Proceedings of the 2011 Winter Simulation Conference
S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, eds.

EFFICIENT RARE EVENT SIMULATION FOR HEAVY-TAILED SYSTEMS VIA CROSS ENTROPY

Jose Blanchet

IEOR Department
Columbia University
New York, NY 10027, USA

Yixi Shi

IEOR Department
Columbia University
New York, NY 10027, USA

ABSTRACT

The cross entropy method is a popular technique that has been used in the context of rare event simulation in order to obtain a good selection (in the sense of variance performance tested empirically) of an importance sampling distribution. This iterative method requires the selection of a suitable parametric family to start with. The selection of the parametric family is very important for the successful application of the method. Two properties must be enforced in such a selection. First, subsequent updates of the parameters in the iterations must be easily computable and, second, the parametric family should be powerful enough to approximate, in some sense, the zero-variance importance sampling distribution. We obtain parametric families for which these two properties are satisfied for a large class of heavy-tailed systems including Pareto and Weibull tails. Our estimators are shown to be strongly efficient in these settings.

1 INTRODUCTION

Tail probabilities of sums of heavy-tailed increments are a fundamental problem in the applied probability field. A large number of applications boils down to these building blocks. In this paper we focus our attention on the tail probabilities of a finite sum of heavy-tailed random variables, and we propose a method to improve variance reduction of an existing class of estimators with proved efficiency.

Let $S_m = X_1 + X_2 + \dots + X_m$ be a sum of independently and identically distributed (i.i.d.) random variables, with $S_0 = 0$ and that the X_n's are suitably heavy-tailed. The primary interest is the design of efficient estimators for the tail probability of the sum

$$u(b) = \mathbb{P}(S_m > b).$$

The basic intuition behind the construction of efficient importance sampling estimators is that one should mimic the behavior of the zero variance change of measure, which coincides with the conditional distribution

$$\mathbb{P}(S \in \cdot | S_m > b) \tag{1}$$

(see for example, Asmussen and Glynn 2008). Therefore, the behavior of the heavy tailed random walk conditional on the rare event becomes the target to be tracked by paths generated under the importance sampling distribution. It is well known from the theory of heavy-tailed large deviations that this "target" is characterized by the so-called "principle of big jump", which states that as $b \nearrow \infty$ the rare event occurs due to the contribution of a single large increment of size $\Omega(b)$ (For non-negative $f(\cdot)$ and $g(\cdot)$ we adopt the notations 1) $f(b) = \mathcal{O}(g(b))$ if $f(b) \leq cg(b)$ for some $c > 0$, 2) $f(b) = \Omega(g(b))$ if $f(b) \geq cg(b)$, and 3) $f(b) = o(g(b))$ as $b \nearrow \infty$ if $f(b)/g(b) \to 0$ as $b \nearrow \infty$.). On the other hand, paths with more than one jumps of order $\Omega(b)$ shall not be neglected in the construction of importance sampler, because of an observation pointed out by Binswanger and Hojgaard. (1997) that the second moment of the estimator for heavy tailed large deviation probabilities is very much sensitive to the likelihood ratio of these paths (see also Example 1 in Section 2).

978-1-4577-2109-0/11/$26.00 ©2011 IEEE

Guided by these observations, it is natural to suggest a mixture based sampler for the increments as the candidate importance sampler. Recently several state-dependent importance sampling estimators based on such mixtures (Dupuis, Leder, and Wang 2006 and Blanchet and Liu 2011) have been developed and shown to be strongly efficient (which means that the number of samples needed to achieve a fixed relative precision is bounded as $b \nearrow \infty$). In simple words, one samples the next increment from different regions of its support with different probabilities. We shall delay the specific form of the mixture to the next section.

Since the zero variance change of measure (1), optimal among all possible sampling distribution, involves the unknown quantity of interest $u(b)$ and is therefore infeasible, the search of global optimal sampling distribution is a futile attempt. But if one restricts optimization within a specific parametric family of sampler, there is hope that an improved change of measure within that family can be obtained. One powerful tool that exactly fits into this setting is *Cross Entropy (CE) minimization* (see for example, Rubinstein and Kroese 2004 and Kroese, Rubinstein, and Glynn 2010). Instead of directly minimizing the variance of the estimator, the CE method minimizes the cross-entropy discrepancy between two densities. The main advantage of the CE method is that, if the parametric family is well chosen, the optimization problem often admits closed-form solutions, as opposed to the variance minimization (VM) method (we refer readers to Chan, Glynn, and Kroese (2011) for an in-depth comparison between these two methods).

The successful application of the CE method is closely tied to the quality of the selected parametric family of densities to start with. Two properties must be enforced in such a selection. First, the parametric family should be powerful enough to approximate, in some sense, the zero-variance importance sampling distribution and, second, subsequent updates of the parameters in the iterations must be easily computable. We shall focus on elaborating these properties on the mixture family of our choice in this paper and demonstrate empirically the performance of this approach applied to the mixture family. We noticed that in existing works, the application of the CE method on estimating tail probabilities of sums of heavy-tailed random variables has been restricted to importance sampling densities that do not capture the "principle of big jump"; for example Chan, Glynn, and Kroese (2011) and Blanchet, Chan, and Kroese (2010) considered importance sampling densities by tilting the scale parameters of the Weibull and log-normal increment distributions, respectively. As expected, the corresponding estimators are asymptotically efficient in a weak sense, as opposed to the strong efficiency criterion that our proposed family satisfies (see Theorem 1 below). Our contribution of this paper is to justify the applicability of the CE method to a parametric family of densities that capture the large deviations behavior of the heavy-tailed sum, and the resulting estimator is *strongly efficient*.

