J. Appl. Prob. Spec. Vol. **48A,** 183–194 (2011)

NEW FRONTIERS IN APPLIED PROBABILITY

A Festschrift for SØREN ASMUSSEN
Edited by P. GLYNN, T. MIKOSCH and T. ROLSKI

Part 4. Simulation

A COMPARISON OF CROSS-ENTROPY AND VARIANCE MINIMIZATION STRATEGIES

JOSHUA C. C. CHAN, *Australian National University*
Research School of Economics, Australian National University, Canberra, ACT 0200, Australia.

PETER W. GLYNN, *Stanford University*
Department of Management Science and Engineering, Huang Engineering Center, Stanford University, 475 Via Ortega, Stanford, CA 94305-4121, USA.

DIRK P. KROESE, *University of Queensland*
Department of Mathematics, University of Queensland, St Lucia, Brisbane, QLD 4072, Australia.
Email address: kroese@maths.uq.edu.au

APPLIED PROBABILITY TRUST
AUGUST 2011

A COMPARISON OF CROSS-ENTROPY AND VARIANCE MINIMIZATION STRATEGIES

BY JOSHUA C. C. CHAN, PETER W. GLYNN AND DIRK P. KROESE

Abstract

The variance minimization (VM) and cross-entropy (CE) methods are two versatile adaptive importance sampling procedures that have been successfully applied to a wide variety of difficult rare-event estimation problems. We compare these two methods via various examples where the optimal VM and CE importance densities can be obtained analytically. We find that in the cases studied both VM and CE methods prescribe the same importance sampling parameters, suggesting that the criterion of minimizing the CE distance is very close, if not asymptotically identical, to minimizing the variance of the associated importance sampling estimator.

Keywords: Variance minimization; cross entropy; importance sampling; rare-event simulation; likelihood ratio degeneracy

2010 Mathematics Subject Classification: Primary 65C05
Secondary 65C60

1. Introduction

In this article we compare two adaptive importance sampling procedures, namely the variance minimization (VM) and cross-entropy (CE) methods [11], [16, pp. 62–83], in the context of rare-event simulation. Both algorithms aim to find an importance density that is optimal in a well-defined sense, though the optimality criteria are different. Under the VM method, the optimal importance density is the one whose associated estimator has minimum variance within a given parametric family. Although this minimum variance criterion is obviously desirable, in practice the minimization problem involves application of time consuming numerical techniques. Instead of directly minimizing the variance of the estimator, the CE method seeks to locate the importance density that is closest in *Kullback–Leibler divergence* or *CE distance* to the zero-variance importance density: the conditional density given the rare event. The main advantage of the CE method is that the optimization problem required to obtain the optimal density often admits closed-form solutions.

To compare these two related but distinct algorithms, we consider various explicit examples where the optimal VM and CE importance densities can be obtained analytically. In all the examples considered, we find that the optimal VM and CE importance densities are asymptotically identical. Although whether this result holds in general or not is an open question, it suggests that the VM and CE criteria are very similar, at least asymptotically. Put differently, the importance density that is the closest—in CE distance—to the zero-variance importance density is also the one whose associated estimator has the minimum asymptotic variance. The significance of this observation is that since CE estimators are typically easier to obtain, this practical adaptive importance sampling strategy is also optimal in the sense that it gives the minimum variance importance sampling estimator. Furthermore, in situations where the VM or CE optimization problem does not admit closed-form solutions, the optimal parameters need to be estimated via a multilevel procedure. We analyze how the variability in the estimates affects the performance of the associated importance sampling estimator.

The rest of this article is organized as follows. In Section 2 we first introduce some background material, and then discuss the classic VM and CE methods as well as two variants proposed recently. We consider in Section 3 the sum of independent but not necessarily identical random variables in the exponential families where the number of parameters is sent to infinity. We show in this scenario that the optimal CE parameters coincide with those suggested by large deviation theory. It is followed by three case studies: we consider the example of the sum of exponential random variables in Section 4, and the cases for Pareto and Weibull random variables in Sections 5 and 6, respectively.

2. Adaptive importance sampling via VM and CE methods

We first introduce some standard notation and efficiency measures in the context of rare-event simulation. We write $a(t) \sim b(t)$ to indicate that $\lim_{t\to\infty} a(t)/b(t) = 1$, and $X_i \overset{\text{i.i.d.}}{=} f, i = 1, \dots, n$, to indicate that $X_1, \dots, X_n$ are independent and identically distributed (i.i.d.) according to the density or distribution f. An unbiased estimator $Z(\gamma)$ for $\ell(\gamma)$ is said to be *logarithmically efficient*, *weakly efficient*, or *asymptotically optimal* if

$$\lim_{\gamma\to\infty} \frac{\log \mathrm{E}\, Z(\gamma)^2}{\log \ell(\gamma)} = 2.$$

This condition is equivalent to the requirement that $\lim_{\gamma\to\infty} \mathrm{E}\, Z(\gamma)^2/\ell(\gamma)^{2-\varepsilon} = 0$ for every $\varepsilon > 0$. The estimator is said to be *strongly efficient* or have *bounded relative error* if $\sup_{\gamma\geq 0} \mathrm{E}\, Z(\gamma)^2/\ell(\gamma)^2 < \infty$. It is readily observed that bounded relative error implies asymptotic optimality. These notions of efficiency are standard in the literature; see, for example, [1, pp. 158–163] and [13]. We are interested in estimating the probability of the form

$$\ell = \mathrm{P}(S(\boldsymbol{X}) > \gamma) = \int \mathbf{1}(S(\boldsymbol{x}) > \gamma) f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x},$$

where S is some real-valued performance function, $\boldsymbol{X}$ is a vector of random variables with probability density function (PDF) f, and γ is a sufficiently large constant such that ℓ is small. Consider estimating ℓ via the *importance sampling* estimator

$$\widehat{\ell}_{\mathrm{IS}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(S(\boldsymbol{X}_i) > \gamma) \frac{f(\boldsymbol{X}_i)}{g(\boldsymbol{X}_i)},$$

