

# IMPROVED CROSS-ENTROPY METHOD FOR ESTIMATION

Joshua C.C. Chan<sup>1</sup>

Dirk P. Kroese<sup>2</sup>

<sup>1</sup>Research School of Economics, Australian National University,  
Canberra, ACT 0200, Australia

<sup>2</sup> Department of Mathematics, University of Queensland,  
Brisbane, QLD 4072, Australia

April 2011

## Abstract

The cross-entropy (CE) method is an adaptive importance sampling procedure that has been successfully applied to a diverse range of complicated simulation problems. However, recent research has shown that in some high-dimensional settings, the likelihood ratio degeneracy problem becomes severe and the importance sampling estimator obtained from the CE algorithm becomes unreliable. We consider a variation of the CE method whose performance does not deteriorate as the dimension of the problem increases. We then illustrate the algorithm via a high-dimensional estimation problem in risk management.

Keywords: cross-entropy, variance minimization, importance sampling, Kullback-Leibler divergence, rare-event simulation, likelihood ratio degeneracy,  $t$  copula.

Acknowledgment: We gratefully acknowledge the financial support by the Australian Research Council under Discovery Grants DP0985177 and DP0987170.

## 1 Introduction

The cross-entropy (CE) method is a versatile adaptive Monte Carlo algorithm originally developed for rare-event simulation by Rubinstein (1997, 1999). Since its inception, it has been applied to a diverse range of difficult simulation problems, such as network reliability estimation in telecommunications (Hui et al., 2005; Ridder, 2005), efficient simulation of buffer overflow probabilities in queuing networks (de Boer et al., 2004), estimation of large portfolio loss probabilities in credit risk models (Chan and Kroese, 2010), adaptive proposal design for Markov chain Monte Carlo (MCMC) methods (Keith et al., 2008) and particle filtering (Cornebise et al., 2008), marginal likelihood computation in Bayesian statistics (Chan and Eisenstat, 2011), and other rare-event probability estimation problems involving light- and heavy-tailed random variables (Kroese and Rubinstein, 2004; Asmussen et al., 2005). A recent review of the CE method and its applications can be found in Kroese (2011); a book-length treatment is given in Rubinstein and Kroese (2004). Despite its wide applicability, recent research has shown that in

some high-dimensional settings, the likelihood ratio degeneracy problem becomes severe and the importance sampling estimator obtained from the CE algorithm is unreliable (e.g., see Rubinstein and Glynn, 2009; Chan and Kroese, 2011). This calls for new approaches that can handle rare-event probability estimation in high-dimensional settings.

The purpose of this paper is threefold. First, we show why the multi-level CE method often breaks down in high-dimensional problems. In fact, we demonstrate that it fails because the importance density obtained from the multi-level procedure is suboptimal. Second, we introduce a new variant of the CE method that can at least ameliorate the degeneracy problem. This is achieved by obtaining the importance density in one single step, thus avoiding the multi-level procedure altogether. We demonstrate that this simple twist of the CE method gives an estimator that is accurate even in high-dimensional settings. We further show that the proposed approach provides a practical way to locate the minimum variance importance sampling estimator within any given parametric family of densities. Lastly, we illustrate the utility of the proposed approach by applying it to a high-dimensional estimation problem in risk management — estimation of large portfolio loss probabilities under the recently proposed  $t$  copula model of Bassamboo et al. (2008). We show that this improved CE estimator outperforms existing importance sampling estimators.

It is worth mentioning that there is a related literature on estimating the normalizing constant of an arbitrary density by MCMC methods; see, for example, Gelfand and Dey (1994), Newton and Raftery (1994), Chib (1995), Chib and Jeliazkov (2001), Gelman and Meng (1998), among many others. Although these methods may be adapted to estimate rare-event probabilities, they are not suitable for these problems. This is because in rare-event simulation a high level of accuracy is typically required. Since MCMC draws often exhibit high autocorrelation, especially in high-dimensional settings, a substantial number of draws is needed to achieve the level of accuracy required. To compound the problem, MCMC draws are generally costly to obtain. Therefore, using these methods in rare-event settings is simply impractical. In contrast, the proposed method is essentially an importance sampling approach and, as such, it circumvents these drawbacks by generating independent draws from some convenient density, where the computational cost of obtaining extra draws is often trivial. Although the proposed approach also requires MCMC draws for obtaining the optimal importance density, the number of draws needed is typically small. It is therefore no surprise that in rare-event simulation, importance sampling is the dominant approach.

The rest of this article is organized as follows. In Section 2 we first describe the conventional CE method for rare-event probability estimation, and highlight the rationale for the multi-level approach. Section 3 suggests reasons why the multi-level CE method often breaks down in high-dimensional problems, and introduces a new variant that is robust against the curse of dimensionality. We then discuss how the proposed approach can be used to locate the importance sampling estimator with minimum variance within any parametric family in Section 4. Finally, in Section 5 we apply the proposed methodology to a high-dimensional estimation problem in risk management.

## 2 The CE Method for Rare-event Probability Estimation

In Monte Carlo simulation, one fundamental problem is the efficient estimation of the possibly high-dimensional integral of the form:

$$\mathbb{E}_f H(\mathbf{X}) = \int H(\mathbf{x})f(\mathbf{x})d\mathbf{x}, \quad (1)$$

where  $\mathbf{X}$  is an  $n \times 1$  vector of random variables with probability density function (pdf)  $f$  and  $H : \mathbb{R}^n \rightarrow \mathbb{R}$  is a real-valued function. In fact, one important class of difficult problems is the estimation of rare-event probabilities, where the real-valued function  $H$  takes the form  $H(\mathbf{x}) = \mathbf{1}(S(\mathbf{x}) \geq \gamma)$  for some sufficiently large threshold  $\gamma$  and performance function  $S$ . Then the estimation problem becomes

$$\ell = \mathbb{P}(S(\mathbf{X}) \geq \gamma) = \int \mathbf{1}(S(\mathbf{x}) \geq \gamma)f(\mathbf{x})d\mathbf{x}. \quad (2)$$

In what follows, we focus on the problem of estimating the rare-event probability in (2). The original estimation problem (1) can be tackled in the same way with minor modifications. One popular approach to solve the estimation problem (2) is via *importance sampling*: take a random sample of size  $M$  from an importance density  $g$  that dominates  $f$ , i.e.,  $g(\mathbf{x}) = 0 \Rightarrow \mathbf{1}(S(\mathbf{x}) \geq \gamma)f(\mathbf{x}) = 0$  for all  $\mathbf{x}$ , and compute

$$\widehat{\ell}_{\text{IS}} = \frac{1}{M} \sum_{i=1}^M \mathbf{1}(S(\mathbf{X}_i) \geq \gamma) \frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)}, \quad (3)$$

where  $\mathbf{X}_1, \dots, \mathbf{X}_M$  are iid draws from the importance density  $g$ , and the ratio  $f(\mathbf{X}_i)/g(\mathbf{X}_i)$  is known as the *likelihood ratio*. Although the estimator  $\widehat{\ell}_{\text{IS}}$  is consistent and unbiased for any given  $g$ , its performance depends critically on its choice.