The rest of the paper is organized as follows. In Section 2 we introduce the assumptions for the heavy-tailed increments, and put forward the parametric family of importance sampling densities to work on. Section 3 justifies the preservation of strong efficiency when switching among the same parametric mixture family. In Section 4 the CE method is reviewed and we discuss how it can be applied to the mixture family under consideration, after which the iterative equations are derived in closed-form. Finally in Section 5 we test the performance of our approach on two examples and give further discussions.

2 ASSUMPTIONS, NOTATIONS AND PARAMETRIC FAMILY OF IS DISTRIBUTIONS

2.1 Heavy-tailed Increment Distributions

Families of heavy-tailed distributions used in practice include regularly varying (Pareto-type tails) Weibull and log-normal. Our two sets of assumptions, discussed next, encompass virtually all models used in practice. We assume the increment distribution satisfies *either* of the following two Assumptions.

Assumption 1 F has a regularly varying right tail with index $\alpha > 1$, i.e.,

$$\bar{F}(x) = 1 - F(x) = L(x)x^{-\alpha},$$

where $L(\cdot)$ is a slowly varying function at infinity, that is, $\lim_{x\to\infty} L(xt)/L(x) = 1$.

Assumption 2 There exists b_0 such that for all $x > b_0$ the following conditions hold.

2a $\lim_{x\to\infty} x\lambda(x) = \infty$.
2b There exists $\beta_0 \in (0,1)$ such that $\partial \log\Lambda(x) = \lambda(x)/\Lambda(x) \leq \beta_0 x^{-1}$ for $x \geq b_0$.
2c $\Lambda(\cdot)$ is concave for all $x \geq b_0$; equivalently, $\lambda(\cdot)$ is assumed to be non-increasing for $x \geq b_0$.

We remark that under Assumption 2, the increment distribution F is essentially assumed to possess a tail at least as heavy as some Weibull distribution with shape parameter $\beta_0 < 1$. Note that under these Assumptions, adopted from Blanchet and Liu (2011), the increments X_i's are *subexponential*, which means that

$$\mathbb{P}(S_m > b) \sim m\mathbb{P}(X_i > b),$$

as $b \nearrow \infty$ (see Lemma 6 of Blanchet and Liu 2011).

2.2 Parametric Family of IS Distributions

State-dependent importance sampler (SDIS) is designed to sample the increments of the system from a distribution that is dependent on the current status of the system being simulated. We consider a mixture based SDIS. Let us denote by $\underline{p}_j = (p_{j,0},...,p_{j,K})$ the vector of mixture probabilities applied to the jth increment, $j = 1,2,...,m-1$, where $K+2$ is the number of mixture determined by the heaviness of the tail (the lighter the tail is, the larger K is). We consider the following family of mixture based densities parameterized by the mixing probabilities

$$\mathbf{p} = \{\underline{p}_1, \underline{p}_2, ..., \underline{p}_{m-1}\} = \{(p_{1,0}, p_{1,1}, ..., p_{1,K}), ..., (p_{m,0}, p_{m,1} ..., p_{m,K})\}$$

where $K \geq 0$, from which we sample the kth increment of the heavy-tailed system:

$$h_k\left(x; \underline{p}_k | S_{k-1} = s\right) = p_{k,0} f_0(x|s) + \sum_{j=1}^{K} p_j f_j(x|s) + \left(1 - \sum_{j=0}^{K} p_j\right) f_\dagger(x|s),$$

where $f_\dagger$ and f_j for $j = 0,1,...,K$ are properly normalized density functions, which have disjoint supports and depend on the current position of the system $S_{k-1} = s$. One can think of the mixture as a mechanism to control the magnitude of the increments based on evaluations of the current status of the system, and therefore it's a natural choice in order to induce the "principle of big jump" in the sampled paths. The two prevalent specifications are from Dupuis, Leder, and Wang (2006) and Blanchet and Liu (2011). The former works for random walks with increments of regularly varying-type tails that satisfy Assumption 1, in which case a mixture of two is used, i.e., $K = 0$. In particular,

$$h_k(x|s) = \left(\frac{I(x > a(b-s))}{\bar{F}(a(b-s))} + \frac{I(x \leq a(b-s))}{F(a(b-s))}\right) f(x),$$

where $a \in (0,1)$ is necessary for analytical reasons and is typically set to be close to 1.

For increments that have distributions covered by Assumption 2, for example Weibull, estimators based on two mixtures might fail to achieve bounded relative error. As discussed in the previous section, this is because the weight of the contribution of those "rogue" paths (i.e., paths with multiple jumps of order $\Omega(b)$) to the relative variance of the estimator is growing increasingly pronounced. Consider the following example.

Example 1 Suppose we are interested in estimating $\mathbb{P}(X_1 + X_2 > b)$, where X_1, X_2 are i.i.d. Weibull with parameter $\beta \in (0,1)$, i.e., $\mathbb{P}(X_i > t) = \bar{F}(t) = \exp\left(-t^\beta\right)$. Note that $\mathbb{P}(X_1 + X_2 > b) \sim \mathbb{P}(X_1 > b) + \mathbb{P}(X_2 > b)$ due to the properties of subexponential distributions. A two-mixture sampler leads to the following importance sampling strategy: sample the increments

$$(Y_1,Y_2)=\begin{cases}\left(X_1,X_2\middle|\left(X_1;\ X_2>b-X_1\right)\right) & w.p.1/2\\ \left(X_1\middle|\left(X_2;\ X_1>b-X_2\right),X_2\right) & w.p.1/2.\end{cases}$$

The corresponding IS estimator is therefore

$$\hat{\mu}_b=\frac{f_{X_1}(y_1)f_{X_2}(y_2)}{f_{X_1,X_2}(y_1,y_2)}=\frac{2\bar{F}(b-y_1)\bar{F}(b-y_2)I(y_1+y_2>b)}{\bar{F}(b-y_1)+\bar{F}(b-y_2)}.$$

