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

FITTING MIXTURE IMPORTANCE SAMPLING DISTRIBUTIONS VIA IMPROVED CROSS-ENTROPY

Tim J. Brereton

Department of Mathematics
University of Queensland
Brisbane 4072, Australia

Joshua C.C. Chan

Research School of Economics
Australian National University
Canberra 0200, Australia

Dirk P. Kroese

Department of Mathematics
University of Queensland
Brisbane 4072, Australia

ABSTRACT

In some rare-event settings, exponentially twisted distributions perform very badly. One solution to this problem is to use mixture distributions. However, it is difficult to select a good mixture distribution for importance sampling. We here introduce a simple adaptive method for choosing good mixture importance sampling distributions.

1 INTRODUCTION

The estimation of rare-event probabilities is an area of Monte Carlo simulation that is of both practical and theoretical interest. It has practical applications in areas as diverse as finance, systems biology and high-energy physics — see for example Rubino and Tuffin (2009) and Kroese, Taimre, and Botev (2011). Rare-event estimation is also of considerable theoretical interest, and has contributed to the development of many novel Monte Carlo methods. One of the central techniques that is employed to improve the speed and accuracy with which rare-event probabilities are estimated is *importance sampling*. The idea is to effect a change of probability measure so that the rare event of interest is no longer rare. A popular choice of change of measure is to employ *exponential twisting*. The problem of efficient importance sampling then reduces to one of finding good twisting parameters. However, as was first demonstrated in Glasserman and Wang (1997), exponential twisting can result in estimators that perform very badly. There are a number of approaches to address this problem, including Dupuis and Wang (2005), Owen and Zhou (2000) and Boots and Mandjes (2002). In this paper, we concentrate on using importance sampling distributions that are *mixtures* of distributions, an approach first suggested by Sadowsky and Bucklew (1990). A major limitation to the use of mixture distributions in importance sampling is the difficulty in quickly and effectively finding good parameters for these distributions. As a result, mixture distributions do not appear to be widely used as importance sampling distributions. In this paper, we demonstrate how the improved Cross-Entropy (CE) idea of Chan (2010) provides a particularly straightforward approach to finding good mixture importance sampling distributions. We give numerical results for several problems where exponential twisting fails. These show that the improved CE estimator performs as well as the asymptotically efficient estimator when it is known, and outperforms a single exponentially twisted distribution.

2 IMPORTANCE SAMPLING AND EXPONENTIAL TWISTING

The primary aim of rare-event simulation is to estimate quantities of the form

$$\ell = \mathbb{P}_f(A) = \mathbb{E}_f \mathbb{I}(A), \tag{1}$$

where A is a rare event pertaining to some random object $\mathbf{X}$ (formally, A is in the σ-algebra generated by $\mathbf{X}$, that is $A \in \sigma(\mathbf{X})$), $\mathbb{I}$ is the indicator function and $\mathbb{E}_f$ indicates that the expectation is calculated with respect to some probability density f (more precisely, $\mathbf{X}$ has the probability density function f). Let

$$I = \mathbb{I}(A).$$

Then, ℓ can be estimated directly by the *Crude Monte Carlo* estimator

$$\hat{\ell} = \frac{1}{N}\sum_{i=1}^{N} I_i,$$

where each I_i is an independent and identically distributed replicate of I. Because the event A is rare, this estimator is highly inefficient. The idea of importance sampling is that (1) can be rewritten to get an estimator under which A is much more likely. The expectation under f is replaced by an expectation under g as follows

$$\ell = \mathbb{E}_f \mathbb{I}(A) = \mathbb{E}_g \frac{f(\mathbf{X})}{g(\mathbf{X})}\mathbb{I}(A),$$

provided that $g(\mathbf{x}) = 0$ whenever $\mathbb{I}(A)f(\mathbf{x}) = 0$. This probability can now be estimated using the importance sampling estimator

$$\hat{\ell} = \frac{1}{N}\sum_{i=1}^{N} \frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)} I_i, \tag{2}$$

where the $\mathbf{X}_i$ (and I_i) are simulated under g. The aim is to choose an importance sampling density g so that the error of the estimator (2) is in some sense minimized. Theoretically, the best case scenario is where the estimator has zero variance. We can achieve this by choosing g to be equal to the conditional density

$$g^*(\mathbf{x}) = \frac{\mathbb{I}(A)f(\mathbf{x})}{\mathbb{P}(A)} = f(\mathbf{x}|A).$$