where $\boldsymbol{X}_i \overset{\text{i.i.d.}}{=} g$, $i = 1, \dots, N$, for some importance sampling PDF g for which $g(\boldsymbol{x}) = 0$ implies that $\mathbf{1}(S(\boldsymbol{x}) > \gamma) f(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$. Although the estimator $\widehat{\ell}_{\mathrm{IS}}$ is consistent and unbiased for any such g, its performance depends critically on the choice of g. Hence, we wish to choose g so that the associated estimator is optimal in a well-defined sense. To this end, consider a parametric family $\mathcal{F} = \{f(\boldsymbol{x}; \boldsymbol{v})\}$ indexed by a parameter vector $\boldsymbol{v}$ that contains the nominal (original) density f. Thus, we can write $f(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{u})$ for some parameter vector $\boldsymbol{u}$. For any given $\boldsymbol{v}$, the general term of the associated importance sampling estimator is $Z(\boldsymbol{v}) = W(\boldsymbol{X}; \boldsymbol{u}, \boldsymbol{v})\, \mathbf{1}(S(\boldsymbol{X}) > \gamma)$, where $W(\boldsymbol{x}; \boldsymbol{u}, \boldsymbol{v})$ is the *likelihood ratio* defined as $W(\boldsymbol{x}; \boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{x}; \boldsymbol{u})/f(\boldsymbol{x}; \boldsymbol{v})$. Now, we wish to choose $\boldsymbol{v}$ so that the associated importance sampling estimator has minimum variance within the parametric family $\mathcal{F}$. The minimizer $\boldsymbol{v}_{\mathrm{vm}}$ is referred to as the *optimal VM parameter vector*. For any unbiased estimator $\widehat{\ell}$ of ℓ, we have $\operatorname{var}\widehat{\ell} = \mathrm{E}\,\widehat{\ell}^2 - \ell^2$. Therefore, $\boldsymbol{v}_{\mathrm{vm}}$ can be written as

$$\boldsymbol{v}_{\mathrm{vm}} = \underset{\boldsymbol{v}}{\operatorname{argmin}}\, \mathrm{E}_{\boldsymbol{v}}\, Z(\boldsymbol{v})^2 = \underset{\boldsymbol{v}}{\operatorname{argmin}}\, \mathrm{E}_{\boldsymbol{u}}\, Z(\boldsymbol{v}) = \underset{\boldsymbol{v}}{\operatorname{argmin}}\, \log \mathrm{E}_{\boldsymbol{u}}\, Z(\boldsymbol{v}), \tag{2.1}$$

where the expectation operators $\mathrm{E}_{\boldsymbol{u}}$ and $\mathrm{E}_{\boldsymbol{v}}$ are taken with respect to the densities $f(\cdot;\boldsymbol{u})$ and $f(\cdot;\boldsymbol{v})$, respectively. A related approach to locating a good importance density involves the *Kullback–Leibler divergence*, or *CE distance*. To motivate the method, first note that the zero-variance importance density for estimating ℓ is simply $g^*(\boldsymbol{x}) = \ell^{-1} f(\boldsymbol{x};\boldsymbol{u})\, \mathbf{1}(S(\boldsymbol{x}) > \gamma)$—the conditional density given the rare event. Obviously, g^* cannot be used directly in practice as it involves the unknown constant ℓ. Nevertheless, this provides a practical criterion to locate a good importance density. Specifically, if we choose the density within $\mathcal{F}$ that is the closest to g^* in the CE distance, then intuitively the associated estimator should have reasonable performance. Let $\boldsymbol{v}_{\mathrm{ce}}$ denote the minimizer, which we refer to as the *optimal CE parameter vector*. It can be shown [16, pp. 67–68] that solving the CE minimization problem is equivalent to finding

$$\boldsymbol{v}_{\mathrm{ce}} = \underset{\mathbf{v}}{\operatorname{argmax}} \int f(\boldsymbol{x};\mathbf{u})\, \mathbf{1}(S(\boldsymbol{x}) > \gamma) \log f(\boldsymbol{x};\mathbf{v})\, \mathrm{d}\boldsymbol{x}. \tag{2.2}$$

Although the optimal CE and VM parameter vectors can be obtained analytically for a few specific cases, in general the optimization problems in (2.1) and (2.2) are difficult to solve. Thus, in practice, we often need to estimate $\boldsymbol{v}_{\mathrm{vm}}$ or $\boldsymbol{v}_{\mathrm{ce}}$ via a multilevel procedure, which we shall call *multilevel VM or CE* (see [11] for a more thorough discussion).

Recent research has shown that in certain high-dimensional cases the estimates for $\boldsymbol{v}_{\mathrm{vm}}$ and $\boldsymbol{v}_{\mathrm{ce}}$ obtained from the multilevel procedures are not accurate, and, as a consequence, the associated estimators perform poorly [8], [9], [14], [15]. A recent variant, called the *screening method*, introduced in [15], aims to reduce the dimension of the likelihood ratio, and is shown to perform better than the multilevel VM and CE methods in various high-dimensional estimation problems. To motivate the method, partition the parameter vector $\boldsymbol{u}$ into two subsets: $\boldsymbol{u} = (\boldsymbol{u}_0, \boldsymbol{u}_1)$, where the occurrence of the rare event $\{S(\boldsymbol{X}) > \gamma\}$ is substantially affected by $\boldsymbol{u}_1$ but not by $\boldsymbol{u}_0$. The vector $\boldsymbol{u}_1$ is referred to as the *bottleneck parameter*. Now consider the parametric family $\mathcal{F}_0 = \{f(\boldsymbol{x};\widetilde{\boldsymbol{v}})\}$ indexed by $\widetilde{\boldsymbol{v}}$, where $\widetilde{\boldsymbol{v}} = (\boldsymbol{u}_0, \boldsymbol{v}_1)$ and $\boldsymbol{u}_0$ is fixed—therefore, $\mathcal{F}_0$ is in fact indexed by $\boldsymbol{v}_1$. The screening method proceeds in the same way as the multilevel CE and VM methods, but instead of twisting the whole parameter vector $\boldsymbol{u}$, we only twist the bottleneck parameter $\boldsymbol{u}_1$.