It's not hard to see that for some choice of $\beta<1$, the relative error is unbounded as $b\nearrow\infty$. In particular, consider the path $(y_1,y_2)=(b/2,b/2)$, one has

$$\begin{aligned}\frac{\mathbb{E}\left(\hat{\mu}_b^2\right)}{\mathbb{P}\left(X_1+X_2>b\right)^2}&=\frac{\mathbb{E}_{\mathbf{p}}\left(\hat{\mu}_b\right)}{\mathbb{P}\left(X_1+X_2>b\right)^2}\\&\geq\frac{1}{\mathbb{P}\left(X_1+X_2>b\right)^2}\frac{f_{X_1}(b/2)f_{X_2}(b/2)}{f_{Y_1,Y_2}(b/2,b/2)}f_{X_1}(b/2)f_{X_2}(b/2)\\&=\frac{\bar{F}(b/2)^2f_{X_1}(b/2)^2}{\mathbb{P}\left(X_1+X_2>b\right)^2\bar{F}(b/2)}\approx\frac{\exp\left(-3\left(b/2\right)^{\beta}+2b^{\beta}\right)}{4},\end{aligned}$$

which grows rapidly as $b\nearrow\infty$ if e.g., $\beta=2/3$.

As the previous example illustrates, more mixtures are needed for the increments covered by Assumption 2 to absorb the impact of such "rogue" paths on the second moment of the estimator. Following this observation, Blanchet and Liu (2011) proposed a multi-point mixture family, which is general enough to cover all the increment types that satisfy Assumption 1 and Assumption 2. The support of the mixture based densities is defined in terms of the hazard function of the increments, and the number of mixtures used is dependent on the tail heaviness of the increments which is expressed in terms of the concavity of the hazard function of the increment distribution. More mixtures are needed when the tails are not as heavy as regularly varying, for example Weibull. More precisely, let $\Lambda(x)=-\log\bar{F}(x)$ be the integrated hazard function of the increments, given $a_*,a_{**}>0$, let

$$f_0(x|s)=f(x)\frac{I\left(x\leq b-s-\Lambda^{-1}\left(\Lambda(b-s)-a_*\right)\right)}{\mathbb{P}\left(x\leq b-s-\Lambda^{-1}\left(\Lambda(b-s)-a_*\right)\right)},$$

and

$$f_{\dagger}(x|s)=f(x)\frac{I\left(x>b-s-\Lambda^{-1}\left(\Lambda(b-s)-a_{**}\right)\right)}{\mathbb{P}\left(x>b-s-\Lambda^{-1}\left(\Lambda(b-s)-a_{**}\right)\right)}.$$

The densities f_j's are defined by a set of cut-off points $c_j=a_j(b-s)$ for $j=1,2,...,K-1$ where $0<a_1<a_2<...<a_{K-1}<1$ is a sequence satisfying, for given $\beta_0\in(0,1)$ and a positive constant σ_1,

$$a_j^{\beta}+(1-a_{j+1})^{\beta}\geq1+\sigma_2,$$

and

$$a_{j+1}-a_j\leq\sigma_1/2,$$

for each $1\leq j\leq k-2$ for some $\sigma_2>0$, and $a_{k-1}\geq1-\sigma_1,a_1\leq\sigma_1$. Set $c_0=b-s-\Lambda^{-1}\left(\Lambda(b-s)-a_*\right)$ and $c_K=b-s-\Lambda^{-1}\left(\Lambda(b-s)-a_{**}\right)$ we define

$$f_j(x)=\begin{cases}f(x)I\left(x\in(c_{j-1},c_j]\right)/\mathbb{P}\left(X\in(c_{j-1},c_j]\right) & 1\leq j\leq K-1\\ f(b-s-x)I\left(x\in(c_{K-1},c_K]\right)/\mathbb{P}\left(X\in(b-s-c_K,b-s-c_{K-1}]\right) & j=K\end{cases}$$

for $j = 1,2,...,K$. Note that the two specifications of the mixtures (by Dupuis, Leder, and Wang 2006 and Blanchet and Liu 2011) have the same spirits when the increments are regularly varying (see equation (14) in Blanchet and Liu 2011). Blanchet and Liu (2011) also showed that this mixture based distribution converges in total variation to the zero-variance distribution in a certain random walk problem, as $b \nearrow \infty$. In what follows, we shall work on a more general form of the mixture given as follows

$$h_k\left(x;\underline{p}_k|S_{k-1}=s\right) = \left(\sum_{j=0}^{K} p_{k,j} I\left(A_j\left(s\right)\right) w_j\left(s,x\right) + \left(1-\sum_{j=0}^{K} p_{k,j}\right) I\left(A_{\dagger}(s)\right) w_{\dagger}\left(s,x\right)\right) f(x),$$

where $A_{\dagger}(s) = \overline{\bigcup_{j=0}^{K} A_j}$, and $w_j\left(s,x\right), w_{\dagger}\left(s,x\right) > 0$ satisfy $\mathbb{E}\left(w_j\left(s,X\right)\right) = \mathbb{E}\left(w_{\dagger}\left(s,X\right)\right) = 1$. Note that the mixture family specified by Dupuis, Leder, and Wang (2006) corresponds to setting

$$w_j\left(s,x\right) = \frac{I\left(x > a(b-s)\right)}{\bar{F}\left(a(b-s)\right)},$$

for $j = 0,\dagger$; and the one proposed by Blanchet and Liu (2011) corresponds to setting

$$w_j\left(s,x\right) = \frac{I\left(A_j(s)\right)}{\mathbb{P}\left(A_j(s)\right)} = \frac{I\left(x \in (c_{j-1},c_j]\right)}{\mathbb{P}\left(x \in (c_{j-1},c_j]\right)},$$

for $j = 0,1,...,K-1,\dagger$ and $c_{-1} = -\infty$ with a slight abuse of notation. And

$$w_K\left(s,x\right) = \frac{f(b-s-x) I\left(x \in (c_{K-1},c_K]\right)}{f(x)\mathbb{P}\left(X \in (b-s-c_K, b-s-c_{K-1}]\right)}.$$

