently, the corresponding estimators are of *bounded relative error* in γ. Comparing (62) with $N \times \kappa^2 = \gamma^n \mathrm{e}^n (2u)^{-n} = (-\ln \eta)^n (\mathrm{e}/2)^n$ in (36), it readily follows that the former (ITLR) is much faster than the latter (SLR), especially when γ is large.

The following proposition illustrates the usefulness of ITLR for estimating small probabilities, for *any* distribution. In the results below the uni-variate ITLR method is used with a $\mathsf{Beta}(\nu,1)$ change of measure. It is important to realize that this CM may not be appropriate for similar problems concerning multi-variate random variables. Indeed the $\mathsf{Beta}(\nu,1)$ CM may give exponential complexity, whereas a $\mathsf{Beta}(1,\nu)$ CM could give polynomial complexity.

Proposition 3.1 Let X be distributed as $L(1-Z)$, with $Z \sim \mathsf{U}(0,1)$, for some monotone increasing function L on $(0,1)$. Then, estimating $\ell = \mathbb{P}(X \geq \gamma)$ via ITLR using the $\{\mathsf{Beta}(\nu,1), \nu > 0\}$ family of distributions gives an LR estimator with bounded relative error.

Proof. The proof uses similar arguments to the ones used in Example 3.2. First, we write $\ell = \mathbb{P}(X \geq \gamma)$ as $\ell = \mathbb{P}(Z \geq 1-\eta)$, with $\eta = L^{-1}(\gamma)$. Hence, if we estimate ℓ via the IS density

$$h(z;\nu) = \nu z^{\nu-1}, \qquad (63)$$

then the optimal CE parameter is given, analogously to (60), by

$$\nu^* = \frac{\eta}{\eta + (1-\eta)\ln(1-\eta)} \approx \frac{2}{\eta},$$

as $\eta \to 0$. Moreover, the corresponding SCV satisfies

$$N \times \kappa^2 \approx \frac{e^2-1}{4} - 1 \approx 0.597264 . \qquad (64)$$

Note that this is independent of η (and hence γ). Thus, the estimator is of *bounded relative error*.

Example 3.3 (Example 2.3 continued) Let $X \sim \mathsf{Weib}(a, u^{-1})$. That is, X has cdf F given by

$$F(x) = 1 - \mathrm{e}^{-(u^{-1}x)^a}, \quad x \geq 0 .$$

We wish to estimate $\ell = \mathbb{P}(X \geq \gamma) = \mathrm{e}^{-(u^{-1}\gamma)^a}$ for large γ. Using

$$L(z) = u\,(-\ln z)^{\frac{1}{a}}, \quad z \in (0,1),$$

we can write $\ell = \mathbb{P}(Z \geq 1 - \eta)$, with $Z \sim \mathsf{U}(0,1)$ and $\eta = \mathrm{e}^{-(u^{-1}\gamma)^a}$. Hence, by Proposition 3.1 we can efficiently estimate ℓ via ITLR using the $\mathsf{Beta}(\nu, 1)$ density, yielding an SLR estimator with bounded relative error given in (64). Note that this is true for *any* shape parameter $a > 0$, including the *heavy-tail* case $0 < a < 1$.

4 Equivalence between SLR and TLR

As we have seen the TLR method can be viewed as a generalization of the SLR method, involving an additional transformation step. In this section we illustrate that seemingly different implementations of SLR and (I)TLR may in fact be completely equivalent.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\mathsf{Weib}(a, u^{-1})$ distributed and consider the estimation of a general rare event probability

$$\ell = \mathbb{P}(S(\boldsymbol{X}) \geq \gamma)$$

for large γ using importance sampling. We consider three methods.

(1) SLR with $\mathsf{Weib}(a, v^{-1})$ twisting, fixed a

The first method is a straightforward change of the Weibull scale parameter, as in Example 2.3. In particular, we consider the change of measure

$$X_n \sim \mathsf{Weib}(a, u^{-1}) \longrightarrow \mathsf{Weib}(a, v^{-1}), \quad v \geq u .$$

Note that the problem is of the form discussed in Remark 2.1; but by symmetry we know that the components of the reference vector must be equal. This leads to slightly different updating formulas, namely:

$$\widehat{v}_t = \left(\frac{\sum_{k=1}^N I_{\{S(\boldsymbol{X}_k) \geq \widehat{\gamma}_t\}} W(\boldsymbol{X}_k; u, \widehat{v}_{t-1})\, n^{-1} \sum_{i=1}^n X_{ki}^a}{\sum_{k=1}^N I_{\{S(\boldsymbol{X}_k) \geq \widehat{\gamma}_t\}} W(\boldsymbol{X}_k; u, \widehat{v}_{t-1})} \right)^{1/a} . \tag{65}$$

(2) ITLR with $\mathsf{Beta}(1, \nu)$ twisting

In the second method we estimate ℓ via the ITLR method. First, write $X_i \sim \mathsf{Weib}(a, u^{-1})$ as

$$X_i = u\,(-\ln(1 - Z_i))^{1/a},$$

with the Z_i i.i.d. $\mathsf{U}(0,1) = \mathsf{Beta}(1,1)$. We now apply a change of measure on the distribution of Z_i:

$$Z_i \sim \mathsf{Beta}(1,1) \longrightarrow \mathsf{Beta}(1,\nu) \quad 0 < \nu \leq 1 \; .$$

Define $\widetilde{S}(\boldsymbol{Z}) = S(\boldsymbol{X})$. The CE updating formula is, similar to (57),

$$\widehat{\nu}_t = -\frac{\sum_{i=1}^{N} I_{\{\widetilde{S}(\boldsymbol{Z}_i) \geq \widehat{\gamma}_t\}} \widetilde{W}(\boldsymbol{Z}_i; 1, \widehat{\nu}_{t-1})}{\sum_{i=1}^{N} I_{\{\widetilde{S}(\boldsymbol{Z}_i) \geq \widehat{\gamma}_t\}} \widetilde{W}(\boldsymbol{Z}_i; 1, \widehat{\nu}_{t-1})\, n^{-1} \sum_{j=1}^{n} \ln(1 - Z_{ij})} \, , \tag{66}$$

where Z_{ij} is the j-th component of $\boldsymbol{Z}_i$.

It is interesting to compare the present ITLR method with the previous Weibull change of measure. Since, Z_i can be written as $Z_i = 1 - (1 - U_i)^{1/\nu}$, with $U_i \sim \mathsf{U}(0,1)$, we have

$$\begin{aligned} X_i &= u \left(- \ln \left(\{1 - U_i\}^{1/\nu} \right) \right)^{1/a} \\ &= \frac{u}{\nu^{1/a}} \left(- \ln (1 - U_i) \right)^{1/a} , \end{aligned}$$

so that under the change of measure $Z_i \sim \mathsf{Beta}(1,1) \longrightarrow \mathsf{Beta}(1,\nu)$ we have that $X_i \sim \mathsf{Weib}(a, u^{-1}\nu^{1/a})$. Let us compare the behavior of the SLR and ITLR estimators for $v = u\nu^{-1/a}$. First of all, observe that

$$\begin{aligned} W(\boldsymbol{X}; u, v) &= \prod_{i=1}^{n} \frac{a u^{-1} (u^{-1} X_i)^{a-1} \mathrm{e}^{-(u^{-1} X_i)^a}}{a v^{-1} (v^{-1} X_i)^{a-1} \mathrm{e}^{-(v^{-1} X_i)^a}} \\ &= \prod_{i=1}^{n} \frac{1}{\nu\, (1 - Z_i)^{\nu - 1}} = \widetilde{W}(\boldsymbol{Z}; 1, \nu) \; . \end{aligned}$$

This shows that

$$\sum_{i=1}^{N} I_{\{S(\boldsymbol{X}_i) \geq \gamma\}} W(\boldsymbol{X}_i; u, v) = \sum_{i=1}^{N} I_{\{\widetilde{S}(\boldsymbol{Z}_i) \geq \gamma\}} \widetilde{W}(\boldsymbol{Z}_i; 1, \nu) \; .$$

In other words, the SLR estimator is *identical* to the ITLR estimator, provided we take $v = u\,\nu^{-1/a}$. Note also that, in the same way, the CE updating formulas and their deterministic counterparts are equivalent, in the sense that $\widehat{v}_t = u(\widehat{\nu}_t)^{-1/a}$ and $v_t = u(\nu_t)^{-1/a}$.