Practically, this choice is not feasible, because the normalizing constant $\mathbb{P}(A)$ is the very quantity we are trying to estimate. It does, however, give insight into what a good importance sampling distribution should look like. In particular, the importance sampling distribution should not change the ways in which a rare event occurs. More precisely, if a rare event can happen in several different ways, the relative proportions with which these different ways occur ought not to be changed under the importance sampling distribution. An often used choice of importance sampling distribution is an *exponentially twisted* version of the original distribution. That is, the importance sampling distribution is taken as

$$f_\theta(\mathbf{x}) = \frac{\exp\{\theta^\intercal \mathbf{x}\} f(\mathbf{x})}{\mathbb{E}\exp\{\theta^\intercal \mathbf{X}\}},$$

where θ is a vector of parameters that allows us to control the shape of the distribution. This is a particularly common choice in a light-tailed setting (where $\mathbb{E}_f \exp\{\theta_i X_i\} < \infty$ for all θ_i in a neighborhood of zero and all X_i that are elements of $\mathbf{X}$). Exponential twisting is an attractive approach because the twisted distributions are straightforward to draw from and the likelihood ratio is straightforward to compute. In simple settings, one can either solve directly for the θ that minimizes the variance of the estimator or find a θ such that the asymptotic rate at which the variance of the estimator goes to zero is maximized. In more complicated settings, adaptive methods such as the CE method or the Variance Minimization method can be employed to find good importance sampling distributions — see Asmussen and Glynn (2007) or Kroese, Taimre, and Botev (2011). However, as described below, exponential twisting may lead to poorly performing estimators.

3 TWO PROBLEMS

The following two problems give examples of settings where exponentially twisting a distribution gives a poorly performing estimator. The common feature is that the rare event can happen in more than one way. In these settings, exponential twisting will put too much emphasis on certain occurrences of the rare event and not enough on others.

3.1 Problem 1: An n-Step Random Walk Hitting a Non-Convex Set

Given n standard normal random variables $X_1,\ldots,X_n$, we define $S_n = X_1+\cdots+X_n$. We wish to estimate

$$\mathbb{P}(A) = \mathbb{P}\left(\{S_n > na\} \cup \{S_n < -n(a+\varepsilon\}\right),$$

where $\varepsilon > 0$. This event naturally decomposes into two events A_1 and A_2, where $A_1 = \{S_n > na\}$ and $A_2 = \{S_n < -n(a+\varepsilon)\}$. Here, twisting the distribution of the X_i corresponds to changing the means of the variables. In doing so, we make one event more probable and the other event less probable. In this case, it makes sense to twist towards the more likely, and thus asymptotically important, event A_1. However, doing this results in a highly skewed distribution for the estimator (2). As A_2 becomes rarer, the likelihood ratio $f(\mathbf{x})/f_\theta(\mathbf{x})$ can take on increasingly large values. This is because $\mathbb{P}_f(A_2) \gg \mathbb{P}_g(A_2)$. If a and ε are such that $a+\varepsilon+\theta^2/2 > 1$ then it is easy to see that when A_2 occurs, the likelihood ratio is bounded from below by e^n. In the worst case, this can result in an estimator with asymptotically infinite variance.