If we write the joint density of the increments under the original measure as

$$\mathbf{f}\left(\mathbf{x}\right) = f\left(x_1\right) f\left(x_2\right) ... f\left(x_m\right),$$

where $\mathbf{x} = (x_1,...,x_m)$, we can express the joint importance sampling density for the mixture based SDIS as

$$\begin{aligned} h\left(\mathbf{x};\mathbf{p}\right) = \prod_{k=1}^{m-1} &\left(\sum_{j=0}^{K} p_{k,j} I\left(A_j\left(s_{k-1}\right)\right) w_j\left(s,x_k\right) + \left(1-\sum_{j=0}^{K} p_{k,j}\right) I\left(A_{\dagger}(s_{k-1})\right) w_{\dagger}\left(s,x_k\right)\right) \\ &\cdot \left(I\left(S_{m-1} < b\right)\mathbb{P}\left(X_m > (b-S_{m-1})\right) + I\left(S_{m-1} \geq b\right)\right)\mathbf{f}\left(\mathbf{x}\right). \end{aligned}$$

3 STRONG EFFICIENCY OF THE FAMILY UNDER CONSIDERATION

The following Theorem highlights the main reason leading to the strong efficiency of the mixture family. The proof, which relates to the techniques studied in Dupuis, Leder, and Wang (2006) and Blanchet and Liu (2011), is given in Blanchet and Shi (2011).

Theorem 1 Let $\mathbb{P}$ and $\mathbb{P}_{\mathbf{p}}$ be the original probability measure and the one induced by the mixture family with mixing probability vector $\mathbf{p}$. If there exists an $\varepsilon > 0$ such that $\mathbf{p} > \varepsilon \cdot \mathbf{1}$, for all $b > 0$, where $\mathbf{1}$ is a vector of ones of dimension $(m-1) \times (K+1)$, then

$$\frac{d\mathbb{P}}{d\mathbb{P}_{\mathbf{p}}} \frac{I\left(S_m > b\right)}{\mathbb{P}\left(S_m > b\right)} = \mathscr{O}\left(1\right),$$

as $b \to \infty$.

The result enables us to comfortably switch to different choices of mixing probabilities within the same parametric family without violating the strong efficiency property of the final estimator, which lays the ground for the applicability of the CE method to be introduced shortly.

4 CROSS ENTROPY METHOD AND THE ITERATIVE EQUATIONS FOR THE MIXTURE FAMILY

4.1 Review of Cross-Entropy Method

If we restrict our search of importance sampler to this particular parametric class, the optimal choice of the vector $\mathbf{p}$ can be obtained by minimizing the so-called *Kullback-Leibler divergence* or the *cross-entropy distance*.

Definition 1 The Kullback-Leibler cross-entropy between two densities g and h is given by

$$\begin{aligned}\mathscr{D}(g,h) &= \int g(\mathbf{x})\log\frac{g(\mathbf{x})}{h(\mathbf{x})}d\mathbf{x}\\ &= \int g(\mathbf{x})\log g(\mathbf{x})d\mathbf{x} - \int g(\mathbf{x})\log h(\mathbf{x})d\mathbf{x}.\end{aligned}$$

If we fix g to be the optimal importance sampling density $g^*(\mathbf{x}) \propto \varphi(S(\mathbf{x};b))f(\mathbf{x})$, where $\varphi(S(\mathbf{x};b))$ is the performance measure of the system (for example, $S(\mathbf{X}) = \sum_{j=1}^{m} X_j$, and $\varphi(S(\mathbf{x};b)) = I(S(\mathbf{x}) > b)$), then our search of the optimal mixture is the output of the following *parametric* optimization problem

$$\begin{aligned}\min_{\mathbf{p}}\mathscr{D}(g^*,h(\cdot,\mathbf{p})) \Longleftrightarrow \max_{\mathbf{p}} D(\mathbf{p}) &= \max_{\mathbf{p}}\mathbb{E}_{\mathbf{p}^\star}\varphi(S(\mathbf{X};b))\log h(\mathbf{X};\mathbf{p})\\ &= \max_{\mathbf{p}}\mathbb{E}_{\tilde{\mathbf{p}}}\varphi(S(\mathbf{X};b))\frac{h(\mathbf{X};\mathbf{p}^\star)}{h(\mathbf{X};\tilde{\mathbf{p}})}\log h(\mathbf{X};\mathbf{p})\\ &= \max_{\mathbf{p}}\mathbb{E}_{\tilde{\mathbf{p}}}\varphi(S(\mathbf{X};b))\frac{\mathbf{f}(\mathbf{X})}{h(\mathbf{X};\tilde{\mathbf{p}})}\log h(\mathbf{X};\mathbf{p}), \qquad (2)\end{aligned}$$

where $\mathbf{f}(\mathbf{X})/h(\mathbf{X};\tilde{\mathbf{p}})$ is the likelihood ratio between the original measure and the measure induced by the mixture based density with some fixed parameter $\tilde{\mathbf{p}}$ (Recall that $\mathbf{X} = (X_1,...,X_m)$). In particular,

$$\begin{aligned}\frac{\mathbf{f}(\mathbf{X})}{h(\mathbf{X};\tilde{\mathbf{p}})} = &\prod_{k=1}^{m-1}\left(\sum_{j=0}^{K}\frac{I(x_k\in A_j(S_{k-1}))}{\tilde{p}_{k,j}w_j(S_{k-1},x_k)} + \frac{I(x_k\in A_\dagger(S_{k-1}))}{\left(1-\sum_{j=0}^{K}\tilde{p}_{k,j}\right)w_\dagger(S_{k-1},x_k)}\right)\\ &\cdot\left(I(S_{m-1}<b)\mathbb{P}(X_m > (b-S_{m-1})) + I(S_{m-1}\geq b)\right). \qquad (3)\end{aligned}$$