(3) TLR with $\mathsf{Exp}(\lambda)$ twisting

Let us finally apply the TLR method with an "exponential change of measure". We now write $X_i \sim \mathsf{Weib}(a, u^{-1})$ as

$$X_i = u Z_i^{1/a},$$

with the Z_i i.i.d. $\mathsf{Exp}(1)$, and apply the change of measure

$$Z_i \sim \mathsf{Exp}(1) \longrightarrow \mathsf{Exp}(\lambda), \quad 0 < \lambda \leq 1 \; .$$

With $\widetilde{S}(\boldsymbol{Z}) = S(\boldsymbol{X})$ the CE updating formula is given by

$$\widehat{\lambda}_t = \frac{\sum_{i=1}^{N} I_{\{\widetilde{S}(\boldsymbol{Z}_i) \geq \widehat{\gamma}_t\}} \widetilde{W}(\boldsymbol{Z}_i; 1, \widehat{\lambda}_{t-1})}{\sum_{i=1}^{N} I_{\{\widetilde{S}(\boldsymbol{Z}_i) \geq \widehat{\gamma}_t\}} \widetilde{W}(\boldsymbol{Z}_i; 1, \widehat{\lambda}_{t-1})\, n^{-1} \sum_{j=1}^{n} Z_{ij}} \, , \tag{67}$$

where Z_{ij} is the j-th component of $\boldsymbol{Z}_i$.

Since, Z_i can be written as $Z_i = \lambda^{-1} \ln(1 - U_i)$, with $U_i \sim \mathsf{U}(0, 1)$, we have

$$X_i = u\, \lambda^{-1/a}\, \ln(1 - U_i)^{1/a}$$

so that under this change of measure $X_i \sim \mathsf{Weib}(a, u^{-1}\lambda^{1/a})$. Repeating the arguments of the ITLR method above, we find that this approach is equivalent to the two methods above, provided that we take $\lambda = \nu = (u/v)^a$.

Remark 4.1 (Sum of independent random variables) The special case where $S(\boldsymbol{X}) = X_1 + \cdots + X_n$, where the X_i are i.i.d. with a sub-exponential distribution was studied in both [3] and [19] via various methods, as explained in the introduction. In particular for the heavy tail Weibull case [19] proved (see their Theorem 3.2) that the change of measure

$$X_i \sim \mathsf{Weib}(a, 1) \longrightarrow \mathsf{Weib}(a, \eta^{1/a}) \tag{68}$$

provides a asymptotically optimal estimator, in the sense of (3), when we choose

$$\eta = c\,\gamma^{-a}, \tag{69}$$

no matter how c is chosen. On the other hand [3] proposed an importance sampling distribution independent of γ which is consistent with the fact that $\eta \to 0$. In the appendix of this paper we prove for the case $n = 2$ the somewhat stronger result that the estimator is in fact polynomial and that the variance of the estimator is minimized for $c = 2$; we conjecture that for general n the variance minimal (VM) parameter is

$${}_*\eta = n\,\gamma^{-a} \; .$$

In a forthcoming paper [4] it is proved that for large γ the optimal CE parameter, η^* say, is indeed given by ${}_*\eta$ above. More precisely, we show that asymptotically

$$\eta^* = \frac{n}{1 + \gamma^a}. \tag{70}$$

Similar results are obtained for the Pareto distribution. Moreover, in that paper we further explore the complexity properties of the SLR estimators applied to various queueing models and provide numerical comparisons with other methods.

5 Stationary waiting time of the GI/G/1 queue

Consider a stable GI/G/1 queue starting with customer $n = 1$ arriving at an empty system. Let the inter-arrival time between customer n and $n + 1$ be denoted by $A_n \sim f^A$, $n = 1, 2, \ldots$ and let the service time of customer n be denoted by $B_n \sim f^B$. We assume that all the service and inter-arrival times are independent. Let S_n denote the *actual waiting time* of the nth customer; hence, by definition $S_1 = 0$. The stochastic process $\{S_n, n \geq 1\}$ satisfies the celebrated *Lindley equation* (see for example [1])

$$S_{n+1} = (S_n + X_n)^+,$$

with $X_n = B_n - A_n$, $i = 1, 2 \ldots$. For a stable system the random variables $\{S_n\}$ converge in distribution to the *steady-state* waiting time, S say.

We are interested in estimating $\ell = \mathbb{P}(S \geq \gamma)$ via importance sampling. We consider two methods.

The regenerative method

Using the regenerative method, see for example [24], we can write

$$\ell = \frac{\mathbb{E} \sum_{n=1}^{\sigma} I_{\{S_n \geq \gamma\}}}{\mathbb{E}\sigma}, \tag{71}$$

where σ is the number of customers during the first busy period, that is

$$\sigma = \inf\{n > 1 \;:\; S_n = 0\} - 1 \;.$$

Define τ as

$$\tau = \inf\{n > 1 \;:\; S_n \geq \gamma\},$$

In other words, τ is the first time that the process $\{S_n\}$ exceeds level γ, if at all.

Consider now the following *switching* change of measure [24].

$$A_n \sim f^A \longrightarrow \widetilde{f}^A \quad \text{and} \quad B_n \sim f^B \longrightarrow \widetilde{f}^B \;, \quad \text{for } n = 1, \ldots, \min(\tau, \sigma) \;.$$

In other words, the IS distribution changes *dynamically* within the cycles. In particular, we initially use the IS densities $\widetilde{f}^A$ and $\widetilde{f}^B$ for the inter-arrival and service times until the process $\{S_n\}$ exceeds level γ, after which we switch back to the original densities, see [24], chapter 9. By doing so the process $\{S_n\}$ naturally returns to the regenerative state.

Under this change of measure the likelihood ratio of a sample $A_1, \ldots, A_n, B_1, \ldots, B_n$ satisfies

$$W_n = \begin{cases} W_{n-1} \dfrac{f^A(A_n) f^B(B_n)}{\widetilde{f}^A(A_n) \widetilde{f}^B(B_n)}, & n \leq \min(\tau, \sigma) \\ W_\tau, & n \geq \min(\tau, \sigma) \;. \end{cases} \tag{72}$$

From [24], we can write

$$\ell = \frac{\mathbb{E} W_\sigma \sum_{n=1}^{\sigma} I_{\{S_n \geq \gamma\}}}{\mathbb{E}\sigma} = \frac{\mathbb{E} \sum_{n=1}^{\sigma} I_{\{S_n \geq \gamma\}} W_n}{\mathbb{E}\sigma} \ . \tag{73}$$

Note that the denominator of (73) can be easily estimated via CMC (no change of measure here). The numerator of (73) (num) can be estimated as

$$\widehat{\text{num}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{n=1}^{\sigma_i} I_{\{S_{in} \geq \gamma\}} W_{in} \ , \tag{74}$$

where, S_{in} and W_{in} are the waiting time of the nth customer and the corresponding likelihood ratio, for iteration i.

Now consider the special case $A_1, A_2, \ldots \sim \mathsf{Weib}(a_1, u_1^{-1})$ and $B_1, B_2, \ldots \sim \mathsf{Weib}(a_2, u_2^{-1})$. Using the TLR method, we may write

$$X_n = u_2 \left(Z_n^{(2)}\right)^{1/a_2} - u_1 \left(Z_n^{(1)}\right)^{1/a_1} ,$$

with $Z_n^{(k)} \sim \mathsf{Exp}(1)$, $k = 1, 2$, $n = 1, 2, \ldots$, so that

$$S_{n+1} = \left(S_n + u_2 \left(Z_n^{(2)}\right)^{1/a_2} - u_1 \left(Z_n^{(1)}\right)^{1/a_1}\right)^{+} , \tag{75}$$

with $S_1 = 0$. Consider the following particular case of the switching change of measure described above:

$$Z_n^{(1)} \sim \mathsf{Exp}(1) \longrightarrow \mathsf{Exp}(v_1^{-1}) \ \text{ and } \ Z_n^{(2)} \sim \mathsf{Exp}(1) \longrightarrow \mathsf{Exp}(v_2^{-1}),\ n \leq \min(\tau, \sigma) \ .$$

Then (72) is given by

$$W_n = \begin{cases} W_{n-1} \prod_{k=1}^{2} v_k \, \mathrm{e}^{-(1-v_k^{-1}) Z_n^{(k)}}, & n \leq \min(\tau, \sigma) \\ W_\tau, & n \geq \min(\tau, \sigma) \ . \end{cases} \tag{76}$$

Since the $Z_n^{(k)}$ are independent and have an exponential distribution we can apply again the standard CE technique to determine/estimate the optimal reference parameters v_1^* and v_2^* for the estimator (74) and achieve variance reduction. In particular, if we define

$$H(\boldsymbol{Z}) = \sum_{n=1}^{\sigma} I_{\{S_n \geq \gamma\}} \ ,$$

with $\boldsymbol{Z} = (Z_1^{(1)}, Z_1^{(2)}, \ldots, Z_\sigma^{(1)}, Z_\sigma^{(2)})$, then, similar to Example 2.1, we have

$$v_k^* = \frac{\mathbb{E}_{\boldsymbol{v}} H(\boldsymbol{Z}) \, W_\tau \sum_{n=1}^{\tau} Z_n^{(k)}}{\mathbb{E}_{\boldsymbol{v}} H(\boldsymbol{Z}) \, W_\tau \, \tau} \ , \quad k = 1, 2,$$

for any reference vector $\boldsymbol{v} = (v_1, v_2)$. Note that in a multi-level CE procedure the updating rule for the level γ_t is not the "usual" quantile rule. Instead γ_t should be chosen such that during each regeneration cycle at least ρ percent of the customers has a waiting time $\geq \gamma$.