Since $\mathcal{F}_0 \subset \mathcal{F}$, the variance of the importance sampling estimator associated with $\boldsymbol{v}_{\mathrm{vm}}$ is at least as small as the variance of the estimator associated with $\widetilde{\boldsymbol{v}}_{\mathrm{vm}}$ simply by definition. Paradoxically, however, the empirical findings in [15] suggest otherwise for the situation where the parameters are estimated via a multilevel procedure. A possible explanation is that the parameter vector obtained via the multilevel procedure, say, $\widehat{\boldsymbol{v}}_{\mathrm{vm},T}$, is not an accurate estimate for $\boldsymbol{v}_{\mathrm{vm}}$. By reducing the dimension of the likelihood ratio via the screening method, we can estimate $\widetilde{\boldsymbol{v}}_{\mathrm{vm}}$—the optimal VM parameter vector within $\mathcal{F}_0$—more accurately. As a result, the importance density $f(\boldsymbol{x};\widehat{\widetilde{\boldsymbol{v}}}_{\mathrm{vm},T})$, where $\widehat{\widetilde{\boldsymbol{v}}}_{\mathrm{vm},T}$ denotes the parameter vector obtained via the screening method, is 'closer' to g^* compared to $f(\boldsymbol{x};\widehat{\boldsymbol{v}}_{\mathrm{vm},T})$, and, thus, the estimator associated with the former density has a smaller variance than that of the latter.

Another improved variant proposed by Chan and Kroese [8] aims to estimate $\boldsymbol{v}_{\mathrm{ce}}$ in one step so as to circumvent likelihood degeneracy of the estimation procedure. Specifically, instead of the multilevel procedure in the classic CE method, they proposed estimating $\boldsymbol{v}_{\text{ce}}$ by finding

$$\widehat{\boldsymbol{v}}_{\text{ce}} = \underset{\mathbf{v}}{\operatorname{argmax}} \sum_{j=1}^{M} \log f(\boldsymbol{X}_j; \mathbf{v}),$$

where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M$ are draws from g^*. They demonstrated that the improved CE method does not only give substantial improvement over the traditional approach but also works well in high-dimensional estimation problems. Generating draws from g^*, however, might not be trivial in general, but with the advent of Markov chain Monte Carlo (MCMC) methods, this problem is well studied and a variety of techniques are available. In Sections 4–6 we consider various concrete examples where we can derive asymptotic expressions for $\boldsymbol{v}_{\text{vm}}$ and $\boldsymbol{v}_{\text{ce}}$, and we show that they are identical asymptotically. We then compute the asymptotic variances of the associated estimators and investigate how they are affected by the estimation errors introduced in the multilevel VM and CE approaches.

3. Sum of independent nonidentical random variables in the exponential families

In this section we consider the rare-event regime where the number of random variables n approaches infinity. To set the stage, suppose that $X_1, X_2, \ldots$ is a sequence of independent but not necessarily identical random variables where each X_j belongs to a one-parameter exponential family parameterized by the mean; that is, the density of each X_j is given by

$$f_j(x; u_j) = \mathrm{e}^{x\theta(u_j) - \zeta(\theta(u_j))} h_j(x). \tag{3.1}$$

Let $\mu_n = \sum_{i=1}^{n} \mathrm{E}\, X_i$ and $\sigma_n^2 = \sum_{i=1}^{n} \operatorname{var} X_i$. We are interested in estimating the probabilities

$$\ell_n = \mathrm{P}(S_n > nb) \quad \text{as } n \to \infty,$$

where $S_n = X_1 + \cdots + X_n$ and $\lim_{n\to\infty}(nb - \mu_n)/\sigma_n = \infty$. We will show that the optimal CE parameters coincide with those suggested by large deviation theory. More specifically, the CE method suggests twisting the means of the random variables such that their sum is equal to the threshold nb.

Proposition 3.1. *Let $X_1, X_2, \ldots$ be a sequence of independent random variables such that X_j belongs to a one-parameter exponential family parameterized by the mean with PDF given in (3.1). Consider estimating $\ell_n = \mathrm{P}(S_n > nb)$ as $n \to \infty$ via the CE method with importance density of the form $\prod_{i=1}^{n} f_i(x_i; v_i)$. Suppose that $\lim_{n\to\infty}(nb - \mu_n)/\sigma_n = \infty$, where $\mu_n = \sum_{i=1}^{n} \mathrm{E}\, X_i$ and $\sigma_n^2 = \sum_{i=1}^{n} \operatorname{var} X_i$. Then the optimal CE parameters $v_{\text{ce},1}^*, v_{\text{ce},2}^*, \ldots$ satisfy*

$$\sum_{i=1}^{n} v_{\text{ce},i}^* \sim nb \quad \text{as } n \to \infty.$$

Proof. First note that to estimate ℓ_n, the optimal CE parameters $v_{\text{ce},i}^*$, $i = 1, \ldots, n$, are given in [17, p. 320]: $v_{\text{ce},i}^* = \mathrm{E}[X_i \mid S_n > nb]$. Therefore, $\sum_{i=1}^{n} v_{\text{ce},i}^* = \mathrm{E}[S_n \mid S_n > nb]$. By the central limit theorem, $(S_n - \mu_n)/\sigma_n$ is asymptotically $\mathsf{N}(0, 1)$ distributed as $n \to \infty$. Therefore,

$$\mathrm{E}[S_n \mid S_n > nb] \sim \mu_n + \frac{\varphi((nb - \mu_n)/\sigma_n)}{1 - \Phi((nb - \mu_n)/\sigma_n)} \sigma_n \sim \mu_n + \frac{nb - \mu_n}{\sigma_n} \sigma_n = nb$$

as $n \to \infty$, where $\varphi(\cdot)$ and $\Phi(\cdot)$ are respectively the PDF and cumulative distribution function (CDF) of the standard normal distribution—hence, the desired result.