In most cases the expectation in (2) is analytically inaccessible. Rubinstein and Kroese (2004) suggested a recursive method based on the following stochastic counterpart of (2)

$$\max_{\mathbf{p}}\hat{D}(\mathbf{p}) = \max_{\mathbf{p}}\frac{1}{N}\sum_{i=1}^{N}\varphi(S(\mathbf{X}(i));b)\frac{\mathbf{f}(\mathbf{X}(i))}{h(\mathbf{X}(i);\tilde{\mathbf{p}})}\log h(\mathbf{X}(i),\mathbf{p}). \qquad (4)$$

Cross Entropy (CE) Algorithm (Rubinstein and Kroese 2004)

1. Choose an initial vector of mixing probabilities $\mathbf{p}^{(0)}$. Set $T=1$.
2. Generate a random sample $\mathbf{X}_1,...,\mathbf{X}_N$ from the joint density $h\left(\cdot;\mathbf{p}^{(T-1)}\right)$.
3. Solve the stochastic optimization program (4). Denote the solution by $\mathbf{p}^{(T)}$, i.e.,

$$\mathbf{p}^{(T)} = \arg\min_{\mathbf{p}}\frac{1}{N}\sum_{i=1}^{N}\varphi(S(\mathbf{X}(i));b)\frac{\mathbf{f}(\mathbf{X}(i))}{h\left(\mathbf{X}(i);\mathbf{p}^{(T-1)}\right)}\log h(\mathbf{X}(i),\mathbf{p}).$$

4. **Stop** if convergence is reached; otherwise, set $T=T+1$, go to Step 2.

It's very convenient to embed the CE algorithm in the main SDIS algorithm to further reduce variance. Let M be the total simulation budget, and τ be the number of recursions in the CE algorithm until convergence of $\mathbf{p}$. If $\tau N < M$, then the SDIS with CE algorithm add-on corresponds to generating τ batches of independent samples from the mixture based importance sampling density parameterized by $\mathbf{p}^{(T)}$, for $T = 0, 1, ..., \tau - 1$, and one batch of size $M - \tau N$ of independent samples from the importance density with optimal CE probability vector $\mathbf{p}^*$. Depending on the size of $M - \tau N$, the final estimator can be obtained by averaging either the last batch of $M - \tau N$ samples, or the entire M samples from different batches. In either case we are able to achieve variance reduction while maintaining strong efficiency property. Even for the case where $\tau N \geq M$, the improved cross-entropy after each iteration typically will reduce the variance of the future samples over those from previous iterations, since each iteration gives us a parameterized density closer to the zero-variance importance density.

4.2 Iterative Equations for the Mixture IS Family

We now proceed to characterize the solution to (4). In the case where we are interested in the tail probability of the sum $\mathbb{P}(S_m > b)$, $\varphi(S(\mathbf{X}); b) = I(S_m > b)$. Note that $\hat{D}$ is concave and differentiable with respect to the components p_k, therefore the solution to (4) is directly given by the first order optimality condition:

$$\sum_{i=1}^{N} I(S_m(i) > b) \frac{\mathbf{f}(\mathbf{X}(i))}{h(\mathbf{X}(i); \tilde{\mathbf{p}})} \bigtriangledown_{\mathbf{p}} \log h(\mathbf{X}(i), \mathbf{p}) = 0. \tag{5}$$

The product structure of the likelihood function is particularly useful because the sensitivity of the likelihood function to the mixing probabilities can be localized. Indeed, a few lines of elementary algebra gives

$$\begin{aligned} \frac{d \log h(\mathbf{X}, \mathbf{p})}{d p_{k,l}} &= \left(I(X_k \in A_l(S_{k-1})) w_l(S_{k-1}, X_k) - I(X_k \in A_\dagger(S_{k-1})) w_\dagger(S_{k-1}, X_k)\right) / \\ &\quad \left(\sum_{j=0}^{K} p_{k,j} I(X_k \in A_j(S_{k-1})) w_j(S_{k-1}, X_k) + \left(1 - \sum_{j=0}^{K} p_{k,j}\right) I(X_k \in A_\dagger(S_{k-1})) w_\dagger(S_{k-1}, X_k) \right) \\ &= \frac{I(X_k \in A_l(S_{k-1}))}{p_{k,l}} - \frac{I(X_k \in A_\dagger(S_{k-1}))}{1 - \sum_{j=0}^{K} p_{k,j}}. \end{aligned}$$

We denote

$$W(\mathbf{X}_{-l}(i); \mathbf{p}^\star, \tilde{\mathbf{p}}) = \prod_{k=1, k \neq l}^{m-1} \frac{h\left(X_k(i); \underline{p}_k^\star\right)}{h\left(X_k(i); \underline{\tilde{p}}_k\right)} \left(I(S_{m-1} < b) \mathbb{P}(X_m(i) > (b - S_{m-1}(i))) + I(S_{m-1}(i) \geq b)\right),$$

where $\underline{p}_k^\star = \{p_{k,0}^\star, ... p_{k,K}^\star\}$, and $\underline{\tilde{p}}_k = \{\tilde{p}_{k,0}, ... \tilde{p}_{k,K}\}$. And further let

$$\Theta_{l,j} = \frac{\sum_{i=1}^{N} W(\mathbf{X}_{-l}(i); \mathbf{p}^\star, \tilde{\mathbf{p}}) \left(1 - \sum_{j=0}^{K} \tilde{p}_{l,j}\right) w_\dagger(S_{l-1}, X_l(i))}{\sum_{i=1}^{N} W(\mathbf{X}_{-l}(i); \mathbf{p}^\star, \tilde{\mathbf{p}}) \tilde{p}_{l,j} w_l(S_{l-1}, X_l(i))}.$$