Random walk

It is well known (see for example page 173 of [24]) that the steady-state waiting time for this queueing system has the same distribution as the supremum of the random walk $\{Y_n,\, n = 1, \ldots\}$, where $Y_1 = 0$ and

$$Y_{n+1} = Y_n + X_n \ , \quad n \geq 1,$$

with $X_i = B_i - A_i$, $i = 1, 2 \ldots$, and the A_i and B_i the same as before. Thus ℓ in (71) is the same as

$$\ell = \mathbb{P}(\sup_n Y_n \geq \gamma) \ . \tag{77}$$

Similar to (75) let us now (re-)define

$$S_{n+1} = S_n + u_2 \left(Z_i^{(2)}\right)^{1/a_2} - u_1 \left(Z_i^{(1)}\right)^{1/a_1} , \tag{78}$$

with $Z_i^{(k)} \sim \mathsf{Exp}(1)$, $k = 1, 2$ Then, with $S = \sup_n S_n$, the estimation of (77) (under the original pdfs $\{\mathsf{Weib}(a_1, u_1^{-1})\}$ and $\{\mathsf{Weib}(a_2, u_2^{-1})\}$) is equivalent to the estimation of

$$\ell = \mathbb{P}(S \geq \gamma) \ .$$

Thus, alternatively to $I_{\{\sup_n Y_n \geq \gamma\}}$, which employs Weibull random variables we can simulate the random variable $I_{\{\sup_n S_n \geq \gamma\}}$ to estimate ℓ, which employs $\mathsf{Exp}(1)$ random variables $Z^{(1)}$ and $Z^{(2)}$. We can apply again the standard CE technique to find the optimal IS reference parameter.

To proceed, define τ as the first time $\{S_n\}$ exceeds level γ or falls below some low level $-L$, that is

$$\tau = \inf\{n > 0 \ : \ S_n \geq \gamma \text{ or } S_n < -L\} \ . \tag{79}$$

Consider, similar to before, the IS change of measure with $Z_i^{(k)} \sim \mathsf{Exp}(v_k^{-1})$. Typically, we seek for an IS change of measure under which the queue has a positive drift. In that case $S_\tau \geq \gamma$ with high probability. For $-L$ small enough we may write to a very close approximation

$$\ell \approx \mathbb{P}(S_\tau \geq \gamma) \ .$$

It will be clear how we estimate the probability above: we run N samples of $S_1, \ldots, S_\tau$ and evaluate the estimator

$$\widehat{\ell} = \frac{1}{N} \sum_{i=1}^{N} I_{\{S_{\tau_i} \geq \gamma\}} W_{\tau_i} \ ,$$

where

$$W_\tau = \prod_{k=1}^{2} \prod_{n=1}^{\tau} v_{t-1,k} \, \mathrm{e}^{-(1 - v_{t-1,k}^{-1}) Z_n^{(k)}} \ .$$

Applying the CE Algorithm 2.1 it is readily seen that the deterministic updating rules for $\boldsymbol{v}_t = (v_{t,1}, v_{t,2})$ are

$$v_{t,k} = \frac{\mathbb{E}_{\boldsymbol{v}_{t-1}} I_{\{S_\tau \geq \gamma_t\}} W_\tau \sum_{n=1}^{\tau} Z_n^{(k)}}{\mathbb{E}_{\boldsymbol{v}_{t-1}} I_{\{S_\tau \geq \gamma_t\}} W_\tau \, \tau} ,$$

with $v_{0,k} = 1$, $k = 1, 2$. This leads to the simulated updating rules

$$\widehat{v}_{t,k} = \frac{\sum_{i=1}^{N} I_{\{S_{i\tau_i} \geq \widehat{\gamma}_t\}} W_{i\tau_i} \sum_{n=1}^{\tau_i} Z_{kn}^{(i)}}{\sum_{i=1}^{N} I_{\{S_{i\tau_i} \geq \widehat{\gamma}_t\}} W_{i\tau_i} \, \tau_i} ,$$

where the simulation is run under $\boldsymbol{v}_{t-1}$. Note that the updating rules for method 1 and 2 are very similar. Indeed, it is reasonable to expect that the optimal CE parameters for the two methods should coincide for large γ; numerical results indicate that this is indeed the case. Finally we remark that some care should be taken with the choice of the low level $-L$. Typically, under the CE optimal parameter the system becomes *unstable* and hence $-L$ can be safely set to $-\infty$, but for the first iteration the system is still stable and hence $-L$ has to be chosen not too small in order to save CPU time.

Remark 5.1 It is important to set L in any simulation involving (79) large enough in order to obtain a valid estimator for the steady state waiting time probabilities. The choice of L is somewhat arbitrary. An alternative approach is to take $L = 0$ and let ℓ correspond to the probability that the waiting time process exceeds level γ during a busy period. This is called the *transient setting* in [24], section 9.3.2. In our numerical results we will consider examples of both cases.

6 Numerical Results

This section presents simulation studies for the rare event probability $\ell = \mathbb{P}(S(\boldsymbol{X}) \geq \gamma)$ for several static and queueing models with both light and heavy tail distributions. We shall employ both the SLR (25) and TLR estimators.

Unless otherwise specified we set in all our experiments with Algorithm 2.1 the rarity parameter $\rho = 0.01$, the sample size for step 2–4 of the algorithm $N = 10^4$ and for the final sample size $N_1 = 5 \cdot 10^5$.

For quite moderate probability like $\ell = 10^{-3}$, we typically compare the CE results with the corresponding CMC results.

6.1 Sum of Weibull random variables

Our first model concerns five i.i.d. $\mathsf{Weib}(a, u^{-1})$ random variables with $a = 5$ and $a = 0.2$, respectively. For both cases we selected $u = 1$. We wish to estimate

$$\mathbb{P}(X_1 + \cdots + X_5 \geq \gamma) .$$

Tables 1 and Tables 2 present, for the cases $a = 5$ and $a = 0.2$, respectively, the performance of Algorithm 2.1 for the TLR method

$$X_i = u Z_i^{1/a}, \quad Z_i \sim \mathsf{Exp}(1) \longrightarrow \mathsf{Exp}(v_i^{-1}) \tag{80}$$

which is equivalent to the (one-parameter) SLR method

$$X_i \sim \mathsf{Weib}(a, u^{-1}) \longrightarrow \mathsf{Weib}(a, v_i^{-1/a}) \ .$$

t	γ_t	v_{1t}	v_{2t}	v_{3t}	v_{4t}	v_{5t}
0	-	1	1	1	1	1
1	5.7	2.37	2.42	2.54	2.49	2.46
2	6.7	5.52	4.91	4.84	4.97	5.20
3	7.0	6.06	6.04	6.03	5.93	5.89
4	7.0	5.99	5.96	6.02	6.00	5.99
5	7.0	5.95	5.90	6.03	6.04	5.98
6	7.0	5.95	5.98	6.04	5.93	5.98
7	7.0	6.03	5.93	6.01	6.01	5.95
8	7.0	6.00	6.08	6.02	5.90	5.95

Table 1: The evolution of the estimate of $\boldsymbol{v}_t$ of the optimal parameters $\boldsymbol{v}^*$ with the TLR method (80), with $a = 5$. The estimated probability is $\widehat{\ell} = 1.6694 \cdot 10^{-9}$, the relative error $RE = 0.011763$ and $\kappa^2 = 62.2$

t	γ_t	v_{1t}	v_{2t}	v_{3t}	v_{4t}	v_{5t}
0	-	1	1	1	1	1
1	9.7e+003	2.45	2.25	2.55	1.97	2.12
2	6.4e+005	3.06	3.70	4.28	3.54	4.62
3	1.0e+006	3.68	5.82	3.92	3.34	4.35
4	1.0e+006	4.37	3.88	4.13	4.62	3.67
5	1.0e+006	4.13	4.47	4.11	3.77	4.37
6	1.0e+006	4.15	4.53	3.98	3.94	3.99
7	1.0e+006	4.10	4.22	4.40	4.11	4.16
8	1.0e+006	4.18	4.39	4.35	4.53	4.11

Table 2: The evolution of the estimate of $\boldsymbol{v}_t$ of the optimal parameters $\boldsymbol{v}^*$ with the TLR method (80), with $a = 0.2$. The estimated probability is $\widehat{\ell} = 6.54 \cdot 10^{-7}$, relative error $RE = 0.0278$ and $\kappa^2 = 386$

Note that in both cases Algorithm 2.1 reaches the desired level γ after three iterations, but we have continued iterating steps 2 – 4 of Algorithm 2.1 in view of Remark 2.2. We see that the parameter vector $\boldsymbol{v}_t$ stabilizes very quickly. Note also that we could have taken the average of the reference parameter at each iteration as a more accurate estimate for the true optimal reference parameter.