In what follows, we consider a different rare-event regime where we keep the number of random variables n fixed and let γ go to infinity.

4. Sum of exponential random variables

Consider the estimation of $\ell = \mathrm{P}(X_1 + \cdots + X_n > \gamma)$ via importance sampling, where $X_i \overset{\text{i.i.d.}}{=} \mathsf{Exp}(1),\ i = 1\ldots, n$; that is, X_i has PDF $f(x) = \mathrm{e}^{-x},\ x \geq 0$. Note that

$$\ell = \mathrm{e}^{-\gamma} \sum_{k=0}^{n-1} \frac{\gamma^k}{k!} = \frac{\Gamma(n,\gamma)}{\Gamma(n)}, \tag{4.1}$$

where $\Gamma(n) = (n-1)!$ and $\Gamma(n,\gamma)$ is the value of the (upper) incomplete gamma function at (n,γ). Suppose that we generate $X_i \overset{\text{i.i.d.}}{=} \mathsf{Exp}(v^{-1}),\ i = 1,\ldots,n$, with PDF $f(x; v^{-1}) = v^{-1}\exp(-v^{-1}x)$. It follows that the general term in the importance sampling estimator is

$$Z(v) = \mathbf{1}(X_1 + \cdots + X_n > \gamma) W(\boldsymbol{X}; 1, v^{-1}),$$

where the likelihood ratio is given by

$$W(\boldsymbol{x}; 1, v^{-1}) = v^n \exp\left(-\left(1 - \frac{1}{v}\right)\sum_{i=1}^{n} x_i\right).$$

We first derive the asymptotic expressions for the optimal VM and CE parameters and show that they are the same. We note that Rubinstein and Kroese [16, p. 69] proved the special case for $n = 1$.

Proposition 4.1. *Let $X_i \overset{\text{i.i.d.}}{=} \mathsf{Exp}(1),\ i = 1,\ldots,n$. To estimate $\ell = \mathrm{P}(X_1+\cdots+X_n > \gamma)$ via importance sampling, suppose that we generate X_i identically from the $\mathsf{Exp}(v^{-1})$ distribution. Then the optimal VM and CE parameters are asymptotically the same. In fact, we have $v_{\mathrm{vm}} \sim \gamma/n$ and $v_{\mathrm{ce}} \sim \gamma/n$.*

Proof. To obtain the optimal VM parameter, we first derive an asymptotic expression for the second moment of the importance sampling estimator $Z(v)$:

$$\begin{aligned}
\mathrm{E}_{1/v}\, Z(v)^2 &= \mathrm{E}_1\, Z(v) \\
&= \int_{\sum x_i > \gamma} v^n \exp\left(-\left(1 - \frac{1}{v}\right)\sum_{i=1}^{n} x_i\right) \prod_{i=1}^{n} \mathrm{e}^{-x_i}\, \mathrm{d}\boldsymbol{x} \\
&= v^n \left(2 - \frac{1}{v}\right)^{-n} \mathrm{P}(Y_1 + \cdots + Y_n > \gamma).
\end{aligned}$$

Here $Y_i \overset{\text{i.i.d.}}{=} \mathsf{Exp}(2 - 1/v),\ i = 1,\ldots,n$, for $v > \frac{1}{2}$. Therefore, we have

$$\mathrm{E}_{1/v}\, Z(v)^2 \sim \frac{\mathrm{e}^{-2\gamma}\gamma^{n-1}}{(n-1)!} \frac{v^n \mathrm{e}^{\gamma/v}}{2 - 1/v} \quad \text{as } \gamma \to \infty.$$

To obtain v_{vm}, we differentiate $\log \mathrm{E}_{1/v}\, Z(v)^2$ with respect to v and solve the equation (with the constraint that $v > \frac{1}{2}$):

$$\frac{\mathrm{d}}{\mathrm{d}v} \log \mathrm{E}_{1/v}\, Z(v)^2 \sim n - \frac{\gamma}{v} - \frac{1}{2v-1} = 0.$$

It follows that

$$v_{\mathrm{vm}} = \frac{\gamma + \sqrt{\gamma^2 - n\gamma + (n+1)^2/4}}{2n} + \frac{n+1}{4n} + \mathcal{O}(\gamma^{-1}) \sim \frac{\gamma}{n}.$$

To compute the optimal CE parameter, we first note that the exponential distribution is a member of the exponential family. Therefore,

$$v_{\mathrm{ce}} = \mathrm{E}[X_1 \mid X_1 + \cdots + X_n > \gamma] = \frac{1}{n}\,\mathrm{E}[Y \mid Y > \gamma],$$

where $Y \overset{\mathrm{D}}{=} \mathsf{Gamma}(n, 1)$. Direct computation shows (see also [16, p. 78]) that $\mathrm{E}[Y \mid Y > \gamma] = \Gamma(n+1, \gamma)/\Gamma(n, \gamma)$. Since

$$\Gamma(n+1, \gamma) = n\Gamma(n, \gamma) + \gamma^n \mathrm{e}^{-\gamma} \quad \text{and} \quad \Gamma(n, \gamma) = \gamma^{n-1}\mathrm{e}^{-\gamma}(1 + \mathcal{O}(\gamma^{-1})),$$

we have

$$\mathrm{E}[Y \mid Y > \gamma] = n + \frac{\gamma^n \mathrm{e}^{-\gamma}}{\Gamma(n, \gamma)} = n + \gamma(1 + \mathcal{O}(\gamma^{-1})).$$

It follows that $v_{\mathrm{ce}} = \gamma/n + 1 + \mathcal{O}(\gamma^{-1}) \sim \gamma/n$ as $\gamma \to \infty$.