3.2 Problem 2: A Weighted Sum of Indicator Functions

Let $X_1,\ldots,X_5$ be $\mathsf{Exp}(1)$ random variables. We wish to estimate the following probability

$$\mathbb{P}(A) = \mathbb{P}\Big(1\mathbb{I}(X_1 > \gamma) + 2\mathbb{I}(X_2 > \gamma) + 3\mathbb{I}(X_3 > \gamma) + 4\mathbb{I}(X_4 > \gamma) + 5\mathbb{I}(X_5 > \gamma) \geq 6\Big),$$

where $\gamma > 0$. This rare-event can be represented as the union of different events in a variety of ways. For example, we will adopt the decomposition $A_1 = \{1\mathbb{I}(X_1 > \gamma) + 5\mathbb{I}(X_5 > \gamma) \geq 6\}$, …, $A_6 = \{3\mathbb{I}(X_3 > \gamma) + 4\mathbb{I}(X_4 > \gamma) \geq 6\}$, where these six events represent the different ways that the event can happen with two indicators being on (equal to 1). Unlike the preceding example, these events have a non-zero intersection. For example, it is possible that both $1\mathbb{I}(X_1 > \gamma) + 2\mathbb{I}(X_6 > \gamma) > 6$ and $3\mathbb{I}(X_3 > \gamma) + 4\mathbb{I}(X_4 > \gamma) \geq 6$. If we use a single twisted distribution in this problem, we will effectively increase the probability of each indicator being on. However, we do so in a way the makes some of the events $A_1,\ldots,A_6$ more likely than others, even though they should clearly happen with equal probabilitity. For example,if $\gamma = 10$ and we use the multi-level CE approach, we get the following probabilities $\mathbb{P}_g(X_1 > 10) \approx 0.06$, $\mathbb{P}_g(X_2 > 10) \approx 0.09$, $\mathbb{P}_g(X_3 > 10) \approx 0.04$, $\mathbb{P}_g(X_4 > 10) \approx 0.17$ and $\mathbb{P}_g(X_5 > 10) \approx 0.28$. This means that under our CE importance sampling distribution, $\mathbb{P}_g(A_1) = \mathbb{P}_g(X_1 > \gamma)\mathbb{P}_g(X_5 > \gamma) \approx .02$, whereas $\mathbb{P}_g(A_4) = \mathbb{P}_g(X_4 > \gamma)\mathbb{P}_g(X_5 > \gamma) = .05$. In addition, none of the events happen sufficiently often. The numerical results given below show that, as a result, the CE estimator has a high relative error compared to an estimator using a mixture distribution.

4 Solutions

When faced with problems where exponential twisting does not work, there are at least three solutions, as described in Blanchet and Mandjes (2009). We can

1. Decompose the rare-event into disjoint events and estimate the probability of each one of these separately. This is the approach in Boots and Mandjes (2002).
2. Use a state dependent estimator, as in Dupuis and Wang (2005).
3. Use a mixture of exponentially twisted distributions.

The first option is attractive if the event naturally decomposes into a finite number of disjoint events, as in the first problem given above. However, if the event does not have a natural disjoint decomposition, as in Problem 2, this method becomes difficult to implement. The second option requires considerable knowledge of the problem at hand and tends to be difficult to implement. We focus here on the third option. It is our belief that this approach is straightforward and generally applicable with minimum knowledge of the structure of the problem. The idea is that, instead of using a single twisted distribution, we use a mixture of L exponentially twisted distributions, with pdf

$$f_\theta(\mathbf{x}) = \sum_{i=1}^{L} \pi_i f_{\theta_i}(\mathbf{x}),$$

where the π_i are positive and sum to 1. In settings where the twisting parameters of the mixture distribution can be found using Large Deviations type arguments, the resulting estimator can be shown to be asymptotically efficient, see Bucklew (2004). Practically, however, it may not be obvious what a good choice of twisting parameters should be. We show in the following, that we are able to use adaptive methods to fit mixture distributions with strong evidence that this approach outperforms existing adaptive methods.