It is well-known that the zero-variance importance density  $g^*$  is simply the conditional density given the rare event, i.e.,

$$g^*(\mathbf{x}) = f(\mathbf{x} | S(\mathbf{x}) \geq \gamma) = \ell^{-1} f(\mathbf{x}) \mathbf{1}(S(\mathbf{x}) \geq \gamma).$$

Since this density involves the unknown constant  $\ell$ , it cannot be used directly. However, one could choose an importance density within a parametric family that is in some sense the “closest” to  $g^*$ . The fundamental insight of the CE method is to formalize this strategy as an optimization problem as follows. Let  $f(\mathbf{x}) = f(\mathbf{x}; \mathbf{u})$  denote the nominal density, where we make explicit the dependence on the parameter vector  $\mathbf{u}$ . Consider the family of pdfs  $\mathcal{F} = \{f(\mathbf{x}; \mathbf{v})\}$  indexed by the parameter vector  $\mathbf{v}$  within which to obtain the optimal CE importance density  $g$ . Various considerations of choosing the parametric family  $\mathcal{F}$  are discussed in Section 4; for now we assume  $\mathcal{F}$  is given. One particularly convenient measure of the “distance” from a density  $h_1$  to another density  $h_2$  is the *Kullback-Leibler divergence*, or *cross-entropy distance*, which is defined as

$$\mathcal{D}(h_1, h_2) = \int h_1(\mathbf{x}) \log \frac{h_1(\mathbf{x})}{h_2(\mathbf{x})} d\mathbf{x}.$$

We then locate the density  $g$  such that  $\mathcal{D}(g^*, g)$  is minimized. Since  $g$  is chosen within the parametric family  $\mathcal{F}$ , we can write  $g(\mathbf{x}) = f(\mathbf{x}; \mathbf{v}_{\text{ce}}^*)$  where  $\mathbf{v}_{\text{ce}}^*$  is referred to as the *optimal*

*CE parameter vector.* Now the functional minimization problem of finding an optimal CE importance density  $g$  reduces to a parametric minimization problem of finding the optimal CE parameter vector  $\mathbf{v}^* = \operatorname{argmin}_{\mathbf{v}} \mathcal{D}(g^*, f(\cdot; \mathbf{v}))$ . Further, note that

$$\mathcal{D}(g^*, f(\cdot; \mathbf{v})) = \int g^*(\mathbf{x}) \log g^*(\mathbf{x}) d\mathbf{x} - \ell^{-1} \int f(\mathbf{x}; \mathbf{u}) \mathbf{1}(S(\mathbf{x}) \geq \gamma) \log f(\mathbf{x}; \mathbf{v}) d\mathbf{x},$$

where the first term on the right-hand side does not depend on  $\mathbf{v}$ . Therefore, solving the CE minimization problem is equivalent to finding

$$\mathbf{v}_{\text{ce}}^* = \operatorname{argmax}_{\mathbf{v}} \int f(\mathbf{x}; \mathbf{u}) \mathbf{1}(S(\mathbf{x}) \geq \gamma) \log f(\mathbf{x}; \mathbf{v}) d\mathbf{x}. \quad (4)$$

The deterministic problem (4) often does not admit an analytic solution. Instead, one can estimate  $\mathbf{v}_{\text{ce}}^*$  by finding

$$\hat{\mathbf{v}}_{\text{ce}}^* = \operatorname{argmax}_{\mathbf{v}} \frac{1}{N} \sum_{i=1}^N \mathbf{1}(S(\mathbf{X}_i) \geq \gamma) \log f(\mathbf{X}_i; \mathbf{v}), \quad (5)$$

where  $\mathbf{X}_1, \dots, \mathbf{X}_N$  are draws from  $f(\cdot; \mathbf{u})$ . One complication arises in solving (5) when  $\{S(\mathbf{X}) \geq \gamma\}$  is a rare event. Specifically, if the event is sufficiently rare, most of the  $\mathbf{1}(S(\mathbf{X}_i) \geq \gamma)$  terms in (5) are zero and the solution would have a high variance. On realizing that we can instead estimate  $\mathbf{v}_{\text{ce}}^*$ , via importance sampling, as the solution to the maximization program

$$\max_{\mathbf{v}} \frac{1}{N} \sum_{i=1}^N \mathbf{1}(S(\mathbf{X}_i) \geq \gamma) \frac{f(\mathbf{X}_i; \mathbf{u})}{f(\mathbf{X}_i; \mathbf{w})} \log f(\mathbf{X}_i; \mathbf{v}), \quad (6)$$

where  $\mathbf{X}_1, \dots, \mathbf{X}_N$  are draws from some arbitrary density  $f(\cdot; \mathbf{w})$  that dominates  $f(\cdot; \mathbf{u})$ , we obtain the following multi-level CE procedure:

**Algorithm 1. Multi-level CE Algorithm for Rare-Event Probability Estimation**

1. Define  $\hat{\mathbf{v}}_0 = \mathbf{u}$ . Let  $N^e = \lfloor \rho N \rfloor$ , where  $\lfloor \cdot \rfloor$  denotes the integer part. Set  $t = 1$ .
2. Generate a random sample  $\mathbf{X}_1, \dots, \mathbf{X}_N$  from the density  $f(\cdot; \hat{\mathbf{v}}_{t-1})$ . Calculate the performances  $S(\mathbf{X}_i)$  for  $i = 1, \dots, N$ , and order them from smallest to largest,  $S_{(1)}, \dots, S_{(N)}$ . Let  $\hat{\gamma}_t$  be the sample  $(1 - \rho)$ -quantile of performances; that is,  $\hat{\gamma}_t = S_{(N-N^e)}$ . If  $\hat{\gamma}_t > \gamma$ , reset  $\hat{\gamma}_t$  to  $\gamma$ .
3. Use the **same** sample  $\mathbf{X}_1, \dots, \mathbf{X}_N$  to solve the stochastic program (6), with  $\mathbf{w} = \hat{\mathbf{v}}_{t-1}$ . Denote the solution by  $\hat{\mathbf{v}}_t$ .
4. If  $\hat{\gamma}_t < \gamma$ , set  $t = t + 1$  and reiterate from Step 2; otherwise, proceed with Step 5.
5. Let  $T$  be the final iteration counter. Generate a sample  $\mathbf{X}_1, \dots, \mathbf{X}_M$  from the density  $f(\cdot; \hat{\mathbf{v}}_T)$  and estimate  $\ell$  via importance sampling, as in (3).

### 3 Improved CE Method

The well-known degeneracy problem notwithstanding, there is always an importance density that gives a zero variance estimator— $g^*$  the conditional density given the rare event. Therefore, intuitively, if the importance density  $g$  is chosen “close enough” to  $g^*$ , the resulting importance sampling estimator should have reasonable accuracy. A natural question is: what goes wrong with the multi-level CE algorithm in high-dimensional settings? A closer look at Algorithm 1 reveals two possibilities: first, the parametric family within which the optimal CE importance density  $g$  is obtained might not be large enough. As a result, even though  $g$  is the “closest” to  $g^*$  within its parametric family, it still does not behave sufficiently like  $g^*$ . Second, it might be the case that the importance density  $g$  located via the multi-level procedure is suboptimal: the parameter vector  $\widehat{\mathbf{v}}_T$  obtained from the multi-level CE procedure is not a good estimator for  $\mathbf{v}_{\text{ce}}^*$  in some settings. We investigate the latter possibility in this article, and propose a natural remedy to the problem. We also discuss various criteria in choosing a good parametric family  $\mathcal{F}$  in the next section.

Heuristically, the parameter vector  $\mathbf{v}_{\text{ce}}^*$  should give the best estimator according to the cross-entropy criterion, as the density  $f(\cdot; \mathbf{v}_{\text{ce}}^*)$  is the “closest” to  $g^*$ , and it should be used when it is available analytically. However, since the deterministic problem (4) is often intractable, we need to estimate  $\mathbf{v}_{\text{ce}}^*$  via Monte Carlo methods in those cases. In many models with moderate dimension,  $\widehat{\mathbf{v}}_T$  is close enough to  $\mathbf{v}_{\text{ce}}^*$ , and the corresponding importance sampling estimator is reasonably accurate. However, in some settings, particularly when the dimension of the problem is large, the likelihood ratio involved in obtaining  $\widehat{\mathbf{v}}_T$  becomes unstable. Therefore, instead of solving (6) sequentially to obtain  $\widehat{\mathbf{v}}_T$ , we consider an alternative estimator, which does not involve any likelihood ratio and can be obtained in one step.