The first order optimality condition (5) therefore yields the following solution $\mathbf{p}^*$ to the stochastic optimization problem (4), we shall call this vector of optimal solution *optimal CE mixing probability vector*:

$$p_{l,j}^* = \frac{\Theta_{l,j}}{1 + \sum_{k=0}^{K} \Theta_{k,j}}, \tag{6}$$

for $j = 0, 1, ..., K$ and $l = 1, 2, ..., m$. It doesn't take long to realize that the previous expression has the following equivalent form

$$p_{l,j}^\star = \frac{\sum_{i=1}^{N} I(S_m(i) > b) W(\mathbf{X}(i); \mathbf{p}^\star, \tilde{\mathbf{p}}) I(X_l \in A_j(S_{l-1}))}{\sum_{i=1}^{N} I(S_m(i) > b) W(\mathbf{X}(i); \mathbf{p}^\star, \tilde{\mathbf{p}})}, \tag{7}$$

for $j = 0, 1, ..., K$ and $k = 1, 2, ..., m$, where $W(\cdot; \mathbf{p}^{\star}, \tilde{\mathbf{p}}) = h(\cdot; \mathbf{p}^{\star}) / h(\cdot; \tilde{\mathbf{p}}) = \mathbf{f}(\cdot) / h(\cdot; \tilde{\mathbf{p}})$ is given by (3). It's worth pointing out that (7) is computationally advantageous over (6), because it avoids dividing by zero in computing $\Theta_{l,j}$, especially when the number of "pilot" runs is small. (Note that the sampling of the mth increment ensures $S_m(i) > b$.) Moreover, the expression (7) entails a nice interpretation: the optimal mixing probability is the proportion of the contribution to the likelihood function from the jth "band" of the kth increment.

For completeness we also include the explicit iteration equations for cases where the increments satisfy Assumption 1 and 2, respectively. We write, for ease of exposition,

$$W_m(i) = \left(I\left(S_{m-1}(i) < b\right) \mathbb{P}\left(X_m(i) > \left(b - S_{m-1}(i)\right)\right) + I\left(S_{m-1}(i) > b\right)\right).$$

For regularly varying increments, the solution for the Tth iteration of the recursive algorithm can be written as

$$p_k^{(T)} = \frac{\sum_{i=1}^{N} I\left(S_m(i) > b; X_k > a(b - s_{k-1})\right) \prod_{k=1}^{m-1} \left(\frac{\mathbb{P}(X_k > a(b-s_{k-1}))}{p_k^{(T-1)} I(X_k > a(b-s_{k-1}))} + \frac{\mathbb{P}(X_k \le a(b-s_{k-1}))}{\left(1-p_k^{(T-1)}\right) I(X_k \le a(b-s_{k-1}))} \right) W_m(i)}{\sum_{i=1}^{N} I\left(S_m(i) > b\right) \prod_{k=1}^{m-1} \left(\frac{\mathbb{P}(X_k > a(b-s_{k-1}))}{p_k^{(T-1)} I(X_k > a(b-s_{k-1}))} + \frac{\mathbb{P}(X_k \le a(b-s_{k-1}))}{\left(1-p_k^{(T-1)}\right) I(X_k \le a(b-s_{k-1}))} \right) W_m(i)}.$$

For increment distributions that satisfy Assumption 2, the likelihood function $W\left(\cdot; \mathbf{p}^{\star}, \mathbf{p}^{(T-1)}\right)$ becomes

$$\begin{aligned} & W\left(\mathbf{X}^{(T-1)}; \mathbf{p}^{\star}, \mathbf{p}^{(T-1)}\right) = \frac{\mathbf{f}\left(\mathbf{x}^{(T-1)}\right)}{h\left(\mathbf{X}^{(T-1)}, \mathbf{p}^{(T-1)}\right)} \\ & = \prod_{k=1}^{m-1} \left(\frac{\mathbb{P}\left(X_k^{(T-1)} \le c_0\right)}{p_{k,0}^{(T-1)} I\left(x_k^{(T-1)} \le c_0\right)} + \frac{\mathbb{P}\left(X_k^{(T-1)} > c_K\right)}{\left(1 - \sum_{j=0}^{K} p_{k,j}^{(T-1)}\right) I\left(X_k^{(T-1)} > c_K\right)} \right. \\ & \quad \left. + \sum_{j=1}^{K-1} \frac{\mathbb{P}\left(X_k^{(T-1)} \in (c_{j-1}, c_j]\right)}{p_{k,j}^{(T-1)} I\left(x_k^{(T-1)} \in (c_{j-1}, c_j]\right)} + \frac{f(b - s - x_k^{(T-1)}) \mathbb{P}\left(X_k^{(T-1)} \in (b - s - c_{K-1}, b - s - c_K]\right)}{p_{k,K}^{(T-1)} f(x_k^{(T-1)}) I\left(x_k^{(T-1)} \in (c_{K-1}, c_K]\right)} \right) W_m(i), \end{aligned}$$