5 THE IMPROVED CROSS-ENTROPY METHOD

The idea of the CE method is to reduce the Cross-Entropy 'distance' between the zero-variance pdf $g^*(\mathbf{x})$ and a proposed importance sampling pdf $g(\mathbf{x};\theta)$. We choose the distribution parameters θ to solve the following stochastic program

$$\mathbf{v}_{CE} = \underset{\mathbf{v}}{\operatorname{argmax}} \, \mathbb{E}_{g^*} \log g(\mathbf{X};\theta).$$

This is usually estimated via a levels approach as in Rubinstein and Kroese (2004). However, with the improved CE method, detailed in Chan, Glynn, and Kroese (2011), we adopt a direct approach. We draw a sample $(\mathbf{Y}_1,\ldots \mathbf{Y}_M)$ of size M approximately from g^*, and then use the estimator

$$\widehat{\mathbf{v}}_{CE} = \underset{\mathbf{v}}{\operatorname{argmax}} \frac{1}{M}\sum_{j=1}^{M} \log g(\mathbf{Y}_j;\mathbf{v}). \tag{3}$$

If the distribution we are trying to fit is a mixture distribution, then this problem does not have a closed-form solution. However, we resolve this problem as follows. We split the rare event of interest A into a series of not necessarily disjoint events $A_1,\ldots,A_L$. We associate a distinct twisted distribution to each one of these events and estimate its twisting parameters via (3), using only those elements of the sample $\mathbf{Y}_1,\ldots,\mathbf{Y}_M$ that correspond to that event occurring. We then use the normalized proportions with which each event occurs as estimates for the weights $\pi_1,\ldots,\pi_l$ of the distribution.

Algorithm 5.1 (Improved CE) Let A be a rare event decomposed into rare-events $A_1,\ldots,A_L$.

1. Draw a sample $\mathbf{Y}_1,\ldots,\mathbf{Y}_M$ that is approximately from g^*.
2. Set $i=1$.
3. Identify the set $Y_{A_i} = \{\text{All elements of } \mathbf{Y}_1,\ldots,\mathbf{Y}_M \text{ such that } A_i \text{ occurs}\}$.
4. Set $\pi_i = |Y_{A_i}|$. Find $\theta_i = \operatorname{argmax}_\theta \sum_{Y_{A_i}} \log f_\theta(\mathbf{Y}_j;\theta)$.
5. If $i<L$, set $i=i+1$ and repeat from Step 3.
6. Normalize the π_i so that $\sum_{i=1}^{L}\pi_i = 1$.
7. Draw $X_1,\ldots,X_N$ from $\sum_i^L \pi_i f_\theta(\mathbf{x};\theta_i)$.
8. Estimate ℓ via
$$\hat{\ell} = \frac{1}{N}\sum_{i=1}^{N} \frac{f(\mathbf{X}_i)}{\sum_{j=1}^{L}\pi_j f_\theta(\mathbf{X}_i;\theta_j)} I_i.$$

6 DRAWING FROM THE ZERO-VARIANCE DISTRIBUTION

The improved CE method assumes that we are able to draw from g^*. In general, this necessitates the use of Markov Chain Monte Carlo (MCMC) methods. However, the structure of g^* means that it is often particularly difficult to sample from. In particular, the zero-variance distributions of the situations that we are looking at are characterized by multiple modes separated by regions of very low probability. In these settings, most MCMC algorithms fail to mix properly. As such, care must be taken that an appropriate MCMC sampler is used. We use a Generalized Splitting algorithm as in Botev and Kroese (2011) to draw from g^*. This algorithm is specially designed to ensure that the chain mixes well in a multi-modal setting.

7 NUMERICAL RESULTS

In the following, we give numerical results comparing the performance of our estimators to a number of other estimators. In Tables 1, 2 and 3, the improved CE estimator is applied to Problem 1. We use a mixture of two distributions, one twisted to make $\{S_n > na\}$ more likely, the other twisted to make $\{S_n < -n(a+\varepsilon)\}$ more likely. We contrast the performance of our estimator with the asymptotically efficient estimator suggested in Glasserman and Wang (1997), which is also a mixture distribution. We also include, as a reference point, the flawed naïve Large Deviations estimator. This estimator significantly underestimates the probability. Note that it also severely underestimates the relative error. This is due to the highly skewed distribtion of the estimator - see Bucklew (2004) for a discussion. The simulation was carried out with a sample size of $N = 10^5$. The numerical results show that the improved CE estimator has a similar level of performance to the asymptotically efficient estimator.