Recall that the reason why solving (5) directly is difficult is that if we generate draws from the nominal density  $f(\cdot; \mathbf{u})$ , most of the  $\mathbf{1}(S(\mathbf{X}_i) \geq \gamma)$  terms are zero if  $\{S(\mathbf{X}) \geq \gamma\}$  is a rare event. Consequently, the estimator  $\widehat{\mathbf{v}}_{\text{ce}}^*$  obtained from (5) would have a high variance. With this in mind, we consider the following small but significant modification: instead of drawing from  $f(\cdot; \mathbf{u})$ , we can generate a random sample  $\mathbf{X}_1, \dots, \mathbf{X}_N$  from  $g^*(\cdot) = \ell^{-1} f(\cdot; \mathbf{u}) \mathbf{1}(S(\cdot) \geq \gamma)$ , and it is easy to see that  $\widehat{\mathbf{v}}_{\text{ce}}^*$  is exactly the solution to the maximization problem

$$\max_{\mathbf{v}} \frac{1}{N} \sum_{i=1}^N \log f(\mathbf{X}_i; \mathbf{v}). \quad (7)$$

One important point to note is that in contrast to (6), the maximization problem (7) does not involve any indicator function or likelihood ratio. As a result, it does not only afford substantial computational saving in high-dimensional settings, but its solution is more robust and numerically stable as well. Generating draws from  $g^*$ , however, requires additional effort, but with the advent of MCMC methods this problem is well studied and a variety of techniques are available to our disposal. In fact, for all the problems considered in this article, efficient samplers exist to generate from  $g^*$ . In addition, the number of draws required to estimate  $\widehat{\mathbf{v}}_{\text{ce}}^*$  is typically much smaller than that required in the multi-level CE algorithm.

**Algorithm 2. Improved CE Algorithm for Rare-Event Probability Estimation**

1. Generate a random sample  $\mathbf{X}_1, \dots, \mathbf{X}_N$  from the density  $g^*(\mathbf{x})$  and find the solution to (7), which is denoted as  $\widehat{\mathbf{v}}_{\text{ce}}^*$ .

2. Generate a sample  $\mathbf{X}_1, \dots, \mathbf{X}_M$  from the density  $f(\cdot; \widehat{\mathbf{v}}_{\text{ce}}^*)$  and estimate  $\ell$  via importance sampling, as in (3).

As mentioned in the introduction, there is an important literature on estimating the normalizing constant of an arbitrary density by MCMC methods. Since the rare-event probability  $\ell$  can be written as a normalizing constant of the zero-variance importance density  $g^*$ , in principle  $\ell$  may be estimated by these methods. However, all these methods involve using MCMC draws to compute certain Monte Carlo averages, which are then used to give an estimate of  $\ell$ . The major drawback of this approach is that MCMC draws are typically costly to obtain, especially in high-dimensional problems. In fact, in complex models where the MCMC draws exhibit high autocorrelation, the computational effort required to obtain enough draws for a sufficiently accurate estimate might be formidable. Therefore, these methods are inherently not suitable for rare-event simulation, where precise estimates are often needed. In contrast, the proposed method is an adaptive importance sampling approach, and it circumvents these drawbacks by generating independent draws from some convenient density. Of course the proposed approach also requires MCMC draws for obtaining the optimal importance density, but the number of draws needed is typically small (a few hundreds to a thousand draws for obtaining the importance density versus tens of thousands draws for the main importance sampling run).

### 3.1 A Toy Example

To investigate the quality of the optimal CE parameter estimators for the multi-level CE and the proposed method, we consider a toy example where we can analytically compute  $\mathbf{v}_{\text{ce}}^*$  by solving independently the deterministic problem (4). Specifically, let  $X_i \sim \text{Ber}(p_i)$  for  $i = 1, \dots, n$ . We wish to estimate  $\mathbb{P}(S_n(\mathbf{X}) \geq \gamma)$ , where  $S_n(\mathbf{X}) = X_1 + \dots + X_n$ ,  $\gamma = 0.6n$ , and  $\mathbf{X} = (X_1, \dots, X_n)$ . The nominal density is

$$f(\mathbf{x}; \mathbf{p}) = \prod_{i=1}^n p_i^{x_i} (1 - p_i)^{(1-x_i)},$$

where  $\mathbf{x} = (x_1, \dots, x_n)$  and  $\mathbf{p} = (p_1, \dots, p_n)$ . It is natural to locate the optimal CE importance density within the parametric family  $f(\mathbf{x}; \mathbf{q})$  indexed by  $\mathbf{q} = (q_1, \dots, q_n)$ , where  $q_i \in (0, 1)$  for  $i = 1, \dots, n$ . Therefore, the deterministic problem (4) becomes

$$\mathbf{q}^* = \underset{\mathbf{q}}{\operatorname{argmax}} \sum_{\mathbf{x}: S_n(\mathbf{x}) \geq \gamma} \left( \prod_{i=1}^n p_i^{x_i} (1 - p_i)^{(1-x_i)} \right) \left( \sum_{i=1}^n x_i \log q_i + (1 - x_i) \log(1 - q_i) \right).$$

It can be shown that the solution  $\mathbf{q}^* = (q_1^*, \dots, q_n^*)$  admits a closed-form expression, with

$$q_j^* = \frac{\sum_{\mathbf{x}: S_n(\mathbf{x}) \geq \gamma} x_j \prod_{i=1}^n p_i^{x_i} (1 - p_i)^{(1-x_i)}}{\sum_{\mathbf{x}: S_n(\mathbf{x}) \geq \gamma} \prod_{i=1}^n p_i^{x_i} (1 - p_i)^{(1-x_i)}},$$

for  $j = 1, \dots, n$ . As a numerical example, we first set  $n = 50$ ,  $\gamma = 30$  and  $p_1 = \dots = p_n = 0.1$ . The rare-event probability can be computed analytically and it is found to be  $\ell = 6.21 \times 10^{-18}$ . We estimate  $\mathbf{q}^*$  via the multi-level CE procedure and the proposed method. For the CE method,

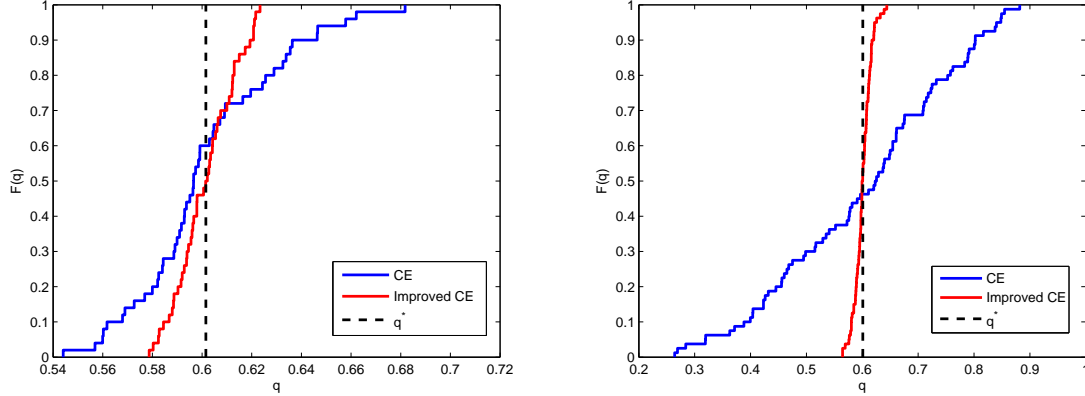


Figure 1: Empirical cdfs of the CE and improved CE estimates for the toy example with  $n = 50$  (left) and  $n = 80$  (right).