where c_j's are the cutoff points of the "bands" and we have explicitly written out the iteration count. Note that at the beginning of iteration T, the only part that is dependent on the unknown parameters $\mathbf{p}$ in the stochastic program (4) is $\log h\left(\mathbf{X}(i), \mathbf{p}^{(T)}\right)$ and hence $\bigtriangledown_{\mathbf{p}} \log h\left(\mathbf{X}(i), \mathbf{p}^{(T)}\right)$ in the optimality condition (5); the likelihood $W\left(\cdot; \mathbf{p}^{\star}, \mathbf{p}^{(T-1)}\right)$ is a function of the probability vector passed from the $(T-1)$st iteration as well as the samples generated from IS density specified by that probability vector. In that regard at the beginning of the Tth iteration, all the ingredients in the expression above are available. The iteration equation for the probability vector at iteration T is therefore given by

$$p_{k,j}^{(T)} = \frac{\sum_{i=1}^{N} I\left(S_m^{(T-1)}(i) > b\right) W\left(\mathbf{X}^{T-1}(i); \mathbf{p}^{\star}, \mathbf{p}^{(T-1)}\right) I\left(x_k^{(T-1)} \in (c_{j-1}, j_k]\right)}{\sum_{i=1}^{N} I\left(S_m^{(T-1)}(i) > b\right) W\left(\mathbf{X}(i)^{(T-1)}; \mathbf{p}^{\star}, \mathbf{p}^{(T-1)}\right)},$$

where $c_{-1} = -\infty$ with a slight abuse of notations.

Note that the iterative equations given so far reveal the ease of implementation of the CE subroutine: one only needs to keep $K+2$ buckets, indicating whether the kth increment falls into the jth band, $j = 1, 2, ..., K+2$, and aggregate the likelihood function for each bucket. The computational cost is of the same order as a vanilla SDIS iteration without the CE routine.

Remark 1 One might consider further guiding the parametric family of samplers using large deviations ideas. For example, in the regularly varying case, one can force the probabilities to have the following structure,

$$p_k = \frac{m-k+1}{m-k} p_{k-1},$$

for $k = 2,...,M-1$, which is equivalent to $p_k = \frac{m-1}{m-k} p$, for $k = 1,2,...,m-1$. This choice reflects the intuition that the chance for the k-th increment to be a large one is roughly proportional to the inverse of the remaining steps to go. Note that this particular structure is very close to the optimal mixture found by Dupuis, Leder, and Wang (2006) using a dynamic programming argument. However, due to the global dependence on the first probability parameter p. It is not difficult to see that the CE iteration equations will involve a root finding procedure, which could increase the computational cost significantly.

5 NUMERICAL EXAMPLES

5.1 Example 1: Regularly Varying Increments

We illustrate the empirical performance of the SDIS with CE routine (SDIS-CE) by considering two examples. In the first example, the increments are regularly varying with index $\alpha = 1/2$, in particular, X_n's have tail distribution

$$\mathbb{P}(X_i > b) = (1+b)^{-1/2}.$$

Following Dupuis, Leder, and Wang (2006), given the parameters of the model, a given number of increments m and a tail parameter b, we estimate $\mathbb{P}(S_m > b)$ and the standard deviation of the estimator as follows. We simulate 20000 replications of our estimator. The estimates are obtained based on averages of the replications. This is the output of a single run. Then we produce 500 independent runs. The results displayed are the averages of the outputs of these runs. We run the experiments with two different sets of input mixing probabilities. In the first case, which we shall later refer to as the "standard choice", we consider the heuristic choice $p_k = \theta/(m-k)$ where $\theta = 0.9$. And for the second set of input we use the optimal choice of the probabilities obtained by Dupuis, Leder, and Wang (2006), i.e.,

$$p_k^* = \frac{a^{-\alpha/2}}{(m-k)a^{-\alpha/2}+1},$$

which we call the "DLW" selection. In both cases we select $a = 0.9$. The results of the experiment are reported in the Table 1 and Table 2.

From the results of Table 1 we observe that even for a reasonable choice of mixing probabilities based on large deviations intuition, the CE algorithm produces a smaller relative error. On the other hand, it is outperformed by the optimal choice of the probabilities obtained in (Dupuis, Leder, and Wang 2006), as can be seen in Table 2, one shall keep in mind, however, that in many applications, the structure of the problem doesn't allow for such analytical solutions easily. We also point out that the optimal solution from Dupuis, Leder, and Wang (2006) hinges on the assumption that b is sufficiently large for large deviations asymptotics to be valid. For smaller exceedance level b, we might expect a better performance using the CE routine, which is underpinned by the results shown in Table 3.

We have mentioned in the previous section that since the recursive CE algorithm is carried out on the pilot sample, it neglects the fact that the increments are simulated in a sequential manner, but rather treats them in an independent way. We averaged the output CE optimal probability vector over the experiments, the near identical mixing probabilities in Table 4 is in line with the expected behavior of the method that each increment has probability at roughly $1/4$ of causing the rare event.

Table 1: Performance of the SDIS-CE estimator compared to the SDIS algorithm without CE procedure where the input mixing probabilities are set to be $p_k = 0.9/(m-k)$ for $k = 1, 2, ..., m-1$.

m	b	Standard	CE	Method
4	1e + 06	3.999E-03	4.000E-03	Average Estimate
		3.148E-05	1.395E-05	Average Std. Error
		0.787%	0.349%	Avg.SE/Avg.Est (%)
	1e + 12	3.999E-06	4.000E-06	
		3.151E-08	1.403E-08	
		0.788%	0.351%	
	1e + 18	4.000E-09	4.000E-09	
		3.153E-11	1.393E-11	
		0.788%	0.348%	
25	1e + 06	2.503E-02	2.498E-02	
		1.525E-03	3.404E-04	
		6.094%	1.363%	
	1e + 12	2.496E-05	2.499E-05	
		1.518E-06	3.458E-07	
		6.082%	1.384%	
	1e + 18	2.496E-08	2.502E-08	
		1.524E-09	3.409E-10	
		6.103%	1.363%	

Table 2: Performance of the SDIS-CE estimator compared to the SDIS without CE procedure where the input mixing probabilities are set to be the optimal choice obtained in Dupuis, Leder and Wang (2006).