The asymptotical value for optimal reference parameter v in the heavy tail case is, see (70), given by

$$\frac{1}{\eta^*} = \frac{1 + \gamma^a}{n}.$$

In particular for Table 2 we obtain a value of $(1 + 10^{1.2})/5 \approx 3.4$, which is not too far from the observed value of around 4.2. Note that for the light tail case the above formula does not hold.

Tables 3 and 4 present, for the same cases $a = 5$ and $a = 0.2$ as above, the performance of Algorithm 2.1 for the two-parameter SLR method

$$X_i \sim \mathsf{Weib}(a, u^{-1}) \longrightarrow \mathsf{Weib}(b_i, v_i^{-1/b_i}) \tag{81}$$

of Remark 2.2.

t	γ_t	b_{1t}	v_{1t}	b_{2t}	v_{2t}	b_{3t}	v_{3t}	b_{4t}	v_{4t}	b_{5t}	v_{5t}
0	-	5	1	5	1	5	1	5	1	5	1
1	5.19	7.57	2.84	7.10	2.69	7.03	2.63	7.40	2.99	7.25	2.67
2	5.87	9.12	8.73	8.93	7.87	9.89	10.38	10.09	10.11	10.47	10.50
3	6.41	11.22	28.51	11.86	37.42	12.10	39.75	11.27	34.33	12.21	34.50
4	6.86	12.25	70.59	12.09	106.44	14.33	153.19	14.69	238.96	14.30	158.26
5	7.00	14.43	231.51	14.13	250.12	12.96	109.01	11.25	92.41	13.88	179.26
6	7.00	14.08	201.95	13.78	206.56	13.63	167.13	12.81	128.66	14.32	246.63
7	7.00	14.04	211.85	13.99	209.57	14.22	206.02	13.33	167.56	14.01	205.50
8	7.00	14.19	202.57	13.22	193.80	13.98	183.36	12.71	133.98	13.43	193.81
9	7.00	14.00	194.39	13.35	195.35	14.25	201.32	13.04	146.14	13.74	195.68
10	7.00	14.24	200.73	13.63	191.78	13.28	185.02	12.59	124.85	14.14	202.64

Table 3: The evolution of the estimates $\boldsymbol{b}_t$ and $\boldsymbol{v}_t$ of the optimal parameters $\boldsymbol{b}^*$ and $\boldsymbol{v}^*$ with the two-parameter SLR method (81). The estimated probability is $\widehat{\ell} = 1.6570 \cdot 10^{-9}$, the relative error $RE = 0.0041$ and $\kappa^2 = 8.4$

t	γ_t	b_{1t}	v_{1t}	b_{2t}	v_{2t}	b_{3t}	v_{3t}	b_{4t}	v_{4t}	b_{5t}	v_{5t}
0	-	0.2	1	0.2	1	0.2	1	0.2	1	0.2	1
1	971.28	0.17	1.55	0.18	1.69	0.18	1.63	0.17	1.48	0.18	1.52
2	28750	0.15	1.76	0.15	2.09	0.15	1.89	0.15	1.75	0.14	1.40
3	461370	0.12	1.86	0.13	1.84	0.12	1.43	0.12	1.53	0.13	2.38
4	1000000	0.12	1.50	0.13	2.17	0.12	1.83	0.13	1.62	0.11	1.93
5	1000000	0.12	1.59	0.12	1.66	0.12	1.92	0.11	1.66	0.12	1.99
6	1000000	0.13	1.68	0.12	2.02	0.12	1.96	0.12	1.83	0.13	1.91
7	1000000	0.12	1.72	0.13	1.97	0.12	1.87	0.12	1.77	0.12	1.87
8	1000000	0.12	1.81	0.12	2.05	0.12	1.90	0.13	1.94	0.12	1.67
9	1000000	0.12	1.95	0.12	1.70	0.13	1.88	0.12	1.66	0.12	1.88

Table 4: The evolution of the estimates $\boldsymbol{b}_t$ and $\boldsymbol{v}_t$ of the optimal parameters $\boldsymbol{b}^*$ and $\boldsymbol{v}^*$ with the two-parameter SLR method (81). The estimated probability is $\widehat{\ell} = 6.5964 \cdot 10^{-7}$, the relative error $RE = 0.014723$ and $\kappa^2 = 108.3$

We see that both the one- and two-parameter methods give very accurate results for both heavy and light tail Weibull distributions, and that the TLR updating performs similar to its two-parameter counterpart, although repeated measurements indicate that for the cases above the RE is about two times smaller for the two-parameter TLR method.

6.2 Sum of Pareto random variables

Here we repeat the experiments of Tables 1 and 2 for the Pareto case. Specifically, we now let the X_i have a Pareto pdf $f(x) = au^{-1}(1 + xu^{-1})^{-(1+a)}$ and consider the TLR change of measure

$$X_i = u\left(e^{Z_i/a} - 1\right), \quad Z_i \sim \mathsf{Exp}(1) \longrightarrow \mathsf{Exp}(v_i^{-1}) \,. \tag{82}$$

Tables 5 and 6 present the performance of the TLR method for $a = 5$ and $a = 0.2$, respectively. For both cases we selected $u = 1$ and took $N = 2 \cdot 10^5$ and $N_1 = 10^6$.

t	γ_t	v_{1t}	v_{2t}	v_{3t}	v_{4t}	v_{5t}
0	-	1	1	1	1	1
1	2.14	1.90	1.88	1.88	1.93	1.93
2	5.56	2.95	2.94	2.93	2.93	2.96
3	13.06	3.67	3.62	3.46	3.68	3.87
4	22.41	4.50	3.99	4.19	4.30	3.89
5	25.00	3.61	5.35	3.92	4.52	3.88
6	25.00	4.02	4.24	4.40	4.36	4.45
7	25.00	4.44	4.30	4.26	4.09	4.27
8	25.00	4.38	4.18	4.11	4.09	4.63
9	25.00	4.27	4.07	4.29	4.47	4.25
10	25.00	4.38	4.33	4.41	4.28	3.92

Table 5: The evolution of the estimate of v_t of the optimal parameters v^* with the TLR method for $a = 5$. The estimated probability is $\widehat{\ell} = 5.22 \cdot 10^{-7}$, the relative error $RE = 0.0238$ and $\kappa^2 = 570.98$

t	γ_t	v_{1t}	v_{2t}	v_{3t}	v_{4t}	v_{5t}
0	-	1	1	1	1	1
1	2.6e+008	1.74	1.75	1.75	1.74	1.74
2	4.6e+014	2.32	2.34	2.33	2.38	2.34
3	4.9e+019	2.78	2.86	2.72	2.89	2.82
4	4.9e+023	3.22	3.24	2.99	3.26	3.23
5	6.7e+026	3.56	3.47	3.26	3.48	3.58
6	1.3e+029	3.80	4.02	3.29	3.74	3.53
7	1.0e+031	3.60	4.09	3.74	4.11	3.78
8	4.5e+032	3.74	4.05	3.91	3.39	4.67
9	2.8e+033	4.00	4.72	3.78	3.81	4.48
10	1e+035	4.48	3.97	4.12	4.57	3.86
11	1e+035	4.16	4.35	4.57	3.99	4.11
12	1e+035	4.37	4.49	4.16	4.13	4.00
13	1e+035	4.14	4.00	4.25	4.11	4.54
14	1e+035	4.12	4.24	4.44	4.16	4.24
15	1e+035	4.30	4.16	4.53	4.18	4.30

Table 6: The evolution of the estimate of v_t of the optimal parameters v^* with the TLR method for $a = 0.2$. The estimated probability is $\widehat{\ell} = 4.86 \cdot 10^{-7}$, relative error $RE = 0.0267$ and $\kappa^2 = 716.74$

Although in this case the TLR change of measure (82) does not seem as "natural" as the SLR one, where a or u is changed, we can see, however, that again a good variance reduction is obtained. In fact, the variance reduction with TLR was very similar to the SLR change of measure, which was also implemented. An advantage of (82) is that only one line of the code for the Weibull case needed to be changed.

6.3 Stochastic shortest path

Our second model concerns a *stochastic shortest path* problem. Consider the weighted graph of Figure 1, with random weights $X_1, \dots, X_5$.

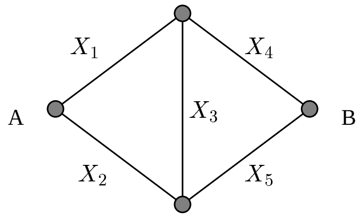

Figure 1: Stochastic shortest path from A to B

Suppose the rv's $X_1, \dots, X_5$ are independent of each other and have a $\mathsf{Weib}(a_i, u_i)$ distribution, $i = 1, \dots, 5$. Let $S(\boldsymbol{X})$ be the length of the shortest path from node A to node B. Note that there are four possible paths. We wish to estimate from simulation the probability $\ell = \mathbb{P}(S(\boldsymbol{X}) \geq \gamma)$ that the length of the shortest path $S(\boldsymbol{X})$ will exceed some fixed γ.

We consider the light- and heavy-tail cases $a_i = 5$ and $a_i = 0.2, i = 1, \dots, 5$. In both cases $\boldsymbol{u} = (0.25, 0.4, 0.1, 0.3, 0.2)$.

