Applied Probability Trust (31 August 2010)

A COMPARISON OF CROSS-ENTROPY AND VARIANCE MINIMIZATION STRATEGIES

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

PETER W. GLYNN,** *Stanford University*

DIRK P. KROESE,*** *University of Queensland*

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 cross-entropy distance might be 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.

2000 Mathematics Subject Classification: Primary 65C05
Secondary 65C60

* Postal address: Research School of Economics, Australian National University, Canberra, ACT 0200, Australia

** Postal address: Department of Management Science and Engineering, 380 Panama Way, Stanford University, Stanford, CA 94305-4026, USA

*** Postal address: Department of Mathematics, University of Queensland, St Lucia, Brisbane, QLD 4072, Australia

1. Introduction

This article compares two adaptive importance sampling procedures, namely the variance minimization (VM) and cross-entropy (CE) methods [11, 15], 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 required to locate the optimal VM importance density is often intractable. 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 *cross-entropy 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 close-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 cross-entropy distance— to the zero-variance importance density is also the one whose associated estimator has the minimum asymptotic variance. The significance of this 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 problems do not admit close-form solutions, the optimal parameters need to be estimated via a multi-level 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. It is followed by three case studies: Section 3 considers

the example of sum of exponential random variables, followed by the cases for Pareto and Weibull random variables in Sections 4 and 5 respectively. We conclude with a scenario in which the number of parameters is sent to infinity.

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 $X_i \overset{iid}{=} f, i = 1, \ldots, n$ to indicate that $X_1, \ldots, X_n$ are independent and identically distributed (iid) 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 $\lim_{\gamma\to\infty} \log \mathbb{E}Z(\gamma)^2 / \log \ell(\gamma) = 2$. This condition is equivalent to the requirement that $\lim_{\gamma\to\infty} \mathbb{E}Z(\gamma)^2/\ell(\gamma)^{2-\epsilon} = 0$, for every $\epsilon > 0$. The estimator is said to be *strongly efficient* or have *bounded relative error* if $\sup_{\gamma\geqslant 0} \mathbb{E}Z(\gamma)^2/\ell(\gamma)^2 < \infty$. It is readily observed that bounded relative error implies asymptotic optimality. These notions of efficiency are standard in the literature; see, for example, [2] and [12].

We are interested in estimating the probability of the form

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

where S is some real-valued performance function, $\mathbf{X}$ 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

$$\widehat{\ell}_{\mathrm{IS}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(S(\mathbf{X}_i) > \gamma) \frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)}, \tag{2}$$

where $\mathbf{X}_i \overset{iid}{=} g$, $i = 1, \ldots, N$ for some importance sampling pdf g for which $g(\mathbf{x}) = 0 \Rightarrow \mathbb{1}(S(\mathbf{x}) > \gamma) f(\mathbf{x}) = 0$ for all $\mathbf{x}$. 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 $\mathcal{F} = \{f(\mathbf{x}; \mathbf{v})\}$ indexed by a parameter vector $\mathbf{v}$ that contains the nominal (original) density f. Thus, we can write $f(\mathbf{x}) = f(\mathbf{x}; \mathbf{u})$ for some parameter vector $\mathbf{u}$. For any given $\mathbf{v}$ the general term of the associated importance sampling estimator is $Z(\mathbf{v}) = W(\mathbf{X}; \mathbf{u}, \mathbf{v}) \mathbb{1}(S(\mathbf{X}) > \gamma)$, where $W(\mathbf{x}; \mathbf{u}, \mathbf{v})$ is

the *likelihood ratio* defined as $W(\mathbf{x};\mathbf{u},\mathbf{v}) = f(\mathbf{x};\mathbf{u})/f(\mathbf{x};\mathbf{v})$. Now, we wish to choose $\mathbf{v}$ so that the associated importance sampling estimator has minimum variance within the parametric family $\mathcal{F}$. The minimizer $\mathbf{v}_{\rm vm}$ is referred to as the *optimal VM parameter vector*. For any unbiased estimator $\widehat{\ell}$ of ℓ, we have $\mathrm{Var}\,\widehat{\ell} = \mathbb{E}\widehat{\ell}^2 - \ell^2$. Therefore $\mathbf{v}_{\rm vm}$ can be written as