m	b	DLW	CE	Method
4	1e + 06	4.000E-03	4.000E-03	Average Estimate
		5.660E-06	1.374E-05	Average Std. Error
		0.141%	0.344%	Avg.SE/Avg.Est (%)
	1e + 12	4.000E-06	4.000E-06	
		5.683E-09	1.382E-08	
		0.142%	0.346%	
	1e + 18	4.000E-09	4.001E-09	
		5.691E-12	1.373E-11	
		0.142%	0.343%	
25	1e + 06	2.499E-02	2.500E-02	
		3.925E-05	1.555E-04	
		0.157%	0.622%	
	1e + 12	2.500E-05	2.500E-05	
		4.032E-08	1.567E-07	
		0.161%	0.627%	
	1e + 18	2.500E-08	2.500E-08	
		4.027E-11	1.568E-10	
		0.161%	0.627%	

Table 3: Comparison of performance between 1) SDIS using CE optimal mixing probabilities and 2) Analytical optimal mixing probabilities from Dupuis, Leder and Wang (2006), $m = 2$.

b	DLW	CE	Method
5	6.999E-01	6.999E-01	Average Estimate
	1.110E-03	5.742E-04	Average Std. Error
	0.159%	0.082%	Avg.SE/Avg.Est (%)
20	4.166E-01	4.166E-01	
	4.727E-04	4.410E-04	
	0.113%	0.106%	

Table 4: Average optimal CE .mixing probabilities, $m = 4,\ b = 10^6$.

k	1	2	3
p_k	0.248	0.253	0.251

5.2 Example 2: Weibull Increments

We now proceed to the second example where the increments are assumed to have the following Weibull-type of distribution,

$$\mathbb{P}(X > b) = e^{-2\sqrt{b+1}},$$

for $t \geq -1$. This corresponds to the case considered by Blanchet and Liu (2011), where the authors use a 5-point mixtures specified by the cut-off points $c_0 = 0.1\sqrt{b-s}, c_1 = 0.1(b-s), c_2 = 0.5(b-s), c_3 = 0.9(b-s)$ and $c_4 = b-s-0.1\sqrt{b-s}$. Since the number of cut-off points increases from the previous mixture sampler, we increase the pilot sample number to 5000; all the other algorithmic parameters (number of runs and number of replications per run) remain the same. The results of the experiments are summarized in Table 5.

Table 5: Performance of the SDIS-CE estimator compared to SDIS without CE procedurein the case of Weibull-type of increments, $m = 4$. We used $p_{k,j} = 1/(K+2)(m-k)$, for $j = 0, 1, ...K$ and $k = 1, 2, ..., m-1$ as the "standard" choice of the mixing probabilities.

b	Standard	CE	Method
150	7.977E-11	7.966E-11	Avg. Est.
	2.580E-12	7.642E-13	Avg. Std. Err.
	3.235%	0.959%	Avg. SE/Avg. Est. (%)
450	1.371E-18	1.372E-18	
	4.835E-20	1.071E-20	
	3.526%	0.781%	
750	6.086E-24	6.069E-24	
	2.209E-25	3.185E-26	
	3.630%	0.525%	

REFERENCES

Asmussen, S., and P. Glynn. 2008. *Stochastic Simulation: Algorithms and Analysis*. New York, NY, USA: Springer-Verlag.

Binswanger, S. A. K., and B. Hojgaard.. 1997. "Rare events simulation for heavy-tailed distributions.". *Bernoulli* 6:303–322.

Blanchet, J., J. C. Chan, and D. Kroese. 2010. “Asymptotics and fast simulation for tail probabilities of the maximum and minimum of sums of lognormals”. working paper.

Blanchet, J., and J. Liu. 2011. “Efficient Simulation and Conditional Functional Limit Theorems for Ruinous Heavy-tailed Random Walks”. *forthcoming*.

Blanchet, J., and Y. Shi. 2011. “Efficient Rare Event Simulation for Heavy-tailed Systems via Cross Entropy”. working paper.

Chan, J. C. C., P. W. Glynn, and D. P. Kroese. 2011. “A Comparison of Cross-Entropy and Variance Minimization Strategies”. *Journal of Applied Probabilities* 48.

Dupuis, P., K. Leder, and H. Wang. 2006. “Importance sampling for sums of random variables with regularly varying tails.”. *ACM TOMACS* 17:Article 14.

Kroese, D. P., R. Y. Rubinstein, and P. W. Glynn. 2010. “The Cross-Entropy Method for Estimation”. In *Handbook of Statistics*, edited by V. Govindaraju and C. R. Rao, Volume 31. Elsevier.

Rubinstein, R. Y., and D. P. Kroese. 2004. *The Cross-Entropy Method*. New York, NY: Springer.

AUTHOR BIOGRAPHIES

JOSE BLANCHET is a faculty member of the IEOR Department at Columbia University. Jose holds a Ph.D. in Management Science and Engineering from Stanford University. Prior to joining Columbia he was a faculty member in the Statistics Department at Harvard University. Jose is a recipient of the 2009 Best Publication Award given by the INFORMS Applied Probability Society and of the 2010 Erlang Prize. He also received a PECASE award given by NSF in 2010. He worked as an analyst in Protego Financial Advisors, a leading investment bank in Mexico. He has research interests in applied probability and Monte Carlo methods. He serves in the editorial board of *Advances in Applied Probability, Journal of Applied Probability, Mathematics of Operations Research* and *QUESTA*.

YIXI SHI is a PhD candidate in Department of Industrial Engineering and Operations Research, School of Engineering and Applied Science at Columbia University. He holds a B.Sc. in Actuarial Science from University of Hong Kong. His email address is ys2347@columbia.edu.