Therefore, by Proposition 4.1, the optimal VM parameter is asymptotically identical to that given by the CE program when $\gamma \to \infty$. We show in the next proposition that either v_{vm} or v_{ce} in fact gives an asymptotically optimal importance sampling estimator for ℓ. In addition, by the definition of v_{vm}, no other importance sampling estimators obtained by generating X_i identically from $\mathsf{Exp}(v^{-1})$ can be strongly efficient. In what follows, we also investigate how the estimation error in obtaining v_{ce} affects the relative error of the associated importance sampling estimator.

Proposition 4.2. *Under the same assumptions as in Proposition 4.1, if we set* $v = \gamma/n + h$ *for some constant* h *then*

$$\frac{\mathrm{E}_{1/v}\, Z(v)^2}{\ell^2} = \frac{\mathrm{e}^n (n-1)!}{2n^n}\gamma\left(1 + \frac{2n-1}{2\gamma} + \frac{n}{4\gamma^2}(2n^2h^2 - 2nh + 3n - 2) + \mathcal{O}(\gamma^{-3})\right) \tag{4.2}$$

as $\gamma \to \infty$. *In particular, the optimal VM/CE parameter gives an asymptotically optimal estimator.*

Proof. Let $v = \gamma/n + h$. By a similar computation as in Proposition 4.1 we have

$$\begin{aligned}
\mathrm{E}_{1/v}\, Z(v)^2 &= \frac{v^n}{(2 - 1/v)^n}\frac{\Gamma(n, (2 - 1/v)\gamma)}{\Gamma(n)} \\
&= \frac{v^n \mathrm{e}^{-2\gamma}\mathrm{e}^{\gamma/v}}{(2 - 1/v)(n-1)!}\gamma^{n-1}\left(1 + \frac{n-1}{2 - 1/v}\gamma^{-1} + \mathcal{O}(\gamma^{-2})\right).
\end{aligned} \tag{4.3}$$

Substituting $v = \gamma/n + h$ and using the expression for ℓ in (4.1) gives (4.2). The final statement in the proposition follows by setting $v = \gamma/n \sim v_{\mathrm{vm}}$.

It is worth noting that in (4.2) only the coefficient of the third order term $1/\gamma^3$ involves h, and that the magnitude of h does not affect the asymptotic efficiency of the importance sampling estimator. However, when the dimension of the estimation problem n is large, h might have a substantial impact on the variance of the importance sampling estimator for any finite γ. This explains why the multilevel CE estimator generally works well in problems with light-tailed random variables, but sometimes breaks down when the dimension of the problem becomes large; see, e.g. [8].

We now investigate what happens when the importance sampling parameter v is obtained via a random procedure, such as in the multilevel VM or CE method. Let us denote the *random* parameter thus obtained by V, which is independent of the $\{Z_i\}$ used in the importance sampling estimator. In the CE procedure the reference parameter V is obtained as

$$V = \frac{\sum_{k=1}^{N} \mathbf{1}(S_k > \gamma) W_k(w) S_k}{n \sum_{k=1}^{N} \mathbf{1}(S_k > \gamma) W_k(w)}, \tag{4.4}$$

where $W_k(w) = W(\boldsymbol{X}_k; 1, 1/w)$ is the kth likelihood ratio corresponding to a reference parameter w obtained in the penultimate iteration, and $S_k \overset{\text{i.i.d.}}{=} \mathsf{Gamma}(n, 1/w),\ k = 1, \dots, N$. The parameter w is usually random as well—for example when obtained via a CE procedure. Suppose, however, that w is some arbitrarily fixed reference parameter. The asymptotic distribution of V as a function of w is given in the next proposition.

Proposition 4.3. *Under the same assumptions as in Proposition 4.1, the CE reference parameter V given in (4.4) is asymptotically normal as $N \to \infty$ with mean v_{ce} and variance $\sigma^2_{\gamma,w}/N$. Furthermore, we have*

$$\sigma^2_{\gamma,w} \sim \left(1 - \frac{1}{n}\right)^2 \frac{w^n (n-1)!}{(2 - 1/w)} \gamma^{-n+3} \mathrm{e}^{\gamma/w} \quad \text{as } \gamma \to \infty.$$

In particular,

$$\sigma^2_{\gamma,\gamma/n} \sim \gamma^3 \frac{(n-1)!\,(1 - 1/n)^2 \mathrm{e}^n}{2n^n}.$$

Proof. First note that V given in (4.4) is a ratio estimator. By the delta method [1, pp. 75–78], the asymptotic distribution is normal with mean

$$\mu = \frac{\mathrm{E}_{1/w}\, \mathbf{1}(S > \gamma) W(w) S}{n\, \mathrm{E}_{1/w}\, \mathbf{1}(S > \gamma) W(w)} = \frac{\mathrm{E}_1\, \mathbf{1}(S > \gamma) S}{n\, \mathrm{E}_1\, \mathbf{1}(S > \gamma)} = v_{\text{ce}}$$

and variance $\sigma^2_{\gamma,w}/N$, with

$$\sigma^2_{\gamma,w} = \frac{\operatorname{var}(A) - 2\mu \operatorname{cov}(A, B) + \mu^2 \operatorname{var}(B)}{\ell^2},$$

where $A = \mathbf{1}(S > \gamma) W(w) S$, $B = \mathbf{1}(S > \gamma) W(w)$, and S is $\mathsf{Gamma}(n, 1/w)$ distributed. The second moment of B is given in (4.3) with w substituted for v. The expectation of A is simply $\mathrm{E}_{1/w} A = \mathrm{E}_1\, \mathbf{1}(S > \gamma) S = \ell\, v_{\text{ce}}$. The second moment of A is

$$\begin{aligned}
\mathrm{E}_{1/w} A^2 &= \int \mathbf{1}(s > \gamma) w^n \mathrm{e}^{-(1-1/w)s} s^2 \frac{1}{\Gamma(n)} s^{n-1} \mathrm{e}^{-s}\, \mathrm{d}s \\
&= \frac{n(n+1)}{(2 - 1/w)^{n+2}} \frac{\Gamma(n+2, (2 - 1/w)\gamma)}{\Gamma(n+2)} \\
&\sim \gamma^2\, \mathrm{E}_w\, B^2.
\end{aligned}$$

Moreover, $\mathrm{E}_{1/w}\, AB = \mathrm{E}_1\, \mathbf{1}(S > \gamma) W(w) S \sim \gamma\, \mathrm{E}_w\, B^2$. It follows, after some algebra, that, for $n > 1$,

$$\sigma^2_{\gamma,w} \sim \left(1 - \frac{1}{n}\right)^2 \frac{w^n (n-1)!}{(2 - 1/w)} \gamma^{-n+3} \mathrm{e}^{\gamma/w}.$$

This completes the proof.