$$\mathbf{v}_{\rm vm} = \operatorname*{argmin}_{\mathbf{v}} \mathbb{E}_{\mathbf{v}} Z(\mathbf{v})^2 = \operatorname*{argmin}_{\mathbf{v}} \mathbb{E}_{\mathbf{u}} Z(\mathbf{v}) = \operatorname*{argmin}_{\mathbf{v}} \log \mathbb{E}_{\mathbf{u}} Z(\mathbf{v}), \tag{3}$$

where the expectation $\mathbb{E}_{\mathbf{w}}$ is taken with respect to some density $f(\cdot;\mathbf{w})$. A related approach to locating a good importance density involves the *Kullback-Leibler divergence*, or *cross-entropy distance*. To motivate the method, first note that the zero-variance importance density for estimating ℓ is simply $g^*(\mathbf{x}) = \ell^{-1} f(\mathbf{x};\mathbf{u})\mathbb{1}(S(\mathbf{x}) > \gamma)$—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 cross-entropy distance, then intuitively the associated estimator should have reasonable performance. Let $\mathbf{v}_{\rm ce}$ denote the minimizer, which we refer to as the *optimal CE parameter vector*. It can be shown [15] that solving the CE minimization problem is equivalent to finding

$$\mathbf{v}_{\rm ce} = \operatorname*{argmax}_{\mathbf{v}} \int f(\mathbf{x};\mathbf{u})\mathbb{1}(S(\mathbf{x}) > \gamma) \log f(\mathbf{x};\mathbf{v}) \mathrm{d}\mathbf{x}. \tag{4}$$

Although the optimal CE and VM parameter vectors can be obtained analytically for a few specific cases, in general the optimization problems in (3) and (4) are difficult to solve. Thus in practice one often needs to estimate $\mathbf{v}_{\rm vm}$ or $\mathbf{v}_{\rm ce}$ via a multi-level procedure, which we shall call *multi-level VM or CE* (see [11] for a more thorough discussion).

Recent research has shown that in certain high-dimensional cases the estimates for $\mathbf{v}_{\rm vm}$ and $\mathbf{v}_{\rm ce}$ obtained from the multi-level procedures are not accurate, and as a consequence the associated estimators perform poorly [8, 9, 13, 14]. A recent variant, called the *screening method*, is introduced in [14] that aims to reduce the dimension of the likelihood ratio, and is shown to perform better than the multi-level VM and CE methods in various high-dimensional estimation problems. To motivate the method, partition the parameter vector $\mathbf{u}$ into two subsets: $\mathbf{u} = (\mathbf{u}_0, \mathbf{u}_1)$, where the occurrence

of the rare event $\{S(\mathbf{X}) > \gamma\}$ is substantially affected by $\mathbf{u}_1$ but not by $\mathbf{u}_0$. The vector $\mathbf{u}_1$ is referred to as the *bottleneck parameter*. Now consider the parametric family $\mathcal{F}_0 = \{f(\mathbf{x}; \widetilde{\mathbf{v}})\}$ indexed by $\widetilde{\mathbf{v}}$, where $\widetilde{\mathbf{v}} = (\mathbf{u}_0, \mathbf{v}_1)$ and $\mathbf{u}_0$ is fixed—therefore, $\mathcal{F}_0$ is in fact indexed by $\mathbf{v}_1$. The screening method proceeds in the same way as the multi-level CE and VM methods, but instead of twisting the whole parameter vector $\mathbf{u}$, one only twists the bottleneck parameter $\mathbf{u}_1$.

Since $\mathcal{F}_0 \subset \mathcal{F}$, the variance of the importance sampling estimator associated with $\mathbf{v}_{\mathrm{vm}}$ is at least as small as the variance of the estimator associated with $\widetilde{\mathbf{v}}_{\mathrm{vm}}$ simply by definition. Paradoxically, however, the empirical findings in [14] 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 multi-level procedure, say, $\widehat{\mathbf{v}}_{\mathrm{vm},T}$ is not an accurate estimate for $\mathbf{v}_{\mathrm{vm}}$. By reducing the dimension of the likelihood ratio via the screening method, one can estimate $\widetilde{\mathbf{v}}_{\mathrm{vm}}$— the optimal VM parameter vector within $\mathcal{F}_0$— more accurately. As a result, the importance density $f(\mathbf{x}; \widehat{\widetilde{\mathbf{v}}}_{\mathrm{vm},T})$ is "closer" to g^* compared to $f(\mathbf{x}; \widehat{\mathbf{v}}_{\mathrm{vm},T})$, and thus the estimator associated with the former density has a smaller variance than that of the latter.