we implement Algorithm 1 with  $N = 10000$  and  $\rho = 0.01$ . More specifically, we initialize with  $\hat{\mathbf{q}}_0 = \mathbf{p}$ . At the  $t$ -th iteration, we first sample  $\mathbf{X}_1, \dots, \mathbf{X}_N$  from  $f(\mathbf{x}; \hat{\mathbf{q}}_{t-1})$ , where  $\mathbf{X}_i = (X_{i1}, \dots, X_{in})$  and  $\hat{\mathbf{q}}_{t-1} = (\hat{q}_{t-1,1}, \dots, \hat{q}_{t-1,n})$ . Then we use (6) to compute  $\hat{\mathbf{q}}_t$ , which involves  $n$  one-dimensional optimizations. In fact, after some algebra, it can be shown that  $\hat{q}_{tj}$ , the  $j$ -th component of  $\hat{\mathbf{q}}_t$ , is given by

$$\hat{q}_{tj} = \frac{\sum_{i=1}^N W_i X_{ij}}{\sum_{i=1}^N W_i},$$

where  $W_i = \mathbf{1}(S(\mathbf{X}_i) \geq \hat{\gamma}_t) f(\mathbf{X}_i; \mathbf{p}) / f(\mathbf{X}_i; \hat{\mathbf{q}}_{t-1})$ . The algorithm terminates at the 4-th iteration, requiring a total of 40000 draws. For the proposed method, we run a Gibbs sampler to sample from  $g^*$  with 10 parallel chains, each has a length of 1000, and the total budget is therefore 10000. Given the draws  $\mathbf{X}_1, \dots, \mathbf{X}_N$  from  $g^*$ , the solution for (7) is simply  $\hat{\mathbf{q}}_{ce}^* = N^{-1} \sum_{i=1}^N \mathbf{X}_i$ . It is also worth mentioning that drawing from  $g^*$  via the Gibbs sampler in this case only requires generating Bernoulli draws. The empirical cumulative distribution functions (cdf) of the CE and improved CE estimates, together with the optimal CE parameter calculated analytically, are presented in Figure 1 (left panel).

For this relatively low-dimensional problem with only 50 parameters, the optimal CE parameters estimated by both methods are reasonably close to those obtained analytically. For instance, the optimal CE parameter calculated analytically is about 0.6 and most of the CE estimates are concentrated between 0.56 and 0.66. However, it is evident that the CE estimates fluctuate more widely compared to those obtained by the improved version, even though the simulation budget for the former is four times as large. Since the CE estimates are not as accurate as the proposed method, it is not surprising that the variance of the resulting estimator from the multi-level CE procedure is about 20% larger. We next perform the same experiment with  $n = 80$  and  $\gamma = 48$ , and the rare-event probability is  $\ell = 8.11 \times 10^{-28}$ . We report the corresponding cdfs of the CE and improved CE estimates in Figure 1 (right). As is apparent in the figure, as the dimension of the problem gets larger, the CE estimates become more unreliable, while those from the proposed method are essentially unaffected by the increase in dimension. In terms of the quality of the importance sampling estimators, the variance of the multi-level CE estimator is more than 100 times larger compared to the improved CE estimator.

The result from this toy example suggests a reason why the multi-level CE method fails to give accurate estimates in high-dimensional settings: the parameter vector obtained is suboptimal, and therefore the resulting importance density does not sufficiently mimic the behavior of  $g^*$ . In principle one can increase the accuracy of the multi-level CE estimates by increasing the sample size  $N$  or the rarity parameter  $\rho$ . In either case, however, the total simulation effort would increase, and in moderately high-dimensional problems, this approach might not be practical. On the other hand, the result also suggests that if we avoid the multi-level maximization procedure and estimate  $\mathbf{v}_{\text{ce}}^*$  directly via (7), we can improve the performance of the standard CE procedure. In Section 5, we will demonstrate the proposed method by visiting a credit risk model that involves hundreds or thousands of random variables. We show that even in this high-dimensional problem, the improved CE method works well and gives estimators that compare favorably to existing importance sampling estimators.

## 4 VM Method and the Choice of Parametric Family

It is of interest to note that the above approach also provides a practical way to locate the minimum variance importance sampling estimator within the parametric family  $\mathcal{F}$ . That is, instead of using  $f(\mathbf{x}; \hat{\mathbf{v}}_{\text{ce}}^*)$  that minimizes the cross-entropy distance to  $g^*$ , we can choose  $f(\mathbf{x}; \mathbf{v}) \in \mathcal{F}$  such that the variance, or equivalently the second moment, of the associated importance sampling estimator is minimized. The minimizer

$$\begin{aligned} \mathbf{v}_{\text{vm}}^* &= \underset{\mathbf{v}}{\operatorname{argmin}} \int f(\mathbf{x}; \mathbf{v}) \mathbf{1}(S(\mathbf{x}) \geq \gamma) \frac{f(\mathbf{x}; \mathbf{u})^2}{f(\mathbf{x}; \mathbf{v})^2} d\mathbf{x} \\ &= \underset{\mathbf{v}}{\operatorname{argmin}} \int \mathbf{1}(S(\mathbf{x}) \geq \gamma) \frac{f(\mathbf{x}; \mathbf{u})^2}{f(\mathbf{x}; \mathbf{v})} d\mathbf{x} \end{aligned} \quad (8)$$

is referred to as the *optimal variance minimization (VM) parameter vector*. Again, the deterministic minimization problem (8) is difficult to solve, and analytic solutions are available only for certain specific examples (e.g. Chan et al., 2011). However, one can estimate  $\mathbf{v}_{\text{vm}}^*$  by first sampling  $\mathbf{X}_1, \dots, \mathbf{X}_N$  from  $g^*$  and then solving

$$\min_{\mathbf{v}} \frac{1}{N} \sum_{i=1}^N \frac{f(\mathbf{X}_i; \mathbf{u})}{f(\mathbf{X}_i; \mathbf{v})}. \quad (9)$$

This leads to the following algorithm.

### Algorithm 3. VM Algorithm for Rare-Event Probability Estimation

1. Generate a random sample  $\mathbf{X}_1, \dots, \mathbf{X}_N$  from the density  $g^*(\mathbf{x})$  and find the solution to (9), which is denoted as  $\hat{\mathbf{v}}_{\text{vm}}^*$ .
2. Generate a sample  $\mathbf{X}_1, \dots, \mathbf{X}_M$  from the density  $f(\cdot; \hat{\mathbf{v}}_{\text{vm}}^*)$  and estimate  $\ell$  via importance sampling, as in (3).