For the second problem, we contrast the performance of the improved CE estimator against the performance of the standard CE estimator (with a single exponentially twisted distribution). We also compare this with a distribution twisted so that for $i = 1,\dots,5$, $\mathbb{E}X_i = \gamma$. For the improved CE estimator, we use a mixture of six distributions - each one corresponding to one of the events $A_1,\dots,A_6$ described above. The simulation was carried out with a sample size of $N = 10^5$. The numerical results, given in Tables 4, 5 and 6, show that the improved CE estimator outperforms the standard CE estimator (which uses a single exponentially twisted distribution) and the other importance sampling estimator in this setting, and does much better as the event gets rarer.

Table 1: Performance of different estimators for Problem 1, with $n = 10$, $a = 1$ and $\varepsilon = .01$.

Estimator	$\hat{\ell}$	Estimated relative error
Actual value	$.00148\dots$	n/a
Asymptotically efficient (mixture)	$.00149$	.006
Improved CE (mixture)	$.00148$	.006
LD (naïve)	7.89×10^{-4}	.006

Table 2: Performance of different estimators for Problem 1, with $n = 10$, $a = 1.5$ and $\varepsilon = .01$.

Estimator	$\hat{\ell}$	Estimated relative error
Actual value	$1.949\dots \times 10^{-6}$	n/a
Asymptotically efficient (mixture)	1.97×10^{-6}	.007
Improved CE (mixture)	1.9×10^{-6}	.007
LD (naïve)	1.053×10^{-6}	.007

Table 3: Performance of different estimators for Problem 1, with $n = 20$, $a = 1$ and $\varepsilon = .01$.

Estimator	$\hat{\ell}$	Estimated relative error
Actual value	$7.01\dots \times 10^{-6}$	n/a
Asymptotically efficient (mixture)	6.96×10^{-6}	.007
Improved CE (mixture)	7.04×10^{-6}	.007
LD (naïve)	3.9×10^{-6}	.007

Table 4: Performance of different estimators for Problem 2, with $\gamma = 6$.

Estimator	$\hat{\ell}$	Estimated relative error
Actual value	$3.67\ldots \times 10^{-5}$	n/a
CE	3.46×10^{-5}	.038
Improved CE (mixture)	3.79×10^{-5}	.027
IS ($\mathbb{E}X_i = \gamma$)	3.67×10^{-5}	.066

Table 5: Performance of different estimators for Problem 2, with $\gamma = 8$.

Estimator	$\hat{\ell}$	Estimated relative error
Actual value	$6.749\ldots \times 10^{-7}$	n/a
CE	6.27×10^{-7}	.062
Improved CE (mixture)	6.78×10^{-7}	.037
IS ($\mathbb{E}X_i = \gamma$)	6.85×10^{-7}	.161

Table 6: Performance of different estimators for Problem 2, with $\gamma = 10$.

Estimator	$\hat{\ell}$	Estimated relative error
Actual value	$1.24\ldots \times 10^{-8}$	n/a
CE	1.59×10^{-8}	.224
Improved CE (mixture)	1.17×10^{-8}	.045
IS ($\mathbb{E}X_i = \gamma$)	1.128×10^{-8}	.215

8 FURTHER RESEARCH

We believe that the adaptive approach to finding mixture importance sampling distributions has yet to be fully explored. There are several areas where further research is warranted. One is finding an optimal way to select the weights in the mixture distribution. The algorithm given above uses a fairly naïve approach, using the normalized empirical proportions with which each event occurs. A better algorithm might take account of higher moments in choosing proportions. Another issue is deciding an optimal approach to splitting a rare event A into sub-events $A_1, \ldots, A_L$. For example, in Problem 2, it might be sufficient to use a mixture of two different distributions. It could be possible to adaptively decide on an optimal partition of A.