Note that the asymptotic variance of the CE reference parameter V is cubic in γ when we set $w = \gamma/n \sim v_{\mathrm{vm}}$. Therefore, even though v_{ce} gives an asymptotically optimal estimator, when γ is sufficiently large, the estimation error in obtaining v_{ce} in the multilevel CE procedure might be so substantial that it renders the resulting importance sampling estimator unreliable.

5. Sum of Pareto random variables

We now consider estimating the tail probability of the sum of heavy-tailed random variables. Specifically, we wish to estimate $\ell = \mathrm{P}(X_1 + \cdots + X_n > \gamma)$ via importance sampling, where $X_i \overset{\text{i.i.d.}}{=} \mathsf{Pareto}(1, 1)$, $i = 1, \ldots, n$. Since the Pareto distribution is subexponential [1, pp. 173–183], we have $\ell \sim n/(1+\gamma)$ as $\gamma \to \infty$. To estimate ℓ via importance sampling, we consider the $\mathsf{Pareto}(\alpha, 1)$ family indexed by $\alpha > 0$ with PDF $f(x;\alpha) = \alpha(1+x)^{-(\alpha+1)}$, $x \geq 0$. Now suppose that we generate $X_i \overset{\text{i.i.d.}}{=} \mathsf{Pareto}(\alpha, 1)$, $i = 1, \ldots, n$. The general term of the likelihood ratio is

$$W(\boldsymbol{x}; 1, \alpha) = \prod_{i=1}^{n} \frac{(1+x_i)^{-2}}{\alpha(1+x_i)^{-(\alpha+1)}} = \alpha^{-n} \prod_{i=1}^{n} (1+x_i)^{-(1-\alpha)},$$

and the corresponding importance sampling estimator is

$$Z(\alpha) = \mathbf{1}(X_1 + \cdots + X_n > \gamma) W(\boldsymbol{X}; 1, \alpha).$$

In the following proposition we show that the optimal VM and CE parameters are identical. In fact, we show that $\alpha_{\mathrm{vm}} \sim n/\log(1+\gamma)$, which gives the minimum variance estimator within the class of importance sampling estimators obtained by generating $X_i \overset{\text{i.i.d.}}{=} \mathsf{Pareto}(\alpha, 1)$ for $i = 1, \ldots, n$. Compare this with the suggestions in [3] and [10].

Proposition 5.1. *Let $X_i \overset{\text{i.i.d.}}{=} \mathsf{Pareto}(1, 1)$, $i = 1, \ldots, n$. Suppose that we wish to estimate $\ell = \mathrm{P}(X_1 + \cdots + X_n > \gamma)$ via importance sampling by generating $X_i \overset{\text{i.i.d.}}{=} \mathsf{Pareto}(\alpha, 1)$. Then the optimal VM and CE parameters for α are asymptotically the same. In fact, we have $\alpha_{\mathrm{vm}} \sim n/\log(1+\gamma)$.*

Proof. Note that the optimal CE parameter for α is given in [4]: $\alpha_{\mathrm{ce}} = (1 + \log(1+\gamma)/n)^{-1} \sim n/\log(1+\gamma)$. To compute the optimal VM parameter, we first derive the second moment of $Z(\alpha)$ with respect to the $\mathsf{Pareto}(\alpha, 1)$ distribution:

$$\begin{aligned} \mathrm{E}_\alpha\, Z(\alpha)^2 &= \mathrm{E}_1\, Z(\alpha) \\ &= \int_{\sum x_i > \gamma} \alpha^{-n} \prod_{i=1}^{n} (1+x_i)^{-(1-\alpha)} (1+x_i)^{-2}\, \mathrm{d}\boldsymbol{x} \\ &= (2\alpha - \alpha^2)^{-n}\, \mathrm{P}(Y_1 + \cdots + Y_n > \gamma). \end{aligned}$$

Here $Y_i \overset{\text{i.i.d.}}{=} \mathsf{Pareto}(2-\alpha, 1)$, $i = 1, \ldots, n$, provided that $\alpha < 2$. Hence,

$$\mathrm{E}_\alpha\, Z(\alpha)^2 \sim (2\alpha - \alpha^2)^{-n} n (1+\gamma)^{-(2-\alpha)}. \tag{5.1}$$

By a computation similar to that in Proposition 4.1 we have

$$\begin{aligned}\alpha_{\mathrm{vm}} &= \frac{(2/n)\log(1+\gamma)+2-\sqrt{(4/n^2)\log^2(1+\gamma)+4}}{(2/n)\log(1+\gamma)} + \mathcal{O}(\log^{-2}(1+\gamma)) \\ &\sim \frac{n}{\log(1+\gamma)}.\end{aligned}$$

Again, the optimal VM parameter is asymptotically identical to that given by the CE program as $\gamma \to \infty$.

We next investigate how the choice of the parameter α affects the growth rate of $\mathrm{E}_\alpha\, Z(\alpha)^2/\ell^2$. As a corollary to Proposition 5.1, we show that $\alpha = \alpha_{\mathrm{ce}} \sim n/\log(1+\gamma)$ gives an importance sampling estimator that is asymptotically optimal. We note that Asmussen and Kroese [2] provided a conditional Monte Carlo estimator that has bounded relative error for the case of the sum of Pareto random variables. In addition, by utilizing a technique based on Lyapunov-type inequalities first introduced in [5], Blanchet and Li [6] were able to derive an importance sampling estimator that achieves bounded relative error for general subexponential distributions.