By construction the VM method gives the minimum variance importance sampling estimator within the parametric family  $\mathcal{F}$ . Therefore, if  $\mathcal{F}$  is chosen to contain the nominal density  $f(\cdot; \mathbf{u})$ ,



then the variance of the VM estimator is no larger than that of the crude Monte Carlo, which is not true for importance sampling estimators in general. For the same reason, the VM estimator has a smaller variance than the CE one. Thus, in situations where the VM optimal parameter vector  $\hat{\mathbf{v}}_{\text{vm}}^*$  can be easily obtained, the VM method is preferable. On the other hand, it is typically much easier to solve (7) to find  $\hat{\mathbf{v}}_{\text{ce}}^*$  than to solve the VM problem in (9) to obtain  $\hat{\mathbf{v}}_{\text{vm}}^*$ . This is because the former problem is “separable” in the sense that one can often reduce the potentially high-dimensional problem in (7) into several low-dimensional ones by, e.g., choosing the parametric family in the form  $\mathcal{F} = \{f(\mathbf{x}; \mathbf{v}) = \prod_{i=1}^k f(\mathbf{x}_i; \mathbf{v}_i)\}$  where  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_k)$  and  $\mathbf{v} = (\mathbf{v}_1, \dots, \mathbf{v}_k)$ . In contrast, the VM problem is not separable, and often involves high-dimensional optimization in typical applications. Moreover, in many settings the VM and CE methods prescribe very similar importance sampling densities (de Boer et al., 2004; Chan et al., 2011). Hence, in applications where it is difficult to obtain  $\hat{\mathbf{v}}_{\text{vm}}^*$ , one might want to consider the CE method instead.

We now discuss various considerations for choosing  $\mathcal{F}$ . It is obvious that if we choose the parametric family to be the singleton  $\mathcal{F} = \{f(\mathbf{x}; \mathbf{u})\}$ , then both the VM and CE estimators reduce to the crude Monte Carlo estimator, and no variance reduction is achieved. On the other extreme, if  $\mathcal{F}$  is chosen to include all pdfs, then both  $f(\mathbf{x}; \mathbf{v}_{\text{vm}}^*)$  and  $f(\mathbf{x}; \mathbf{v}_{\text{ce}}^*)$  become  $g^*$ , and neither estimators can be used in practice. Hence, the fundamental trade-off is between system complexity (in obtaining the VM/CE optimal parameter vectors, generating random samples from the importance density and evaluating the likelihood ratio) and matching  $g^*$  sufficiently well. The objective is to select a parametric family such that (1) the family is sufficiently large/diverse, and (2) each member in  $\mathcal{F}$  is easy to sample from and easy to evaluate. One “default” choice that is often easy to implement is to select  $\mathcal{F}$  to be the same family of pdfs as the nominal density  $f(\mathbf{x}; \mathbf{u})$ . For example, if  $f(\mathbf{x}; \mathbf{u})$  is a product of  $k$   $\text{Exp}(u_i)$  densities, then  $\mathcal{F}$  can be chosen to be a product of  $k$   $\text{Exp}(v_i)$  densities indexed by  $\mathbf{v} = (v_1, \dots, v_k)$ . This simple strategy is found to be adequate in typical applications. An additional advantage of selecting the product form  $\mathcal{F} = \{f(\mathbf{x}; \mathbf{v}) = \prod_{i=1}^k f(\mathbf{x}_i; \mathbf{v}_i)\}$ , as discussed earlier, is that the CE optimization problem (7) becomes separable, and the potentially high-dimensional problem is reduced into several low-dimensional ones. For more discussions on various strategies and trade-offs in approximating  $g^*$ , we refer the readers to the review in Smith et al. (1997).

Other more sophisticated strategies can be implemented if one has an asymptotic description of how the rare event  $\{S(\mathbf{X}) \geq \gamma\}$  occurs. For instance, Orsak (1993) shows that a necessary condition to achieve variance reduction is that the rare event  $\{S(\mathbf{X}) \geq \gamma\}$  occurs more frequently under the importance density than that obtained using the nominal density. This suggests that special attention should be paid to random variables that determine the occurrence of the rare event  $\{S(\mathbf{X}) \geq \gamma\}$ . For instance, in applications where the rare event  $\{S(\mathbf{X}) \geq \gamma\}$  occurs primarily when one or a few random variables attain certain values, one could change only the distributions of those variables, while keeping the distributions of other variables fixed. In this way, the dimension of the optimization problems in (7) and (9) is reduced, while the quality of the estimators are not substantially affected.

## 5 Application: Large Portfolio Loss in the $t$ Copula Model

We illustrate the utility of the proposed approach by estimating an important measure of risk—the probability of large portfolio losses—under the recently proposed  $t$  copula model of Bassamboo et al. (2008). Suppose we have a portfolio of loans consisting of  $n$  obligors, each of them has a given probability of defaulting, which we denote as  $p_i \in (0, 1)$ ,  $i = 1, \dots, n$ . Introduce a vector of underlying latent variables  $\mathbf{X} = (X_1, \dots, X_n)$  such that the  $i$ -th obligor defaults if  $X_i$  exceeds some given threshold level  $x_i$ , i.e.,  $p_i = \mathbb{P}(X_i > x_i)$ . We define the portfolio loss incurred from defaults as

$$L(\mathbf{X}) = c_1 \mathbf{1}(X_1 > x_1) + \dots + c_n \mathbf{1}(X_n > x_n),$$

where  $c_i$  is the monetary loss associated with the default of the  $i$ -th obligor. A natural risk measure of the portfolio is the probability of large losses of the form

$$\ell(\gamma) = \mathbb{P}(L(\mathbf{X}) > \gamma), \quad (10)$$

where  $\gamma = bn$  for some  $b > 0$ . To complete the model specifications, one needs to specify the joint distribution of  $\mathbf{X}$ . One popular model that is widely used in the financial industry is the *normal copula* model that forms the basis of the CreditMetrics and other related models. Specifically, the underlying correlations are specified through a linear factor model:  $X_i = w_{i1}Z_1 + \dots + w_{im}Z_m + w_i\eta_i$ ,  $i = 1, \dots, n$ , where  $Z_1, \dots, Z_m$  are iid standard normal variables known as *factors* and  $\eta_i$  is a normal random variable independent of the factors that captures the idiosyncratic risk of the  $i$ -th obligor. In addition, we assume (without loss of generality) that  $w_{i1}^2 + \dots + w_{im}^2 + w_i^2 = 1$ .

One of the potential problems of the normal copula model is that it might assign too low a probability to the event of many simultaneous defaults. In view of this inadequacy of the normal copula, Bassamboo et al. (2008) propose the *t-copula model*, based on the multivariate  $t$ -distribution, that attempts to capture the relatively frequent occurrences of extremal comovements of financial variables. Following Bassamboo et al. (2008) we restrict our attention to the single factor model ( $m=1$ ) to keep the notations simple. It is important to realize that the techniques developed here can be easily generalized to a general  $m$ -factor model. As in the normal copula model, the factors and the individuals' idiosyncratic risks are modeled as independent normally distributed random variables. More precisely,  $Z \sim \mathbf{N}(0, 1)$  and  $\eta_i \stackrel{iid}{\sim} \mathbf{N}(0, \sigma_\eta^2)$ ,  $i = 1, \dots, n$ . To induce a  $t$  structure, we introduce a *shock variable*  $\lambda > 0$  that is independent of  $Z$  and  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)$  such that  $\lambda \sim \text{Gamma}(\nu/2, \nu/2)$  for some  $\nu > 0$ . Define

$$X_i = \left( \rho Z + \sqrt{1 - \rho^2} \eta_i \right) \lambda^{-\frac{1}{2}}, \quad i = 1, \dots, n. \quad (11)$$

It is well-known that if  $\lambda \sim \text{Gamma}(\nu/2, \nu/2)$ , then *marginally*  $\mathbf{X} = (X_1, \dots, X_n)$  follows a multivariate  $t$  distribution with degree of freedom  $\nu$ . Bassamboo et al. (2008) propose two importance sampling algorithms to estimate the probability that the portfolio incurs large losses. The first estimator uses importance sampling based on an *exponential change of measure* (ECM) (see, e.g., Asmussen and Glynn, 2007) and has bounded relative error; the second uses a variant of *hazard rate twisting* (HRT) (Juneja and Shahabuddin, 2002), which is shown to be logarithmically efficient. An extensive simulation study shows that while both estimators offer substantial variance reduction, the former provides 6 to 10 times higher variance reduction than the latter.