ACKNOWLEDGEMENTS

Dirk Kroese acknowledges the support of the Australian Research Council under grant number DP0985177.

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic simulation*. New York: Springer-Verlag.

Blanchet, J., and M. Mandjes. 2009. Rare event simulation for queues. In *Rare Event Simulation*, ed. G. Rubino and B. Tuffin, 87–124. Hoboken, New Jersey: John Wiley & Sons.

Boots, N.-K., and M. Mandjes. 2002. Fast simulation of a queue fed by a superposition of many (heavy-tailed) sources. *Probability in the Engineering and Informational Sciences* 40:1104–1128.

Botev, Z., and D. P. Kroese. 2011. Efficient Monte Carlo simulation via the generalized splitting method. *Statistics and Computing*.

Bucklew, J. A. 2004. *Introduction to rare event simulation*. New York: Springer-Verlag.

Chan, J. C. C. 2010. *Advanced Monte Carlo methods with applications in finance*. Ph. D. thesis, University of Queensland.

Chan, J. C. C., P. W. Glynn, and D. P. Kroese. 2011. A comparison of cross-entropy and variance minimization strategies. *Journal of Applied Probability*.

Dupuis, P., and H. Wang. 2005. Dynamic importance sampling for uniformly recurrent Markov chains. *The Annals of Applied Probability* 15 (1A): 1–38.

Glasserman, P., and Y. Wang. 1997. Counterexamples in importance sampling for large deviations probabilities. *The Annals of Applied Probability* 7 (3): 731–746.

Kroese, D. P., T. Taimre, and Z. I. Botev. 2011. *Handbook of Monte Carlo methods*. New York: John Wiley & Sons.

Owen, A., and Y. Zhou. 2000. Safe and effective importance sampling. *Journal of the American Statistical Association* 95 (449): 135–143.

Rubino, G., and B. Tuffin. (Eds.) 2009. *Rare event simulation using Monte Carlo methods*. Hoboken, New Jersey: John Wiley & Sons.

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

Sadowsky, J., and J. Bucklew. 1990. On large deviations theory and asymptotically efficient Monte Carlo estimation. *IEEE transactions on Information Theory* 36:579–588.

AUTHOR BIOGRAPHIES

Tim Brereton is a PhD student at the University of Queensland. He has a Bachelor of Science (Honours) in Mathematics and a Masters of International Economics and Finance, both from the University of Queensland. His research interests include simulation, computational statistics and mathematical finance. His email address is tim.brereton@uqconnect.edu.au.

Joshua C. C. Chan is a lecturer at the Research School of Economics, Australian National University. He completed his Ph.D. at the Department of Mathematics, University of Queensland. His research interests lie in the field of Monte Carlo simulation, especially the methodological and computational issues of adaptive importance sampling and Markov chain Monte Carlo methods, with an emphasis on economics and finance applications. His email address is joshua.chan@anu.edu.au.

Dirk Kroese is an Australian Professorial Fellow in Statistics at the School of Mathematics and Physics of the University of Queensland. He has held teaching and research positions at Princeton University, the University of Twente, the University of Melbourne, and the University of Adelaide. His research interests include Monte Carlo methods, adaptive importance sampling, randomized optimization, and rare-event simulation. He has over 70 peer-reviewed publications, including three monographs: Simulation and the Monte Carlo Method, 2nd Edition, 2007, John Wiley & Sons (with R.Y. Rubinstein), The Cross-Entropy Method, 2004, Springer-Verlag, (with R.Y. Rubinstein), and the Handbook of Monte Carlo Methods, 2011, John Wiley & Sons (with T. Taimre and Z.I. Botev). He is serving on the editorial boards of Methodology and Computing in Applied Probability and The Annals of Operations Research. His website is http://www.maths.uq.edu.au. His email address is kroese@maths.uq.edu.au