Tables 7 and 8 present the performance of Algorithm 2.1 with the TLR method (80), for the cases $a = 5$ and $a = 0.2$ respectively. The results are self-explanatory.

t	γ_t	v_{1t}	v_{2t}	v_{3t}	v_{4t}	v_{5t}
0	-	1	1	1	1	1
1	0.568	2.491	1.530	1.267	1.748	1.931
2	0.650	4.257	2.152	1.543	2.431	2.977
3	0.706	6.052	2.705	1.896	3.294	4.153
4	0.752	8.125	3.476	2.260	4.128	5.360
5	0.792	10.356	4.074	2.630	4.994	6.687
6	0.800	10.293	4.126	2.850	5.519	7.460
7	0.800	10.712	4.265	2.520	5.090	7.109
8	0.800	10.550	4.125	2.565	5.310	7.383
9	0.800	10.897	4.377	2.577	5.277	7.096

Table 7: The evolution of the estimate $\boldsymbol{v}_t$ of the optimal parameter $\boldsymbol{v}^*$ with the TLR method and $a = 5$. The estimated probability is $\widehat{\ell} = 1.20 \cdot 10^{-10}$, the relative error 0.044.

t	γ_t	v_{1t}	v_{2t}	v_{3t}	v_{4t}	v_{5t}
0	-	1	1	1	1	1
1	6.760	2.005	1.906	1.166	1.857	1.912
2	159.419	3.067	2.911	1.038	2.499	2.619
3	1070.002	4.226	3.940	1.052	3.029	3.211
4	4173.601	5.320	4.930	0.854	3.598	3.901
5	11663.017	6.877	6.333	1.118	3.730	3.867
6	34307.081	9.237	8.434	1.078	3.461	3.548
7	100000.000	7.030	6.623	0.842	7.762	7.658
8	100000.000	11.309	10.660	1.043	3.227	3.474
9	100000.000	14.038	13.035	0.981	1.126	1.189
10	100000.000	14.261	13.008	0.979	1.066	1.035

Table 8: The evolution of the estimate $\boldsymbol{v}_t$ of the optimal parameter vector $\boldsymbol{v}^*$ with the TLR method and $a = 0.2$. The estimated probability is $\widehat{\ell} = 1.09 \cdot 10^{-11}$ the relative error 0.026.

6.4 GI/G/1 queue

Our third model is the GI/G/1 queue with inter-arrival time distribution $\mathsf{Weib}(a_1, u_1^{-1})$ and service time distribution $\mathsf{Weib}(a_2, u_1^{-2})$. Note that the traffic intensity of the queue is thus given by

$$\frac{u_2 \Gamma(1 + 1/a_2)}{u_1 \Gamma(1 + 1/a_1)} \ .$$

We first consider the estimation of the probability that the stationary waiting time in the queue exceeds some fixed level γ, using the *random walk* method described in Section 5.

In particular, with A_i and B_i the inter-arrival and service times, we use the TLR change of measure

$$\begin{aligned} A_i &= u_1 \left(Z_i^{(1)}\right)^{1/a_1}, \qquad Z_i^{(1)} \sim \mathsf{Exp}(1) \longrightarrow \mathsf{Exp}(v_1^{-1}) \\ B_i &= u_2 \left(Z_i^{(2)}\right)^{1/a_2}, \qquad Z_i^{(2)} \sim \mathsf{Exp}(1) \longrightarrow \mathsf{Exp}(v_2^{-1}) \ . \end{aligned} \tag{83}$$

Table 9 illustrates the evolution of Algorithm 2.1 for determining the CE optimal parameters v_1 and v_2 to be used in the TLR estimator. In this particular case the parameters are $a_1 = 0.5$, $u_1 = 1$, $a_2 = 0.5$ and $u_2 = 0.5$, which gives a traffic intensity of 0.5. The level to be exceeded is $\gamma = 80$. The sample size used in steps 1–4 was $N = 50,000$. The rarity parameter ρ was set to 0.01.

We have repeated steps 2–4 four more times after reaching γ in order to show the accuracy of the estimation of the true optimal CE parameter. (The corresponding estimate and RE for this case are given in Table 10.)

t	γ_t	v_1	v_2
0	-	1	1
1	39.5	0.774073	1.39477
2	80	0.796896	1.44949
3	80	0.813729	1.42962
4	80	0.810056	1.40465
5	80	0.799487	1.43608
6	80	0.801236	1.44118

Table 9: The evolution of Algorithm 2.1 using the TLR method for the $GI/G/1$ with the following parameters: $a_1 = 0.5$, $v_1 = 1$, $a_2 = 0.5$, $v_2 = 0.5$

It is interesting to note that after one iteration the system becomes *unstable*, so that γ_t in step 2 of the CE algorithm reaches level γ in just *two* iterations. This is in accordance with the *instability property* of the CE algorithm described and analyzed in [6]. As a consequence, the choice of the rarity parameter does not matter very much.

Tables 10 – 13 summarize some performance characteristics of the TLR estimation procedure as a function of γ, for various light and heavy-tail cases. In all cases we set $N = 10^4$ and $N_1 = 5 \cdot 10^5$. Also, the rarity parameter ρ was set to 0.1 (in fact any parameter $\rho < 1$ would be ok) and the level $-L$ was set low enough to -100.

In all tables we report the optimal CE parameters (recall that the original ones are 1), the estimate of the probability, the relative error and the CPU time in seconds.

$a_1 = 0.5, u_1 = 1, \quad a_2 = 0.5, u_2 = 0.5$						
γ	20	40	60	80	100	120
v_1^*	0.78	0.79	0.80	0.80	0.80	0.81
v_2^*	1.36	1.38	1.40	1.41	1.43	1.45
$\widehat{\ell}$	$7.139 \cdot 10^{-2}$	$1.152 \cdot 10^{-2}$	$2.08 \cdot 10^{-3}$	$4.25 \cdot 10^{-4}$	$8.99 \cdot 10^{-5}$	$2.08 \cdot 10^{-5}$
RE	0.002	0.0036	0.0067	0.016	0.020	0.045
sec	149	264	396	467	587	696

Table 10: Simulation results for method 2 for the waiting time probabilities of a GI/G/1 queue with heavy tail inter-arrival and service time distributions, as a function of γ. The traffic intensity is 0.5.

$a_1 = 2, u_1 = 1, \quad a_2 = 2, u_2 = 0.75$						
γ	3	6	9	12	15	18
v_1^*	0.56	0.56	0.56	0.56	0.56	0.56
v_2^*	1.57	1.58	1.58	1.58	1.58	1.59
$\widehat{\ell}$	$1.031 \cdot 10^{-2}$	$1.63 \cdot 10^{-4}$	$2.60 \cdot 10^{-6}$	$4.15 \cdot 10^{-8}$	$6.63 \cdot 10^{-10}$	$1.56 \cdot 10^{-11}$
RE	0.0017	0.0027	0.0040	0.0053	0.013	0.016
sec	101	154	210	274	338	398

Table 11: Simulation results for method 2 for the waiting time probabilities of a GI/G/1 queue with light tail inter-arrival and service time distributions, as a function of γ. The traffic intensity is 0.75.

$a_1 = 1, u_1 = 2,\quad a_2 = 1, u_2 = 1.5$						
γ	20	40	60	80	100	120
v_1^*	0.75	0.75	0.75	0.75	0.75	0.75
v_2^*	1.33	1.33	1.33	1.33	1.33	1.33
$\widehat{\ell}$	$2.676 \cdot 10^{-2}$	$9.539 \cdot 10^{-4}$	$3.404 \cdot 10^{-5}$	$1.214 \cdot 10^{-6}$	$4.333 \cdot 10^{-8}$	$1.546 \cdot 10^{-9}$
RE	0.00036	0.00039	0.00040	0.00040	0.00038	0.00053
sec	160	429	509	558	691	828

Table 12: Simulation results for method 2 for the waiting time probabilities of an M/M/1 queue, as a function of γ. The traffic intensity is 0.75.

$a_1 = 1, u_1 = 1,\quad a_2 = 0.5, u_2 = 0.25$						
γ	10	20	30	40	50	60
v_1^*	0.81	0.83	0.84	0.85	0.85	0.89
v_2^*	1.64	1.68	1.71	1.65	1.73	1.77
$\widehat{\ell}$	$2.83 \cdot 10^{-2}$	$3.55 \cdot 10^{-3}$	$5.63 \cdot 10^{-4}$	$1.05 \cdot 10^{-4}$	$2.60 \cdot 10^{-5}$	$7.07 \cdot 10^{-6}$
RE	0.003	0.0067	0.012	0.017	0.047	0.093
sec	108	190	224	306	335	407

Table 13: Simulation results for method 2 for the waiting time probabilities of an M/G/1 queue, with heavy tail service distribution, as a function of γ. The traffic intensity is 0.5.