Another improved variant is proposed in [8] and aims to estimate $\mathbf{v}_{\mathrm{ce}}$ in one step so as to circumvent likelihood degeneracy om the estimation procedure. Specifically, instead of the multi-level procedure in the classic CE method, they propose estimating $\mathbf{v}_{\mathrm{ce}}$ by finding

$$\widehat{\mathbf{v}}_{\mathrm{ce}} = \underset{\mathbf{v}}{\operatorname{argmax}} \sum_{j=1}^{M} \log f(\mathbf{X}_j; \mathbf{v}), \tag{5}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_M$ are draws from g^* via, say, Markov chain Monte Carlo (MCMC) methods. They demonstrate that the improved CE method does not only give substantial improvement over the traditional approach but also works well in high-dimensional estimation problems. In what follows, we consider various concrete examples where one can derive asymptotic expressions for $\mathbf{v}_{\mathrm{vm}}$ and $\mathbf{v}_{\mathrm{ce}}$, 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 multi-level VM and CE approaches.

3. Sum of Exponential Random Variables

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

$$\ell = \mathrm{e}^{-\gamma} \sum_{k=0}^{n-1} \frac{\gamma^k}{k!} = \frac{\Gamma(n,\gamma)}{\Gamma(n)}, \tag{6}$$

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 $X_i \overset{iid}{=} \mathsf{Exp}(v^{-1})$, $i = 1, \ldots, n$ with pdf $f(x; v^{-1}) = v^{-1}\mathsf{Exp}(-v^{-1}x)$. It follows that the general term in the importance sampling estimator is $Z(v) = \mathbb{1}\{X_1 + \cdots + X_n > \gamma\} W(\mathbf{X}; 1, v^{-1})$, where the likelihood ratio is given by $W(\mathbf{x}; 1, v^{-1}) = v^n \exp\left(-(1-1/v)\sum_{i=1}^n x_i\right)$. We first derive the asymptotic expressions for the optimal VM and CE parameters and show that they are the same.

Proposition 3.1. *Let $X_i \overset{iid}{=} \mathsf{Exp}(1)$, $i = 1, \ldots, n$. To estimate $\ell = \mathbb{P}(X_1 + \cdots + X_n > \gamma)$ via importance sampling, suppose 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 $Z(v)$:

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

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

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

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

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

It follows that

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

For the optimal CE parameter, note that the solution of the maximization problem (4) is given by

$$v_{\mathrm{ce}} = \frac{\int_{\sum x_i > \gamma} \mathrm{e}^{-\sum_{i=1}^n x_i} \sum_{i=1}^n x_i \mathrm{d}\mathbf{x}}{n \int_{\sum x_i > \gamma} \mathrm{e}^{-\sum x_i} \mathrm{d}\mathbf{x}} = \frac{1}{n}\mathbb{E}[Y \,|\, Y > \gamma] \,,$$

where $Y \stackrel{d}{=} \mathsf{Gamma}(n,1)$. Direct computation shows that $\mathbb{E}[Y \,|\, Y > \gamma] = \Gamma(n+1,\gamma)/\Gamma(n,\gamma)$. Since $\Gamma(n+1,\gamma) = n\,\Gamma(n,\gamma) + \gamma^n \mathrm{e}^{-\gamma}$ and $\Gamma(n,\gamma) = \gamma^{n-1}\mathrm{e}^{-\gamma}(1+\mathcal{O}(\gamma^{-1}))$, we have

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

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

Therefore, by Proposition 3.1, 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 3.2. *Under the same assumptions as in Proposition 3.1, if one sets* $v = \gamma/n + h$ *for some constant* h*, then*

$$\mathbb{E}_{1/v} Z(v)^2/\ell^2 = \frac{\mathrm{e}^n (n-1)!}{2n^n} \gamma \left(1 + \frac{2n-1}{2\gamma} + \frac{n}{4\gamma^2}(2n^2h^2 - 2nh + 3n - 2) + \mathcal{O}(\gamma^{-3})\right) \tag{7}$$

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 3.1, we have

$$\begin{aligned} \mathbb{E}_{1/v} Z(v)^2 &= \frac{v^n}{(2-1/v)^n} \frac{\Gamma(n, (2-1/v)\gamma)}{\Gamma(n)} \\ &= \frac{v^n \mathrm{e}^{-2\gamma} \mathrm{e}^{\gamma/v}}{(2-1/v)(n-1)!} \gamma^{n-1} \left(1 + \frac{n-1}{2-1/v}\gamma^{-1} + \mathcal{O}(\gamma^{-2})\right). \end{aligned} \tag{8}$$