Nevertheless, the more efficient ECM algorithm involves generating random variables from a nonstandard distribution, which takes on average three times more time compared to naive Monte Carlo simulation. In addition, the normalizing constant of the proposal density is not known, and has to be computed by numerical routines in order to be used in the likelihood ratio evaluation.

We now apply the proposed methods to estimate the probability of large portfolio loss in (10). First, we obtain a sample from the zero-variance importance density  $g^*$  via the Gibbs sampler. Second, given the draws, we locate the optimal CE and VM importance densities within an appropriate family of distributions. To this end, let  $\mathring{f}(z, \boldsymbol{\eta}, \lambda)$  denote the joint density of  $(z, \boldsymbol{\eta}, \lambda)$ , i.e.,

$$\mathring{f}(z, \boldsymbol{\eta}, \lambda) = f_{\mathbf{N}}(z; 0, 1) f_{\mathbf{G}}(\lambda; \nu/2, \nu/2) \prod_{i=1}^n f_{\mathbf{N}}(\eta_i; 0, \sigma_{\boldsymbol{\eta}}^2),$$

where  $f_{\mathbf{N}}(\cdot; a, b)$  denotes the density of  $\mathbf{N}(a, b)$  and  $f_{\mathbf{G}}(\cdot; c, d)$  represents the density of  $\mathbf{Gamma}(c, d)$ . Note that the zero-variance importance density is

$$g^*(z, \boldsymbol{\eta}, \lambda) = \mathring{f}(z, \boldsymbol{\eta}, \lambda | L(\mathbf{x}) > \gamma) \propto \mathring{f}(z, \boldsymbol{\eta}, \lambda) \mathbf{1}(L(\mathbf{x}) > \gamma),$$

where  $\mathbf{x}$  is defined in (11). A Gibbs sampler can be constructed by sequentially drawing from  $g^*(z | \boldsymbol{\eta}, \lambda)$ ,  $g^*(\lambda | z, \boldsymbol{\eta})$  and  $g^*(\boldsymbol{\eta} | z, \lambda)$ . Two points on implementation are worth mentioning. First, the Gibbs sampler involves only drawing from univariate truncated normal and right truncated gamma distributions, and a draw from either distribution can be obtained by the inverse-transform method or various efficient rejection methods (e.g., Robert, 1995; Philippe, 1997). Second, since the performance of the proposed estimator is relatively insensitive to the autocorrelation of the MCMC draws, even though more efficient sampling scheme might exist, the gain in efficiency might not worth the extra effort. The detailed implementation of the Gibbs sampler is discussed in the appendix.

Now suppose we have a sample  $\{Z_i, \boldsymbol{\eta}_i, \lambda_i\}_{i=1}^N$  from  $g^*$ . We consider the following family of distributions within which to locate the optimal CE and VM importance densities:

$$\mathcal{F} = \left\{ f(z, \boldsymbol{\eta}, \lambda; \mathbf{v}) = f_{\mathbf{N}}(z; \mu_z, \sigma_z^2) f_{\mathbf{G}}(\lambda; \alpha_\lambda, \beta_\lambda) \prod_{i=1}^n f_{\mathbf{N}}(\eta_i; \mu_{\boldsymbol{\eta}}, \sigma_{\boldsymbol{\eta}}^2) \right\},$$

where the family is indexed by  $\mathbf{v} = (\mu_z, \sigma_z^2, \alpha_\lambda, \beta_\lambda, \mu_{\boldsymbol{\eta}})$  with  $\mu_z, \mu_{\boldsymbol{\eta}} \in \mathbb{R}$  and  $\sigma_z^2, \alpha_\lambda, \beta_\lambda > 0$ . In particular, we have  $\mathring{f}(\cdot) = f(\cdot; \mathbf{u})$  where  $\mathbf{u} = (0, 1, \nu/2, \nu/2, 0)$ . Note that  $\mathcal{F}$  is the “default” choice discussed in Section 4 (i.e., the same family as the nominal density), except that we fix the variance of  $\eta_i$  at  $\sigma_{\boldsymbol{\eta}}^2$ . This is because the rare event  $\{L(\mathbf{X}) > \gamma\}$  is determined primarily by the values of  $\lambda$  and  $Z$  (see Bassamboo et al., 2008, for the precise statements of the asymptotics), while the random variables  $\eta_i$  are relatively unimportant. It is therefore sufficient to allow the mean of  $\eta_i$  to change but not its variance.

Since any member of  $\mathcal{F}$  is a product of densities, standard techniques of obtaining the maximum likelihood estimator (MLE) can be applied to estimate the optimal CE parameter vector  $\mathbf{v}_{\text{ce}}^*$ . In fact, it is easy to solve the maximization problem in (7) analytically for  $(\hat{\mu}_z^*, \hat{\sigma}_z^{2*}, \hat{\mu}_{\boldsymbol{\eta}}^*)$ :

$$\hat{\mu}_z^* = \frac{1}{N} \sum_{i=1}^N Z_i, \quad \hat{\sigma}_z^{2*} = \frac{1}{N} \sum_{i=1}^N (Z_i - \hat{\mu}_z^*)^2, \quad \hat{\mu}_{\boldsymbol{\eta}}^* = \frac{1}{nN} \sum_{i=1}^N \sum_{j=1}^n \eta_{i,j},$$

where  $\eta_{i,j}$  is the  $j$ -th element of  $\boldsymbol{\eta}_i$ . Moreover,  $(\hat{\alpha}_\lambda^*, \hat{\beta}_\lambda^*)$  can be obtained, for example, by the Newton-Raphson method. Alternatively, they can be approximated by the method of moments estimates:  $\tilde{\alpha} = \bar{\mu}_\lambda^2 / S_\lambda^2$  and  $\tilde{\beta} = \bar{\mu}_\lambda / S_\lambda^2$ , where  $\bar{\mu}_\lambda$  and  $S_\lambda^2$  are respectively the sample mean and sample variance of  $\lambda_1, \dots, \lambda_N$ . The latter approach is the one we adopt here. Once we obtain the optimal CE importance density  $f(\cdot; \hat{\mathbf{v}}_{ce}^*)$ , we then deliver the importance sampling estimator:

$$\frac{1}{M} \sum_{i=1}^M \mathbf{1}(L(\mathbf{X}_i) > \gamma) \frac{f(Z_i, \boldsymbol{\eta}_i, \lambda_i; \mathbf{u})}{f(Z_i, \boldsymbol{\eta}_i, \lambda_i; \hat{\mathbf{v}}_{ce}^*)}, \quad (12)$$

where  $(Z_i, \boldsymbol{\eta}_i, \lambda_i)$ ,  $i = 1, \dots, M$  are generated from the importance density  $f(\cdot; \hat{\mathbf{v}}_{ce}^*)$ .

To obtain the VM optimal parameter vector  $\hat{\mathbf{v}}_{vm}^*$ , we use a numerical routine to solve (9). Although this is a high-dimensional application involving hundreds of random variables, the optimization problem (9) involves only five parameters

$\mathbf{v} = (\mu_z, \sigma_z^2, \alpha_\lambda, \beta_\lambda, \mu_\eta)$ . Therefore, finding  $\hat{\mathbf{v}}_{vm}^*$  numerically in this example is still feasible. While  $\hat{\mathbf{v}}_{ce}^*$  can be obtained instantly, it takes the optimization routine about 2 seconds on a Dual-core 2.6 GHz desktop to find  $\hat{\mathbf{v}}_{vm}^*$ . Once we have  $\hat{\mathbf{v}}_{vm}^*$ , we deliver the estimator as in (12) with the importance density  $f(Z_i, \boldsymbol{\eta}_i, \lambda_i; \hat{\mathbf{v}}_{vm}^*)$  instead.