The results seem to indicate that the RE increases (sub)linearly, but there is not sufficient evidence to conclude that the estimators are polynomial, except in the M/M/1 case, where the RE remains constant. In the latter case we have the well-known optimal (exponential) change of measure where the service and inter-arrival rates are interchanged. What is clearer is that for the light tail case we can estimate much smaller probabilities than for the heavy tail case, for a given accuracy (RE) and simulation effort. It is interesting to note that for the second experiment (with $a_1 = a_2 = 2$) quite small probabilities can be efficiently estimated despite the fact that the TLR estimator is not asymptotically optimal. Namely, the only asymptotically optimal estimator is obtained by an exponential change of measure, see Sadowsky [26] and Asmussen and Rubinstein [5], and the TLR change of measure for this case is obviously not an exponential change of measure.

Note also that for both light-tail cases the reference parameters seem to have "converged", but not yet for the two heavy-tail cases. Also the estimates for the reference parameters seem more noisy in the heavy tail case. In both the light and heavy tail case we observed that the estimates for the probabilities stabilized quite quickly (for moderate sample sizes). However, we also observed that accurate estimates for the variance of the estimator were much more difficult to obtain in the heavy-tail case than in the light-tail case.

We have repeated the experiments in Tables 10–13 for method 1, the *switching regenerative method*, using $N_1 = 5 \cdot 10^5$ regeneration cycles and using exactly the same CE parameters v_1^* and v_2^* as reported for method 2. The results were very similar to those of method 2. Tables 14 and 15 give the results for two of these experiments. We also ran the model with crude Monte Carlo, that is method 1 with $v_1 = v_2 = 1$, increasing the number of cycles to $5 \cdot 10^6$ in order to obtain execution times of the same order as the other methods. The SMC estimates were in exact agreement with the IS estimates, and the IS estimates consistently gave a significant variance reduction, although less pronounced in the heavy-tail case.

$a_1 = 0.5, u_1 = 1, \quad a_2 = 0.5, u_2 = 0.5$						
γ	20	40	60	80	100	120
$\widehat{\ell}$	$7.19 \cdot 10^{-2}$	$1.161 \cdot 10^{-2}$	$2.08 \cdot 10^{-3}$	$4.38 \cdot 10^{-4}$	$8.63 \cdot 10^{-5}$	$2.01 \cdot 10^{-5}$
RE	0.0087	0.011	0.017	0.029	0.034	0.071
sec	59	80	109	135	170	200

Table 14: Simulation results for method 1 for the waiting time probabilities of a GI/G/1 queue with heavy tail inter-arrival and service time distributions, as a function of γ. The traffic intensity is 0.5.

$a_1 = 2, u_1 = 1, \quad a_2 = 2, u_2 = 0.75$						
γ	3	6	9	12	15	18
$\widehat{\ell}$	$1.028 \cdot 10^{-2}$	$1.63 \cdot 10^{-4}$	$2.58 \cdot 10^{-6}$	$4.12 \cdot 10^{-8}$	$6.76 \cdot 10^{-10}$	$1.05 \cdot 10^{-11}$
RE	0.0064	0.0084	0.011	0.019	0.020	0.021
sec	51	91	167	173	212	398

Table 15: Simulation results for method 1 for the waiting time probabilities of a GI/G/1 queue with light tail inter-arrival and service time distributions, as a function of γ. The traffic intensity is 0.75.

We also conducted various experiments in the *transient setting* (that is taking $L = 0$, see Remark 5.1, and using Pareto arrival and service times. Tables 16 – 17 present two examples. Table 18 presents an example using Pareto arrival and Weibull service time. For the Pareto case a similar TLR change of measure as in (82) was used. In all tables τ is as in (79) with $L = 0$ and SCV stands for the squared coefficient of variation for the random variable IW in the TLR estimator.

γ	40	120	160	240	300	360
N	$5 \cdot 10^4$	$5 \cdot 10^4$	$5 \cdot 10^4$	$5 \cdot 10^4$	$5 \cdot 10^4$	$5 \cdot 10^4$
N_1	$5 \cdot 10^5$	$5 \cdot 10^5$	$5 \cdot 10^5$	$5 \cdot 10^5$	10^6	10^6
$\widehat{\ell}$	1.76e-002	1.49e-003	4.78e-004	5.32e-005	1.00e-005	1.81e-006
RE	0.0068	0.013	0.016	0.023	0.021	0.026
$\widehat{\tau}$	94.26	655.96	1137.86	2075.22	3305.66	3703.51
SCV	23.46	87.73	129.57	283.18	430.81	664.02

Table 16: Transient simulation results as function of γ for a GI/G/1 queue with the inter-arrival distribution $\mathsf{Pareto}(0.5, 0.4)$ and service distribution $\mathsf{Pareto}(0.5, 0.36)$. The traffic intensity is 0.9. For $\gamma = 40$ the probability was checked by CMC estimator: $\widehat{\ell} = 1.78 \cdot 10^{-2}$

γ	25	50	80	120	250	350
N	10^5	10^5	10^5	10^5	10^5	10^5
N_1	$5 \cdot 10^5$	$5 \cdot 10^5$	$5 \cdot 10^5$	$5 \cdot 10^5$	$5 \cdot 10^5$	$5 \cdot 10^5$
$\widehat{\ell}$	1.25e-003	8.72e-005	9.66e-006	2.51e-006	2.59e-007	7.54e-008
RE	0.011	0.029	0.041	0.047	0.053	0.051
$\widehat{\tau}$	75.60	88.26	69.75	79.20	92.18	93.76
SCV	61.81	427.33	841.23	1082.71	1387.39	1301.01

Table 17: Transient simulation results as function of γ for GI/G/1 queue with the inter-arrival distribution $\mathsf{Pareto}(3, 0.75)$ and service distribution $\mathsf{Pareto}(3, 1)$. The traffic intensity is 0.75.

γ	20	50	130	160	300	400
N	$5 \cdot 10^4$	$5 \cdot 10^4$	$5 \cdot 10^4$	$5 \cdot 10^4$	$5 \cdot 10^4$	$5 \cdot 10^4$
N_1	$2 \cdot 10^5$	$3 \cdot 10^5$	$3 \cdot 10^5$	$3 \cdot 10^5$	$3 \cdot 10^5$	$3 \cdot 10^5$
$\widehat{\ell}$	2.37e-003	2.20e-004	1.05e-005	8.25e-006	1.33e-006	5.75e-007
RE	0.016	0.030	0.033	0.029	0.028	0.027
$\widehat{\tau}$	18.38	17.21	16.03	14.44	16.08	14.92
SCV	51.94	275.31	326.70	258.69	239.94	220.05

Table 18: Transient simulation results as function of γ for GI/G/1 queue with the inter-arrival distribution $\mathsf{Weib}(2, 1)$ and service distribution $\mathsf{Pareto}(2.5, 1)$. The traffic intensity is 0.75225.

6.5 Two non-Markovian queues with feedback

As a final example, we consider the network depicted in Figure 2. It consists of two queues in tandem, where customers departing from the second queue either leave the network (with probability p), or go back to the first queue (with probability $1-p$). We are interested in estimating the transient probability that the total number of customers in the network exceeds some high level, 50 in this example, during one busy cycle. This model was also considered in [6], using only light-tail distributions and applying IS with exponential twisting.

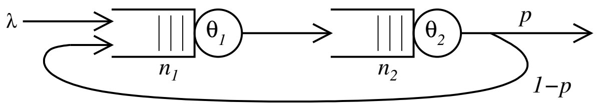

Figure 2: Two queues in tandem with feedback

In the experiments reported below the inter-arrival time distribution is a two-stage Erlang distribution, with exponential parameter $\lambda = 0.2$. The service time distributions of the first queue is uniform on $[0, 3.333]$. In the second queue the service time distribution is $\mathsf{Weib}(a, c)$. In Table 19 we consider the light tail case with $a = 2$ and $c = 0.354491$, which gives a mean service time of 2.5, while in Table 20 we consider the heavy tail case with $a = 0.8$ and $c = 0.453201$, which gives again mean service time of 2.5. We note that this is the same mean service time as in [6]. In the tables, θ is the the exponential twisting parameter for the uniform distribution. The λ column gives the evolution of reference parameter for the Erlang inter-arrivals, and similar for U and p.

t	γ_t	λ	θ	c	p
0	3.0	0.200000	0.000000	0.354491	0.5
1	50	0.342317	-0.023671	0.294095	0.177778
2	50	0.363233	0.000000	0.315648	0.225282
3	50	0.360159	0.000000	0.320599	0.234336
4	50	0.360873	-0.003051	0.320986	0.234113
5	50	0.358857	-0.003623	0.320894	0.235779
6	50	0.360186	-0.000707	0.320591	0.234769
7	50	0.359469	-0.003483	0.320718	0.234796

Table 19: Simulation results for the non-Markovian network for $\ell = 50$. Here $N = N_1 = 10^4$. The estimated probability is $\widehat{\ell} = 1.62e - 25$, the relative error $RE = 0.018$

We see that the optimal CE parameters are estimated quite accurately for a relatively small N. Since the second queue is the bottleneck *state independent tilting*, changing the parameters irrespective of the state of the queue, seems to work nicely, and the TRL method seems to deliver an accurate estimate of a very small probability. No numerical results are available for validation; therefore, we repeated the experiment various times. The fact that we obtained similar estimates gives confidence.

t	γ_t	λ	θ	c	p
0	3.0	0.200000	0.000000	0.453201	0.5
1	50	0.300620	0.000000	0.263503	0.3019
2	50	0.301135	0.000000	0.263982	0.3031
3	50	0.301291	-0.000000	0.264346	0.3026
4	50	0.300832	0.000000	0.263580	0.3031
5	50	0.301350	-0.000000	0.263770	0.3029
6	50	0.300620	0.000000	0.263503	0.3019
7	50	0.301135	0.000000	0.263982	0.3031

Table 20: Simulation results for the non-Markovian network for $\ell = 50$. Here $N = N_1 = 10^5$The estimated probability is $\widehat{\ell} = 4.323e - 18$, the relative error $RE = 0.0079$

For this heavy tail case a similar picture emerges: the estimates for the reference parameters are quite stable a small probability can be estimated with reasonable accuracy. However, when we repeat this for a smaller a ($a = 0.5$) the results were not so satisfactory, indicating that a (much) larger sample size is required.