Substituting $v = \gamma/n + h$ and using the expression for ℓ in (6) gives (7). The final statement in the proposition follows by setting $v = \gamma/n \sim v_{\mathrm{vm}}$. □

It is worth noting that in (7) 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 multi-level CE estimator generally works well in problems with light-tailed random variables, but sometimes breaks down when the dimension of the problem gets large, e.g., see [8].

We now investigate what happens when the importance sampling parameter v is obtained via a random procedure, such as in the multi-level VM or CE methods. 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

$$V = \frac{\sum_{k=1}^{N} \mathbb{1}\{S_k > \gamma\} W_k(w) S_k}{n \sum_{k=1}^{N} \mathbb{1}\{S_k > \gamma\} W_k(w)}, \tag{9}$$

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

Proposition 3.3. *Under the same assumptions as in Proposition 3.1, the CE reference parameter V given in* (9) *is asymptotically normal as $N \to \infty$ with mean v_{ce} and variance $\sigma^2_{\gamma,w}/N$. Furthermore, we have*

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

In particular,

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

Proof. First note that V given in (9) is a ratio estimator. By the delta method [7], the asymptotic distribution is normal with mean

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

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

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

where $A = \mathbb{1}\{S > \gamma\}W(w)S$, $B = \mathbb{1}\{S > \gamma\}W(w)$, and S is $\mathsf{Gamma}(n, 1/w)$ distributed. The second moment of B is given in (8) with w substituted for v. The expectation of A is simply $\mathbb{E}_{1/w}A = \mathbb{E}_1\mathbb{1}\{S > \gamma\}S = \ell\, v_{\mathrm{ce}}$. The second moment of A is

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

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

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

□

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 multi-level CE procedure might be so substantial that it renders the resulting importance sampling estimator unreliable.

4. 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 $\ell = \mathbb{P}(X_1 + \cdots + X_n > \gamma)$ via importance sampling, where $X_i \overset{iid}{=} \mathsf{Pareto}(1,1)$, $i = 1,\ldots,n$. Since the Pareto distribution is subexponential [2], 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 $f(x;\alpha) = \alpha(1+x)^{\alpha+1}, x \geqslant 0$. Now suppose that we generate $X_i \overset{iid}{=} \mathsf{Pareto}(\alpha,1)$, $i = 1,\ldots,n$. The general term of the likelihood ratio is

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

and the corresponding importance sampling estimator is $Z(\alpha) = \mathbb{1}\{X_1 + \cdots + X_n > \gamma\}W(\mathbf{X};1,\alpha)$. In the following proposition we show that the optimal VM and CE parameters are the identical. In fact, we show $\alpha_{\mathrm{vm}} \sim n/\log(1+\gamma)$, which gives the minimum variance estimator within the the class of importance sampling estimators

obtained by generating $X_i \overset{iid}{=} \mathsf{Pareto}(\alpha, 1)$ for $i = 1, \ldots, n$. Compare this with the suggestions of [1] and [10].

Proposition 4.1. *Let $X_i \overset{iid}{=} \mathsf{Pareto}(1, 1)$, $i = 1, \ldots, n$. Suppose we wish to estimate $\ell = \mathbb{P}(X_1 + \cdots + X_n > \gamma)$ via importance sampling by generating $X_i \overset{iid}{=} \mathsf{Pareto}(\alpha, 1)$. 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 [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:

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

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

$$\mathbb{E}_\alpha Z(\alpha)^2 \sim (2\alpha - \alpha^2)^{-n} n (1+\gamma)^{-(2-\alpha)}. \tag{10}$$

By a computation similar to that in Proposition 3.1, we have

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

Again, the optimal VM parameter is asymptotically identically 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 $\mathbb{E}_\alpha Z(\alpha)^2/\ell^2$. As a corollary to Proposition 4.1 we show that $\alpha = \alpha_{\mathrm{ce}} \sim n/\log(1+\gamma)$ gives an importance sampling estimator that is asymptotically optimal. We note that [3] provide 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], [6] are able to derive an importance sampling estimator that achieves bounded relative error for general subexponential distributions.