## 5.1 Numerical Results

We demonstrate the performance of the proposed importance sampling estimators via simulation studies similar to those in Bassamboo et al. (2008). The broad conclusions drawn from these experiments are that even though the  $t$  copula model involves hundreds of random variables, the proposed estimators perform remarkably well and offer accurate estimates for a relatively small replication sample size ( $M = 50000$ ). In addition, both compare favorably to the two other importance sampling estimators, ECM and HRT, proposed in Bassamboo et al. (2008). Except in one scenario, they outperform the ECM algorithm, offering up to 8 times higher variance reduction, and are more efficient than the HRT algorithm in all scenarios, providing 2 to 16 times higher variance reduction. Another factor that is in favor of the proposed estimators is that they only involve generating from standard distributions. In contrast, the ECM estimator involves generating from a nonstandard distribution, where the normalizing constant is not known, and has to be computed by numerical routines. In addition, it involves accept-reject sampling, which takes on average three times longer than naive simulation, thus making the algorithm slower and more difficult to implement.

For comparison purposes, we consider the same sets of parameter values as those in Bassamboo et al. (2008) Tables 1–4. In all the experiments in this subsection we set  $\sigma_\eta^2 = 9$  and  $l = b \times n$ . The individual threshold  $x_i$  and monetary loss associated to the  $i$ -th obligor  $c_i$  are the same across the obligors:  $x_1 = \dots = x_n = \sqrt{n} \times 0.5$ , and  $c_1 = \dots = c_n = 1$ . For each set of specified parameters, we run 5 parallel chains via the Gibbs sampler described in the appendix. Each chain is of length 1000 and we discard the first 50 draws in each chain as “burn-in”. We use the Gibbs output to estimate the optimal CE and VM parameters. Then we generate  $M = 50000$  samples for the main run. Table 1 shows the relative errors (in %) of the proposed estimators, as well those of the ECM and HRT, for various values of the degree of freedom parameter  $\nu$ . The estimated probability  $\hat{\ell}(\gamma)$  is obtained by the CE estimator. Other model parameters are chosen to be  $n = 250$ ,  $\rho = 0.25$  and  $b = 0.25$ . In Table 2 we perform the same comparison but

now we vary the correlation parameter  $\rho$  while keeping  $\nu$  fixed at 12.

Table 1: Relative errors (in %) of the improved CE and VM estimators for various values of  $\nu$ .

$\nu$	$\widehat{\ell}(\gamma)$	CE	VM	ECM	HRT
4	$8.14 \times 10^{-3}$	0.5	0.5	0.6	1.1
8	$2.41 \times 10^{-4}$	0.8	0.7	0.9	1.8
12	$1.08 \times 10^{-5}$	1.1	1.0	1.7	2.6
16	$6.08 \times 10^{-7}$	1.4	1.3	2.8	3.6
20	$4.43 \times 10^{-8}$	1.8	1.7	3.7	5.4

Table 2: Relative errors (in %) of the improved CE and VM estimators for various values of  $\rho$ .

$\rho$	$\widehat{\ell}(\gamma)$	CE	VM	ECM	HRT
0.1	$8.52 \times 10^{-6}$	1.1	1.0	0.9	1.8
0.2	$9.77 \times 10^{-6}$	1.2	1.0	1.2	2.3
0.3	$1.17 \times 10^{-5}$	1.1	1.0	1.7	3.2
0.4	$1.37 \times 10^{-5}$	1.1	1.0	3.1	4.0

In Table 3 we report the relative errors (in %) of the proposed estimator as well as those of ECM and HRT for various values of  $n$ , the number of obligors. Other model parameters are chosen to be  $\nu = 12$ ,  $\rho = 0.25$  and  $b = 0.25$ . Table 4 shows the results of a similar analysis but now we vary  $b$ , the proportion of defaults in the portfolio, while keeping  $n$  fixed at 250. The results suggest that the improved CE estimator performs remarkably well even when  $n$  is large, where the model contains hundreds of random variables. It is of interest to note that even though the VM estimator performs better than the CE one in all scenario as expected, their performances are remarkably similar. Also note that in Bassamboo et al. (2008) Tables 3–4, the authors actually computed  $\mathbb{P}(L(\mathbf{X}) \geq \gamma)$  instead of  $\mathbb{P}(L(\mathbf{X}) > \gamma)$  as stated. As a result, the estimated rare-event probabilities there are slightly larger than those we report in the corresponding tables.

Table 3: Relative errors (in %) of the improved CE and VM estimators for various values of  $n$ .

$n$	$\widehat{\ell}(\gamma)$	CE	VM	ECM	HRT
100	$1.86 \times 10^{-3}$	1.3	1.1	1.6	1.8
250	$1.08 \times 10^{-5}$	1.1	1.0	1.7	2.6
500	$1.47 \times 10^{-7}$	1.0	0.9	1.5	3.4
1000	$2.28 \times 10^{-9}$	0.9	0.8	1.6	3.6

Table 4: Relative errors (in %) of the improved CE and VM estimators for various values of  $b$ .

$b$	$\widehat{\ell}(\gamma)$	CE	VM	ECM	HRT
0.1	$3.47 \times 10^{-3}$	0.8	0.7	0.9	1.6
0.2	$7.44 \times 10^{-5}$	1.0	0.9	1.2	2.5
0.3	$1.12 \times 10^{-6}$	1.4	1.2	2.0	3.4

## 6 Concluding Remarks and Future Research

In this article we first document the main reason why the standard CE method fails in certain high-dimensional settings: the importance density obtained from the multi-level procedure is suboptimal. We therefore introduce a small but significant modification to the standard CE method, and demonstrate that it gives substantial improvement over the traditional approach. We then apply the proposed method to a high-dimensional estimation problem under a recently proposed credit risk model, and show that it outperforms existing importance sampling estimators.

The proposed approach gives the best importance density within the class of densities considered, and therefore in a sense it is the optimal importance sampling strategy. Many, if not all, of the problems previously considered with the multi-level CE and VM approach, particularly those mentioned in the introduction, can be tackled by the improved variant, which is expected to give better results. Moreover, its applicability is not limited to rare-event simulation, but it can be applied to a wide variety of problems, ranging from pricing exotic options to estimating normalizing constants of an arbitrary density, particularly the marginal likelihood in Bayesian statistics.

## Appendix: Gibbs Sampler for the $t$ copula Model

In this appendix we discuss the implementation of the Gibbs sampler of drawing from  $g^*(z, \boldsymbol{\eta}, \lambda)$ , the zero-variance importance density for estimating the rare-event probability  $\mathbb{P}(L(\mathbf{X}) > \gamma)$  under the  $t$  copula model. The Gibbs sampler is constructed by sequentially drawing from  $g^*(z | \boldsymbol{\eta}, \lambda)$ ,  $g^*(\lambda | z, \boldsymbol{\eta})$  and  $g^*(\boldsymbol{\eta} | z, \lambda)$ . The first conditional density  $g^*(z | \boldsymbol{\eta}, \lambda)$  is a univariate truncated normal. To see this, first define  $G_i = \rho^{-1}(x_i \lambda^{1/2} - \sqrt{1 - \rho^2} \eta_i)$ . Arrange  $G_1, \dots, G_n$  in ascending order, let  $G_{(i)}$  denote the  $i$ -th ordered value, and  $c_{(i)}$  the corresponding ordered monetary loss. Then the event  $\{L(\mathbf{X}) > \gamma\}$  occurs if and only if  $Z > G_{(k)}$  where  $k = \min\{l : \gamma < \sum_{i=1}^l c_{(i)}\}$ . In particular, if  $c_i = c$  for all  $i = 1, \dots, n$ , then  $k = \lfloor \gamma/c \rfloor + 1$ , where  $\lfloor \cdot \rfloor$  indicates the integer part. Hence, the conditional density of  $Z$  is a univariate truncated normal distribution:

$$g^*(z | \boldsymbol{\eta}, \lambda) \propto f_{\mathbf{N}}(z; 0, 1) \mathbf{1}(z > G_{(k)}),$$

and a draw from this distribution can be obtained either by the inverse-transform method or various efficient rejection methods (e.g., Robert, 1995). We use the inverse-transform method to generate draws from the truncated normal distribution.