A The sum of two Weibulls

As noted in Remark (4.1) for the sum of n heavy-tail Weibulls, the change of measure given by (68) for any constant c in (69) gives an SLR estimator which is asymptoticall optimal. A proof of this is given in Theorem 3.2 of [19]. In this appendix we prove that for the case $n = 2$ and for large γ the best, that is, minimum variance, choice for c is $c = n = 2$ and that the estimator is not only asymptotically optimal, but in fact polynomial. We conjecture that in general

$c = n$. We show explictly that the relative error grows (for $n = 2$) as γ^{2a}, and we conjecture that in general it grows as γ^{na}. The proof below uses the TLR representation of the change of measure, but it could as easily have been given via an SLR approach. Most of the result hold for the light ($a \geq 1$) and heavy tail $a < 1$ case, except when the subexponentiality property is used for the heavy-tail case. Without loss of generality we take $u = 1$.

Thus the problem is as follows: Let X_1, X_2 be i.i.d. $\mathsf{Weib}(a, 1)$ distributed; estimate

$$\ell = \mathbb{P}(X_1 + X_2 \geq \gamma) = \mathbb{P}(Z_1^{1/a} + Z_2^{1/a} \geq \gamma) \; ,$$

with $Z_i \sim \mathsf{Exp}(1)$, independent. Consider the exponential change of measure $Z_i \sim \mathsf{Exp}(1) \longrightarrow \mathsf{Exp}(1-\theta)$, where $0 \leq \theta < 1$ is the exponential twisting parameter. Let $\mathbb{E}_\theta$ denote the corresponding expectation operator. Thus $\mathbb{E}_0$ corresponds to the original $\mathsf{Exp}(1)$ distribution. We have

$$\ell = \mathbb{E}_\theta I_{\{Z_1^{1/a} + Z_2^{1/a} \geq \gamma\}} W \; .$$

Here $W = W(\theta)$ is shorthand notation for the likelihood ratio

$$W(\theta) = \mathrm{e}^{-\theta(Z_1+Z_2)+2\zeta(\theta)} = \frac{\mathrm{e}^{-\theta(Z_1+Z_2)}}{(1-\theta)^2},$$

where we have used the fact that the cumulant function for this exponential family is given by $\zeta(\theta) = \ln(1/(1-\theta)) = -\ln(1-\theta)$.

There does not exist a simple formula for ℓ as a function of a and γ, but it is not difficult to verify that

$$\begin{aligned} \ell(\gamma) &= \left(2 \iint_{A_1} + \iint_{A_2} + 2 \iint_{A_3}\right) \mathrm{e}^{-(z_1+z_2)}\, dz_1 dz_2 \\ &= \exp\left(-\gamma^a\, 2^{1-a}\right) + 2 \int_0^{(\gamma/2)^a} \exp\left(-\left\{\gamma - x^{1/a}\right\}^a - x\right)\, dx \; , \end{aligned}$$

where the regions A_1, A_2 and A_3 are given in Figure 3.

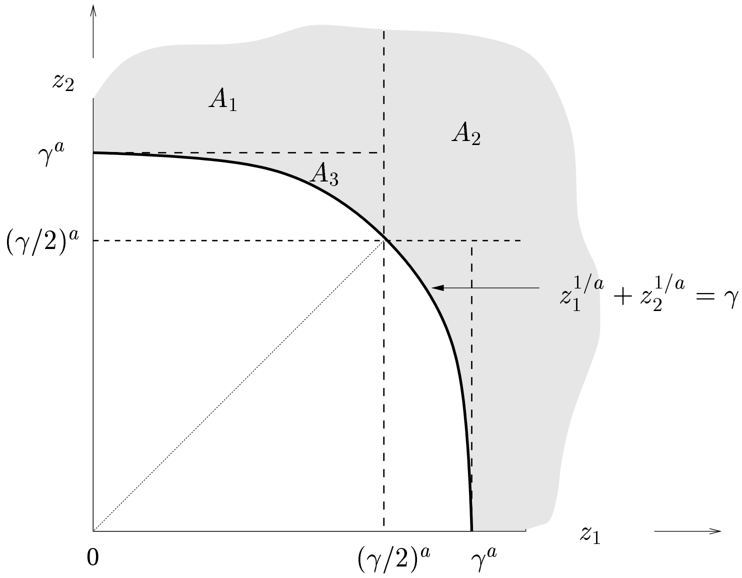

Figure 3: ℓ is equal to the integral of $\mathrm{e}^{-(z_1+z_2)}$ over the shaded region.

Let us mention some known facts about ℓ. First, for the heavy-tail case $a < 1$ it is well-known that the Weibull distribution is *sub-exponential*, which means that the sum of n i.i.d. Weibull random variables satisfies

$$\lim_{\gamma \to \infty} \frac{\mathbb{P}(X_1 + \cdots + X_n \geq \gamma)}{\mathbb{P}(X_1 \geq \gamma)} = n \ .$$

In particular, for our $n = 2$ case we have that

$$\lim_{\gamma \to \infty} \frac{\ell(\gamma)}{2\mathrm{e}^{-\gamma^a}} = 1 \ .$$

For $a = 1$ it is not difficult to see that

$$\ell = \mathrm{e}^{-\gamma}(\gamma + 1) \ .$$

For $a > 1$ one can show that

$$\lim_{\gamma \to \infty} \frac{\ell(\gamma)}{\mathrm{e}^{-2(\gamma/2)^a}\ \gamma^{a/2}} = c(a) \ ,$$

for some constant $c(a)$, decreasing as a increases. For example, for $a = 2$, $c(a) = \sqrt{\pi/2}$ and for $a = 3$, $c(a) = \sqrt{3\pi}/4$.

Let us now turn to the complexity properties of the TLR estimator, as a function of γ. This is, as always, determined by the second moment (under θ) of the random variable $IW = I_{\{Z_1^{1/a} + Z_2^{1/a} \geq \gamma\}} W(\theta)$. Using a simplified notation we have

$$\begin{aligned}
\mathbb{E}_\theta\,(IW)^2 &= \mathbb{E}_\theta\; IW^2 \\
&= \mathbb{E}_0\; IW \\
&= \mathbb{E}_0\; I\frac{\mathrm{e}^{-\theta(Z_1+Z_2)}}{(1-\theta)^2} \\
&= \left(2\iint_{A_1} + \iint_{A_2} + 2\iint_{A_3}\right) \frac{\mathrm{e}^{-(1+\theta)(z_1+z_2)}}{(1-\theta)^2}\, d\boldsymbol{z}\ .
\end{aligned} \tag{84}$$

We wish to show that the SCV increases at most polynomially in γ, for a certain choice of θ. This is equivalent to showing that $\mathbb{E}_\theta\,(IW)^2/\ell^2$ increases at most polynomially in γ. We do this by considering the contributions of the three integrals in (84) individually.

Define $D_i = \iint_{A_i} \frac{\mathrm{e}^{-(1+\theta)(z_1+z_2)}}{(1-\theta)^2}\, d\boldsymbol{z}, i = 1,2,3$. The easiest of these is D_2; namely

$$D_2 = \left(\frac{\mathrm{e}^{-(1+\theta)(\gamma/2)^a}}{1-\theta^2}\right)^2 .$$

It follows that for fixed θ

$$\lim_{\gamma\to\infty} \frac{D_2}{\ell^2} = \frac{4}{(1-\theta^2)^2} \lim_{\gamma\to\infty} \mathrm{e}^{-\gamma^a\left\{(1+\theta)2^{1-a}-2\right\}} = 0\ ,$$

provided that $1+\theta > 2^a$, or equivalently $1-\theta < 2-2^a$.