Proposition 4.2. *Under the same assumptions as in Proposition 4.1, if one sets* $\alpha = n/\log(1+\gamma) + h$ *for some constant* h *such that* $0 < \alpha < 2$*, then*

$$\frac{\mathbb{E}_\alpha Z(\alpha)^2}{\ell^2} \sim \frac{\mathrm{e}^n}{n}\gamma^h \left(h(2-h) + \frac{2n}{\log(1+\gamma)} - \frac{n^2}{\log^2(1+\gamma)} \right)^{-n} \quad \text{as } \gamma \to \infty. \tag{11}$$

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

Proof. By (10) and the fact that $\ell \sim n/(1+\gamma)$, we have

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

Hence, if we set $\alpha = n/\log(1+\gamma) + h$, then (11) follows. As a result, for $\alpha = \alpha_{\mathrm{ce}} \sim n/\log(1+\gamma)$ we have

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

□

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/CE reference parameters is expected to be more severe in the heavy-tailed case.

5. Sum of Weibull Random Variables

Consider the same estimation problem as in the last section, but now $X_i \overset{iid}{=} \mathsf{Weib}(\beta, 1)$, $i = 1, \ldots, n$ 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 $f(x;\beta,\theta) = \theta\beta x^{\beta-1}\mathrm{e}^{-\theta x^\beta}, x \geqslant 0$ indexed by $\theta > 0$ while keeping β fixed. It follows that the general term of the importance sampling estimator is $Z(\theta) = \mathbb{1}\{X_1 + \cdots + X_n > \gamma\} W(\mathbf{X}; 1, \theta)$, with likelihood ratio $W(\mathbf{x}; 1, \theta) = \theta^{-n} \exp\left(-(1-\theta)\sum_{i=1}^n x_i^\beta\right)$. Again, for this sum of Weibull random variables case, the optimal VM and CE parameters coincide asymptotically.

Proposition 5.1. *Let* $X_i \overset{iid}{=} \mathsf{Weib}(\beta, 1)$*,* $i = 1, \ldots, n$ *with* $0 < \beta < 1$*. Suppose we wish to estimate* $\ell = \mathbb{P}(X_1 + \cdots + X_n > \gamma)$ *via importance sampling by generating* $X_i \overset{iid}{=} \mathsf{Weib}(\beta, \theta)$*. 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 [4]: $\theta_{\text{ce}} = n/(n+\gamma^\beta) \sim n/\gamma^\beta$. Next we compute the optimal VM parameter as follows.

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

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

$$\mathbb{E}_\theta Z(\theta)^2 \sim \frac{n}{\theta^n(2-\theta)^n}\mathbb{P}(Y_1 > \gamma) = \frac{n}{\theta^n(2-\theta)^n}\mathrm{e}^{-(2-\theta)\gamma^\beta} \quad \text{as } \gamma \to \infty. \tag{12}$$

By a similar computation as in Proposition 3.1, it can be shown that

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

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 (12), it can be shown that if one chooses θ such that $\theta\gamma^\beta = c$ for some constant c, the associated importance sampling estimator is asymptotically optimal. In particular, the choice $\theta = \theta_{\text{ce}} \sim \theta_{\text{vm}}$ 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 $X_i \overset{iid}{=} \mathsf{Weib}(\beta, \theta), i = 1, \ldots, n$.

Proposition 5.2. *Under the same assumptions as in Proposition 5.1, if one sets $\theta = n/\gamma^\beta$, then*

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

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

Proof. Since the $\mathsf{Weib}(\beta, 1)$ distribution is subexponential for $\beta < 1$, we have $\ell \sim n\mathrm{e}^{-\beta\gamma}$. It follows from (12) that $\mathbb{E}_\theta Z(\theta)^2/\ell^2 \sim n^{-1}\theta^{-n}(2-\theta)^{-n}\mathrm{e}^{\theta\gamma^\beta}$, as $\gamma \to \infty$. The desired result follows by letting $\theta = n/\gamma^\beta$.