Proposition 5.2. *Under the same assumptions as in Proposition 5.1, if we set* $\alpha = n/\log(1+\gamma)+h$ *for some constant* h *such that* $0<\alpha<2$, *then*

$$\frac{\mathrm{E}_\alpha\, Z(\alpha)^2}{\ell^2} \sim \frac{\mathrm{e}^n}{n}\gamma^h\bigg(h(2-h)+\frac{2n}{\log(1+\gamma)}-\frac{n^2}{\log^2(1+\gamma)}\bigg)^{-n} \quad \textit{as } \gamma\to\infty. \tag{5.2}$$

In particular, the optimal VM/CE parameter gives an asymptotically optimal estimator.

Proof. By (5.1) and the fact that $\ell \sim n/(1+\gamma)$, we have

$$\frac{\mathrm{E}_\alpha\, Z(\alpha)^2}{\ell^2} \sim \frac{1}{n(2\alpha-\alpha^2)^n}(1+\gamma)^\alpha.$$

Hence, if we set $\alpha = n/\log(1+\gamma)+h$ then (5.2) follows. As a result, for $\alpha=\alpha_{\mathrm{ce}} \sim n/\log(1+\gamma)$, we have

$$\frac{\mathrm{E}_\alpha\, Z(\alpha)^2}{\ell^2} \sim \frac{\mathrm{e}^n}{n}\bigg(\frac{2n}{\log(1+\gamma)}-\frac{n^2}{\log^2(1+\gamma)}\bigg)^{-n} \quad \text{as } \gamma\to\infty.$$

This completes the proof.

It is of interest to note that in contrast to the light-tailed case, the estimation error h does increase the asymptotic variance of the importance sampling estimator. Therefore, the problem of suboptimal VM and CE reference parameters is expected to be more severe in the heavy-tailed case.

6. Sum of Weibull random variables

Consider the same estimation problem as in the last section, but now $X_i \overset{\text{i.i.d.}}{=} \mathsf{Weib}(\beta,1)$, $i=1,\dots,n$, for $0<\beta<1$; that is, X_i has PDF $f(x;\beta)=\beta x^{\beta-1}\mathrm{e}^{-x^\beta}$. We wish to estimate the tail probability ℓ via importance sampling by tilting the scale parameter. That is, we locate the importance density within the parametric family $\mathsf{Weib}(\beta,\theta)$ with PDF $f(x;\beta,\theta)=\theta\beta x^{\beta-1}\mathrm{e}^{-\theta x^\beta}$, $x\ge 0$, indexed by $\theta>0$ while keeping β fixed. It follows that the general term of the importance sampling estimator is

$$Z(\theta) = \mathbf{1}(X_1 + \cdots + X_n > \gamma) W(\boldsymbol{X}; 1, \theta),$$

with likelihood ratio

$$W(\boldsymbol{x}; 1, \theta) = \theta^{-n} \exp\biggl(-(1-\theta)\sum_{i=1}^{n} x_i^{\beta}\biggr).$$

Again, for this sum of Weibull random variables case, the optimal VM and CE parameters coincide asymptotically. Kroese and Rubinstein [12] proved that, for the $n = 2$ case, the optimal VM parameter is asymptotically $\theta_{\mathrm{vm}} \sim 2/\gamma^{\beta}$. They conjectured that, for general n, $\theta_{\mathrm{vm}} \sim n/\gamma^{\beta}$, and the squared relative error of the resulting estimator increases proportionally to $\gamma^{n\beta}$. The following propositions prove these two conjectures.

Proposition 6.1. *Let* $X_i \overset{\text{i.i.d.}}{=} \mathsf{Weib}(\beta, 1)$, $i = 1, \ldots, n$, *with* $0 < \beta < 1$. *Suppose that we wish to estimate* $\ell = \mathrm{P}(X_1 + \cdots + X_n > \gamma)$ *via importance sampling by generating* $X_i \overset{\text{i.i.d.}}{=} \mathsf{Weib}(\beta, \theta)$. *Then the optimal VM and CE parameters for* θ *are asymptotically identical. In fact, we have* $\theta_{\mathrm{vm}} \sim n/\gamma^{\beta}$.

Proof. First note that the optimal CE parameter for θ is given in [4] and [16]: $\theta_{\mathrm{ce}} = n/(n + \gamma^{\beta}) \sim n/\gamma^{\beta}$. Next we compute the optimal VM parameter as follows:

$$\begin{aligned}
\mathrm{E}_{\theta}\, Z(\theta)^2 &= \mathrm{E}_1\, Z(\theta) \\
&= \int_{\sum x_i > \gamma} \theta^{-n} \exp\biggl(-(1-\theta)\sum_{i=1}^{n} x_i^{\beta}\biggr) \prod_{i=1}^{n} \beta x_i^{\beta-1} \mathrm{e}^{-x_i^{\beta}}\, \mathrm{d}\boldsymbol{x} \\
&= \theta^{-n}(2-\theta)^{-n}\, \mathrm{P}(Y_1 + \cdots + Y_n > \gamma).
\end{aligned}$$

Here $Y_i \overset{\text{i.i.d.}}{=} \mathsf{Weib}(\beta, 2-\theta)$, provided that $\theta < 2$. Since the $\mathsf{Weib}(\beta, \theta)$ distribution is subexponential for $\beta < 1$, we have

$$\mathrm{E}_{\theta}\, Z(\theta)^2 \sim \frac{n}{\theta^n (2-\theta)^n}\, \mathrm{P}(Y_1 > \gamma) = \frac{n}{\theta^n (2-\theta)^n} \mathrm{e}^{-(2-\theta)\gamma^{\beta}} \quad \text{as } \gamma \to \infty. \tag{6.1}$$

By a similar computation as in Proposition 4.1, it can be shown that

$$\theta_{\mathrm{vm}} = \frac{n}{\gamma^{\beta}} + 1 - \sqrt{1 + \frac{n^2}{\gamma^{2\beta}}} + \mathcal{O}(\gamma^{-(\beta+1)}) \sim \frac{n}{\gamma^{\beta}} \quad \text{as } \gamma \to \infty.$$

Therefore, the optimal CE and VM parameters for θ are identical asymptotically.