Second, we have

$$D_1 \le \widetilde{D}_1 = \int_0^\infty \int_{\gamma^a}^\infty \frac{\mathrm{e}^{-(1+\theta)(z_1+z_2)}}{(1-\theta)^2}\, d\boldsymbol{z} = \frac{1}{(1-\theta^2)^2}\,\mathrm{e}^{-\gamma^a(1+\theta)}\ .$$

The contribution of D_1 to the SCV is therefore bounded by

$$\frac{\widetilde{D}_1}{\ell^2} \approx \frac{1}{2\,(1-\theta^2)^2}\,\mathrm{e}^{-\gamma^a(1+\theta-2)}\ .$$

As a consequence, this contribution remains polynomial in γ if we choose $\theta = 1 - c\gamma^{-a}$, for any c. In that case

$$\frac{\widetilde{D}_1}{\ell^2} \approx \frac{\mathrm{e}^c\gamma^{4a}}{4c^2(c-2\gamma^a)^2}\ .$$

If we minimize this with respect to c, we obtain for fixed γ the minimal argument

$$c^* = \gamma^a - \sqrt{\gamma^{2a}+4} + 2\ .$$

For large γ we have thus $c \approx 2$. This suggests we take

$$\theta = 1 - 2\gamma^{-a}\ .$$

It is obvious that with this choice of θ the contribution of D_2 to the SCV is tends 0, as γ increases. It follows that

$$\frac{2D_1 + D_2}{\ell^2} = \gamma^{2a}\frac{\mathrm{e}^2}{64} + o\left(\gamma^{2a}\right) .$$

It remains to show that the contribution of D_3 remains polynomial. We have

$$\begin{aligned}\frac{D_3}{\ell^2} &\approx \frac{\mathrm{e}^{2\gamma^a}}{2}\int_0^{(\gamma/2)^a}\int_{\left(\gamma - z_1^{1/a}\right)^a}^{\gamma^a} \frac{\mathrm{e}^{-(1+\theta)(z_1+z_2)}}{(1-\theta)^2}\, d\boldsymbol{z} \\ &= \frac{d_3}{2(1-\theta)^2(1+\theta)},\end{aligned}$$

where

$$d_3 = \int_0^{(\gamma/2)^a} \mathrm{e}^{2\gamma^a}\,\mathrm{e}^{-(1+\theta)z}\left\{\mathrm{e}^{-(1+\theta)(\gamma - z^{1/a})^a} - \mathrm{e}^{-(1+\theta)\gamma^a}\right\} dz > 0\,.$$

For fixed z and $\theta = 1 - 2\gamma^{-a}$ write the integrand of d_3 as $\mathrm{e}^{-(1+\theta)z}g(z,\gamma)$, where

$$\begin{aligned}g(z,\gamma) &= \mathrm{e}^{2\gamma^a}\left\{\mathrm{e}^{-(2-2\gamma^{-a})\gamma^a\left(1 - z^{1/a}/\gamma\right)^a} - \mathrm{e}^{-(2-2\gamma^{-a})\gamma^a}\right\} \\ &= \exp\left\{\gamma^a\left[2 - 2\left(1 - \frac{z^{1/a}}{\gamma}\right)^a\right] + 2\left(1 - \frac{z^{1/a}}{\gamma}\right)^a\right\} - \mathrm{e}^2\end{aligned}$$

decreases monotone to 0 as $\gamma \to \infty$. By the monotone convergence theorem, it follows that $d_3 \to 0$ as well, as $\gamma \to \infty$. Hence, we have $D_3/\ell^2 = o\left(\gamma^{2a}\right)$.

Concluding, for $a < 1$ we have proved that with the exponential twist $\theta = 1 - 2\gamma^{-a}$ the SCV of the TLR estimator increases proportionally to γ^{2a}, as $\gamma \to \infty$, that is

$$\kappa^2(\gamma) = O(\gamma^{2a}) \text{ as } \gamma \to \infty. \tag{85}$$

It is interesting to note that κ^2 decreases with a, that is as the tail of Weibull pdf becomes heavier.

We conjecture that for arbitrary n the optimal twisting parameter is asymptotically $\theta^* \approx 1 - n\gamma^{-a}$ and that the SCV increases proportionally to γ^{na}, as $\gamma \to \infty$.

Acknowledgement We would like to thank Rostislav Man from the Technion for performing most of the computational part of this work. We would also like to thank Pieter-Tjerk de Boer for implementing and running the simulations for the tandem system.

References

[1] S. Asmussen. *Applied probability and queues.* John Wiley and Sons, 1987.

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

[3] S. Asmussen, K. Binswanger, and B. Højgaard. Rare events simulations for heavy-tailed distributions. *Bernoulli*, 6:303 – 322, 2000.

[4] S. Asmussen, D.P. Kroese, and R.Y. Rubinstein. Complexity result for rare event simulation with heavy tails. In preparation.

[5] S. Asmussen and R.Y. Rubinstein. Complexity properties of steady-state rare-events simulation in queueing models. In *Advances in Queueing: Theory, Methods and Open Problems*, pages 429–462. CRC Press, 1995.

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

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

[8] P. Embrechts and N. Veraverbeke. Estimates for the probability of ruin with special empahsis on the possibility of large claims. *Insurance Mathematics and Economics*, 1:55–72, 1982.

[9] M.J.J. Garvels and D.P. Kroese. A comparison of RESTART implementations. In *Proceedings of the 1998 Winter Simulation Conference*, pages 601–609, Washington, DC, 1998.

[10] M.J.J. Garvels, D.P. Kroese, and J.C.W. van Ommeren. On the importance function in RESTART simulation. *European Transactions on Telecommunications*, 13(4), 2002.

[11] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. A look at multilevel splitting. In H. Niederreiter, editor, *Monte Carlo and Quasi Monte Carlo Methods 1996, Lecture Notes in Statistics*, volume 127, pages 99–108. Springer Verlag, 1996.

[12] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. A large deviations perspective on the efficiency of multilevel splitting. *IEEE Transactions on Automatic Control*, 43(12):1666–1679, 1998.

[13] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Multilevel splitting for estimating rare event probabilities. *Operations Research*, 47(4):585–600, 1999.

[14] P.W. Glynn and D.L. Iglehart. Importance sampling for stochastic simulations. *Management Science*, 35:1367–1392, 1989.

[15] C. Görg. Simulating rare event details of ATM delay time distributions with RESTART/LRE. In *Proceedings of the RESIM Workshop*. University of Twente, The Netherlands, March 1999.

[16] C. Görg and O. Fuß. Comparison and optimization of RESTART run time strategies. *AEÜ*, 52(3):197–204, 1998.

[17] Z. Haraszti and J. Townsend. Rare event simulation of delay in packet switching networks using DPR-based splitting. In *Proceedings of the RESIM Workshop, 11-12 March 1999*, pages 185–190. University of Twente, the Netherlands, 1999.

[18] P. Heidelberger. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation*, 5:43–85, 1995.

[19] S. Juneja and P. Shahabuddin. Simulating heavy tailed processes using delayed hazard rate twisting. *ACM Transactions on Modeling and Computer Simulation*, 12:94–118, 2002.

[20] H. Kahn and T.E. Harris. *Estimation of Particle Transmission by Random Sampling.* National Bureau of Standards Applied Mathematics Series, 1951.

[21] I. Kovalenko. Approximations of queues via small parameter method. In J.H. Dshalalow, editor, *Advances in Queueing: Theory, Methods and Open Problems*, pages 481 – 509. CRC Press, New York, 1995.

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

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

[24] R.Y. Rubinstein and B. Melamed. *Modern simulation and modeling.* Wiley series in probability and Statistics, 1998.

[25] R.Y. Rubinstein and A. Shapiro. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization via the score function method.* Wiley, 1993.

[26] J. S. Sadowsky. On the optimality and stability of exponential twisting in Monte Carlo simulation. *IEEE Trans. Info. Theory*, IT–39:119–128, 1993.

[27] F. Schreiber and C. Görg. A modified RESTART method using the LRE-algorithm. In North Holland, editor, *Proceedings of the 14th International Teletraffic Congress*, pages 787 – 796, 1994.

[28] F. Schreiber and C. Görg. The RESTART/LRE method for rare event simulation. In *Proceedings of the 1996 Winter Simulation Conference*, pages 390–397, 1996.

[29] P. Shahabuddin. Rare event simulation of stochastic systems. In *Proceedings of the 1995 Winter Simulation Conference*, pages 178–185, Washington, D.C., 1995. IEEE Press.

[30] D. Siegmund. Importance sampling in the Monte Carlo study of sequential tests. *Annals of Statistics*, 4:673–684, 1976.

[31] M. Villén-Altamirano and J. Villén-Altamirano. RESTART: A method for accelerating rare event simulations. In J.W. Cohen and C.D. Pack, editors, *Proceedings of the 13th International Teletraffic Congress, Queueing, Performance and Control in ATM*, pages 71–76, 1991.

[32] M. Villén-Altamirano and J. Villén-Altamirano. About the efficiency of RESTART. In *Proceedings of the RESIM'99 Workshop*, pages 99–128. University of Twente, the Netherlands, 1999.