6. Sum of Independent Non-identical 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 $X_1, X_2, \ldots$ 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

$$f_j(x;u_j) = \mathrm{e}^{x\theta(u_j)-\zeta(\theta(u_j))}h_j(x). \tag{13}$$

Let $\mu_n = \sum_{i=1}^n \mathbb{E}X_i$ and $\sigma_n^2 = \sum_{i=1}^n \mathrm{Var}\ X_i$. We are interested in estimating the probabilities $\ell_n = \mathbb{P}(S_n > n\,b)$ as $n \to \infty$, where $S_n = X_1 + \cdots + X_n$ and $\lim_{n\to\infty}(n\,b - \mu_n)/\sigma_n = \infty$. We will show that the optimal CE parameters coincide with the ones suggested by large deviation theory. More specifically, the CE method suggests twisting the means of the random variables such that their sum equal to the threshold $n\,b$.

Proposition 6.1. *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* (13)*. Consider estimating $\ell_n = \mathbb{P}(S_n > n\,b)$ 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 $\lim_{n\to\infty}(n\,b - \mu_n)/\sigma_n = \infty$, where $\mu_n = \sum_{i=1}^n \mathbb{E}X_i$ and $\sigma_n^2 = \sum_{i=1}^n \mathrm{Var}\ X_i$. Then the optimal CE parameters $v^*_{\mathrm{ce},1}, v^*_{\mathrm{ce},2}, \ldots$ satisfy*

$$\sum_{i=1}^n v^*_{\mathrm{ce},i} \sim n\,b \quad \text{as } n \to \infty.$$

Proof. First note that to estimate ℓ_n, the optimal CE parameters $v^*_{\mathrm{ce},i},\ i = 1, \ldots, n$ are given by [16, page 320]: $v^*_{\mathrm{ce},i} = \mathbb{E}[X_i \mid S_n > n\,b]$. Therefore, $\sum_{i=1}^n v^*_{\mathrm{ce},i} = \mathbb{E}[S_n \mid S_n > n\,b]$. By the central limit theorem, $(S_n - \mu_n)/\sigma_n$ is asymptotically $\mathsf{N}(0,1)$ distributed as $n \to \infty$. Therefore,

$$\mathbb{E}[S_n \mid S_n > n\,b] \sim \mu_n + \frac{\varphi\left(\frac{n\,b-\mu_n}{\sigma_n}\right)}{1-\Phi\left(\frac{n\,b-\mu_n}{\sigma_n}\right)}\sigma_n \sim \mu_n + \frac{n\,b-\mu_n}{\sigma_n}\sigma_n = n\,b,$$

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. □

7. Concluding Remarks and Future Research

We compare the VM and CE methods through various concrete examples and we find 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 obtained, 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 further study 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., Binswanger, K. and Højgaard, B. (2000). Rare events simulation for heavy-tailed distributions. *Bernoulli* **6,** 303–322.

[2] Asmussen, S. and Glynn, P. W. (2007). *Stochastic simulation: algorithms and analysis*. Springer-Verlag, New York.

[3] Asmussen, S. and Kroese, D. P. (2006). Improved algorithms for rare event simulation with heavy tails. *Advances in Applied Probability* **38,** 545–558.

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

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

[6] Blanchet, J. and Li, C. (2011). Efficient rare event simulation for heavy-tailed compound sums. *ACM Transactions on Modeling and Computer Simulation* **21,** forthcoming.

[7] Casella, G. and Berger, R. L. (2001). *Statistical Inference* second ed. Duxbury Press.

[8] Chan, J. C. C. and Kroese, D. P. (2010). Improved cross-entropy method for estimation. *Technical report*. The University of Queensland Brisbane, Australia.

[9] Chan, J. C. C. and Kroese, D. P. (2010). Rare-event probability estimation with conditional Monte Carlo. *Annals of Operations Research*. Forthcoming.

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

[11] Kroese, D. P. (2010). The cross-entropy method. In *Wiley Encyclopedia of Operations Research and Management Science*. Forthcoming.

[12] L'Ecuyer, P., Blanchet, J. H., Tuffin, B. and Glynn, P. W. (2010). Asymptotic robustness of estimators in rare-event simulation. *ACM Transactions on Modeling and Computer Simulation* **20,** 1–41.

[13] Rubinstein, R. Y. (2009). The Gibbs cloner for combinatorial optimization, counting and sampling. *Methodology and Computing in Applied Probability* **11,** 491–549.

[14] Rubinstein, R. Y. and Glynn, P. W. (2009). How to deal with the curse of dimensionality of likelihood ratios in Monte Carlo simulation. *Stochastic Models* **25,** 547 – 568.

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

[16] Rubinstein, R. Y. and Kroese, D. P. (2007). *Simulation and the Monte Carlo Method Second Edition*. John Wiley & Sons, New York.