The choice of θ_{vm} is to be compared with the suggestion in [10] to take $\theta = b/\gamma^{\beta}$, where $b > 0$ is some arbitrary constant. Since the $\mathsf{Weib}(\beta, 1)$ distribution is subexponential for $\beta < 1$, it follows that $\ell \sim n\mathrm{e}^{-\gamma^{\beta}}$. Therefore, using the expression in (6.1), it can be shown that if we choose θ such that $\theta\gamma^{\beta} = c$ for some constant c, the associated importance sampling estimator is asymptotically optimal. In particular, the choice $\theta = \theta_{\mathrm{ce}} \sim \theta_{\mathrm{vm}}$ gives an asymptotically optimal importance sampling estimator, which also has the minimum asymptotic variance within the class of importance sampling estimators with importance densities under which $X_i \overset{\text{i.i.d.}}{=} \mathsf{Weib}(\beta, \theta)$, $i = 1, \ldots, n$.

Proposition 6.2. *Under the same assumptions as in Proposition 6.1, if we set* $\theta = n/\gamma^\beta$ *then*

$$\frac{\mathrm{E}_\theta\, Z(\theta)^2}{\ell^2} \sim \frac{\mathrm{e}^n}{2^n n^{n+1}} \gamma^{n\beta},$$

i.e. the optimal VM/CE parameter gives an asymptotically optimal estimator.

Proof. Since the Weib$(\beta, 1)$ distribution is subexponential for $\beta < 1$, we have $\ell \sim n\mathrm{e}^{-\beta\gamma}$. It follows from (6.1) that

$$\frac{\mathrm{E}_\theta\, Z(\theta)^2}{\ell^2} \sim n^{-1}\theta^{-n}(2-\theta)^{-n}\mathrm{e}^{\theta\gamma^\beta} \quad \text{as } \gamma \to \infty.$$

The desired result follows by letting $\theta = n/\gamma^\beta$.

7. Concluding remarks and future research

We compared the VM and CE methods through various concrete examples and we found that in the three examples considered the optimal VM and CE parameters are asymptotically identical. It would be of considerable interest to determine under what conditions this is the case. Since CE estimators are typically easy to obtain, this would provide a practical approach to locate the importance sampling estimator with the minimum variance within a given parametric class. Moreover, it is worthwhile to study further the impact of CE parameter estimation on the quality of the associated importance sampling estimator.

Acknowledgements

Joshua Chan and Dirk Kroese would like to acknowledge financial support from the Australian Research Council through Discovery Grants DP0985177 and DP0987170, respectively.

References

[1] Asmussen, S. and Glynn, P. W. (2007). *Stochastic Simulation: Algorithms and Analysis*. Springer, New York.

[2] Asmussen, S. and Kroese, D. P. (2006). Improved algorithms for rare event simulation with heavy tails. *Adv. Appl. Prob.* **38,** 545–558.

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

[4] Asmussen, S., Kroese, D. P. and Rubinstein, R. Y. (2005). Heavy tails, importance sampling and cross-entropy. *Stoch. Models* **21,** 57–76.

[5] Blanchet, J. and Glynn, P. (2008). Efficient rare-event simulation for the maximum of heavy-tailed random walks. *Ann. Appl. Prob.* **18,** 1351–1378.

[6] Blanchet, J. and Li, C. (2011). Efficient rare event simulation for heavy-tailed compound sums. To appear in *ACM Trans. Model. Comput. Simul.*

[7] Casella, G. and Berger, R. L. (2001). *Statistical Inference*, 2nd edn. Duxbury Press.

[8] Chan, J. C. C. and Kroese, D. P. (2011). Improved cross-entropy method for estimation. Submitted.

[9] Chan, J. C. C. and Kroese, D. P. (2009). Rare-event probability estimation with conditional Monte Carlo. *Ann. Operat. Res.*, 19pp.

[10] Juneja, S. and Shahabuddin, P. (2002). Simulating heavy tailed processes using delayed hazard rate twisting. *ACM Trans. Model. Comput. Simul.* **12,** 94–118.

[11] Kroese, D. P. (2011). The cross-entropy method. In *Wiley Encyclopedia of Operations Research and Management Science*.

[12] Kroese, D. P. and Rubinstein, R. Y. (2004). The transform likelihood ratio method for rare event simulation with heavy tails. *Queueing Systems* **46,** 317–351.

[13] L'Ecuyer, P., Blanchet, J. H., Tuffin, B. and Glynn, P. W. (2010). Asymptotic robustness of estimators in rare-event simulation. *ACM Trans. Model. Comput. Simul.* **20,** 1–41.

[14] Rubinstein, R. Y. (2009). The Gibbs cloner for combinatorial optimization, counting and sampling. *Methodology Comput. Appl. Prob.* **11,** 491–549.

[15] Rubinstein, R. Y. and Glynn, P. W. (2009). How to deal with the curse of dimensionality of likelihood ratios in Monte Carlo simulation. *Stoch. Models* **25,** 547 – 568.
[16] Rubinstein, R. Y. and Kroese, D. P. (2004). *The Cross-Entropy Method*. Springer, New York.
[17] Rubinstein, R. Y. and Kroese, D. P. (2007). *Simulation and the Monte Carlo Method*, 2nd edn. John Wiley, New York.

JOSHUA C. C. CHAN, *Australian National University*

Research School of Economics, Australian National University, Canberra, ACT 0200, Australia.

PETER W. GLYNN, *Stanford University*

Department of Management Science and Engineering, Huang Engineering Center, Stanford University, 475 Via Ortega, Stanford, CA 94305-4121, USA.

DIRK P. KROESE, *University of Queensland*

Department of Mathematics, University of Queensland, St Lucia, Brisbane, QLD 4072, Australia.
Email address: kroese@maths.uq.edu.au