Next define  $H_i = (\rho Z + \sqrt{1 - \rho^2} \eta_i) x_i^{-1}$  and let  $H_{(i)}$  be the  $i$ -th ordered value of  $H_1, \dots, H_n$  and  $c_{(i)}$  the corresponding ordered monetary loss. Since the event  $\{L(\mathbf{X}) > \gamma\}$  occurs if and only if  $\sqrt{\lambda} < H_{(n-k)}$  where  $k = \min\{l : \gamma < \sum_{i=1}^l c_{(i)}\}$ , the conditional density  $g^*(\lambda | z, \boldsymbol{\eta})$  is a right-truncated gamma distribution:

$$g^*(\lambda | z, \boldsymbol{\eta}) \propto f_G(\lambda; \nu/2, \nu/2) \mathbf{1}(\lambda < \min(H_{(n-k)}^2, 0)),$$

and a draw from this distribution can be obtained either by the inverse-transform method or the rejection method described in Philippe (1997). We adopt the latter approach to generate draws from the right-truncated gamma distribution.

Lastly, we need to obtain a draw from  $g^*(\boldsymbol{\eta} | z, \lambda)$ , which is a truncated multivariate normal distribution. A feasible approach is to sequentially draw from  $g^*(\eta_i | z, \lambda, \boldsymbol{\eta}_{-i})$  for  $i = 1, \dots, n$ , each of which is a univariate truncated normal density, where  $\boldsymbol{\eta}_{-i}$  denotes the vector  $\boldsymbol{\eta}$  except the  $i$ -th element, i.e.,  $\boldsymbol{\eta}_{-i} = (\eta_1, \dots, \eta_{i-1}, \eta_{i+1}, \dots, \eta_n)$ . More specifically, given  $(\boldsymbol{\eta}_{-i}, Z, \lambda)$ , if

$$\sum_{j \neq i} c_j \mathbf{1}((\rho Z + \sqrt{1 - \rho^2} \eta_j) \lambda^{-1/2}) > \gamma,$$

then there is no restriction on  $\eta_i$  and

$$g^*(\eta_i | Z, \lambda, \boldsymbol{\eta}_{-i}) = f_N(\eta_i; 0, \sigma_{\boldsymbol{\eta}}^2);$$

otherwise,

$$g^*(\eta_i | Z, \lambda, \boldsymbol{\eta}_{-i}) = f_N(\eta_i; 0, \sigma_{\boldsymbol{\eta}}^2) \mathbf{1}\left(\eta_i > \frac{x_i \lambda^{1/2} - \rho Z}{\sqrt{1 - \rho^2}}\right).$$

Alternatively, one can simply generate  $\eta_i^c \stackrel{iid}{\sim} \mathbf{N}(0, \sigma_{\boldsymbol{\eta}}^2)$ , for  $i = 1, \dots, n$ , and compute the corresponding  $L(\mathbf{X})$ . If  $L(\mathbf{X}) > \gamma$ , set  $\boldsymbol{\eta} = \boldsymbol{\eta}^c$ ; otherwise, repeat the process until a draw is accepted. We adopt the latter approach. In the numerical examples, the acceptance rate is over 0.8.

## References

- S. Asmussen and P. W. Glynn. *Stochastic simulation: algorithms and analysis*. Springer, New York, 2007.
- S. Asmussen, R. Y. Rubinstein, and D. P. Kroese. Heavy tails, importance sampling and cross-entropy. *Stochastic Models*, 21:57–76, 2005.
- A. Bassamboo, S. Juneja, and A. Zeevi. Portfolio credit risk with extremal dependence: Asymptotic analysis and efficient simulation. *Operations Research*, 56(3):593–606, 2008.
- J. C. C. Chan and E. Eisenstat. Marginal likelihood estimation with the cross-entropy method. 2011. Submitted.
- J. C. C. Chan and D. P. Kroese. Efficient estimation of large portfolio loss probabilities in  $t$ -copula models. *European Journal of Operational Research*, 205:361–367, 2010.
- J. C. C. Chan and D. P. Kroese. Rare-event probability estimation with conditional Monte Carlo. *Annals of Operations Research*, 2011. Forthcoming.

- J. C. C. Chan, P. W. Glynn, and D. P. Kroese. A comparison of cross-entropy and variance minimization strategies. *Journal of Applied Probability*, 2011. Forthcoming.
- S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321, 1995.
- S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96:270–281, 2001.
- J. Cornebise, É. Moulines, and J. Olsson. Adaptive methods for sequential importance sampling with application to state space models. *Statistics and Computing*, 18:461–480, 2008.
- P. T. de Boer, D. P. Kroese, and R. Y. Rubinstein. A fast cross-entropy method for estimating buffer overflows in queueing networks. *Management Science*, 50:883–895, 2004.
- A. E. Gelfand and D. K. Dey. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3):501–514, 1994.
- A. Gelman and X. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13:163–185, 1998.
- K-P. Hui, N. Bean, M. Kraetzl, and D. P. Kroese. The cross-entropy method for network reliability estimation. *Annals of Operations Research*, 134:101–118, 2005.
- S. Juneja and P. Shahabuddin. Simulating heavy tailed processes using delayed hazard rate twisting. *ACM Transactions on Modeling and Computer Simulation*, 12:94–118, 2002.
- J. M. Keith, D. P. Kroese, and G. Y. Sofronov. Adaptive independence samplers. *Statistics and Computing*, 18:409–420, 2008.
- D. P. Kroese. The cross-entropy method. In *Wiley Encyclopedia of Operations Research and Management Science*. Wiley, 2011.
- D. P. Kroese and R. Y. Rubinstein. The transform likelihood ratio method for rare event simulation with heavy tails. *Queueing Systems*, 46:317–351, 2004.
- M. A. Newton and A. E. Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B*, 56:3–48, 1994.
- G. C. Orsak. A note on estimating false alarm rates via importance sampling. *IEEE Transactions on Communications*, 41(9):1275–1277, 1993.
- A. Philippe. Simulation of right and left truncated gamma distribution by mixtures. *Statistics and Computing*, 7:173–181, 1997.
- A. Ridder. Importance sampling simulations of Markovian reliability systems using cross-entropy. *Annals of Operations Research*, 134:119–136, 2005.
- C. P. Robert. Simulation of truncated normal variables. *Statistics and Computing*, 5:121–125, 1995.
- R. Y. Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99:89–112, 1997.



- R. Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 1(2):127–190, 1999.
- R. Y. Rubinstein and P. W. Glynn. How to deal with the curse of dimensionality of likelihood ratios in Monte Carlo simulation. *Stochastic Models*, 25:547–568, 2009.
- R. Y. Rubinstein and D. P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization Monte-Carlo Simulation, and Machine Learning*. Springer-Verlag, New York, 2004.
- P. J. Smith, M. Shafi, and H. Gao. Quick simulation: a review of importance sampling techniques in communications systems. *IEEE Journal on Selected Areas in Communications*, 15(4):597–613, 1997.