Statistics and Computing manuscript No.
(will be inserted by the editor)

Markov Chain Importance Sampling

with Applications to Rare Event Probability Estimation

Zdravko I. Botev · Pierre L'Ecuyer · Bruno Tuffin

Received: date / Accepted: date

Abstract We present a versatile Monte Carlo method for estimating multidimensional integrals, with applications to rare-event probability estimation. The method fuses two distinct and popular Monte Carlo simulation methods — Markov chain Monte Carlo and importance sampling — into a single algorithm. We show that for some illustrative and applied numerical examples the proposed Markov Chain importance sampling algorithm performs better than methods based solely on importance sampling or MCMC.

Keywords MCMC · importance sampling · non-parametric · minimum variance density · rare-event probability · variance reduction

Mathematics Subject Classification (2000) 65C05 · 65C60 · 65C40

1 Introduction

1.1 Background

A hallmark problem of Monte Carlo simulation is the efficient estimation of high-dimensional integrals of the form:

$$\mathcal{Z} = \int f(\mathbf{x}) H(\mathbf{x})\, \mathrm{d}\mathbf{x} = \mathbb{E}_f H(\mathbf{X}) \,, \tag{1}$$

where the function $H : \mathbb{R}^d \to \mathbb{R}$ and $\mathbf{X}$ is a d-dimensional random variable with probability density function (pdf) f. Such high-dimensional integrals arise in Bayesian statistics, financial mathematics, statistical mechanics, reliability analysis, queuing analysis, and combinatorial counting problems in computer science; see [2, 20, 23].

When direct simulation methods are impractical, a popular method for estimating $\mathcal{Z}$ efficiently is the importance sampling method [2, 26]. A common problem in applying the method is that the selection of a good importance sampling density is not easy, and a poor choice, even if a priori appealing, may lead to mediocre estimators with infinite variance [26]. For specific problems it is possible to construct an efficient importance sampling density derived from large deviations and other asymptotic approximations; see, for example, [1, 10, 16, 21, 22]. The resulting importance sampling schemes are thus always problem-specific and rely on analytical tractability in a some asymptotic regime. In contrast, adaptive importance sampling schemes are more broadly applicable and typically do not require any preliminary asymptotic analysis. There are many proposals for the adaptive (or automatic) selection of the importance sampling density. These include the cross-entropy and the variance minimization methods [2, 28], which have recently been shown (under some strong assumptions) to have similar asymptotic performance on many prominent rare-event probability estimation problems [13]. All of these methods employ an importance sampling density from a given parametric family. As a result, these approaches do not resolve the selection problem satisfactorily, because the choice of the

Z. I. Botev
School of Mathematics and Statistics
University of New South Wales
Sydney NSW 2052 Australia

P. L'Ecyuer
Department of Computer Science and Operations Research, Université de Montréal, Montréal QC H3C 3J7, Canada, Pavillon André-Aisenstadt CP 6128 succ Centre-Ville E-mail: lecuyer@iro.umontreal.ca

Bruno Tuffin
INRIA Rennes Bretagne-Atlantique Campus de Beaulieu 35042 RENNES Cedex, France E-mail: bruno.tuffin@inria.fr

parametric family remains subjective and rarely contains the minimum variance density for the estimation of (1) (which is the reason why problem-specific methods are used if possible).

Given these shortcomings of importance sampling, a number of alternatives have been proposed that use MCMC. The literature on MCMC methods for estimating $\mathcal{Z}$ is vast, but it appears that one of the most efficient and popular methods is Chib's method [15], which has found applications in Bayesian statistics and beyond. A lingering problem with Chib's and other MCMC estimation methods is that the resulting estimators are always biased, because the Markov chain does not sample perfectly from the target density. In addition, MCMC sampling does not generate iid samples and this makes assessment of the estimation error and construction of confidence intervals difficult and convoluted.

Inspired by Chib's approach and new advances in importance sampling, we formulate a novel method called Markov Chain Importance Sampling (MCIS), which combines importance sampling and MCMC sampling in an elegant framework that takes advantage of both. In particular, the MCIS method makes the selection of the importance sampling density more specific to the estimation problem and less subjective. Unlike traditional importance sampling methods [2], the MCIS importance sampling density is model free, that is, it is not part of an arbitrarily selected parametric family of densities. Unlike other MCMC estimators, the MCIS estimator is unbiased and the calculation of the confidence intervals and other error criteria is straightforward.

The rest of the paper is organized as follows. In Section 1.2 we review the standard procedures for constructing importance sampling schemes, and in Section 1.3 we review the currently used MCMC estimation methods. In Section 2 we describe how MCMC and importance sampling can be fused into a single MCIS framework. This is followed by an illustrative numerical example comparing the MCIS method with well-known Monte Carlo integration methods. In Section 3 we focus on applying the MCIS procedure to rare-event probability estimation. Finally, in Section 4 we give concluding remarks and point to directions for future research.

1.2 Importance sampling

One of the most popular variance reduction techniques in Monte Carlo estimation of (1) is *importance sampling* [2,29]. Briefly, the method works as follows. Let g be another probability density such that $H(\mathbf{x})\,f(\mathbf{x})$ is dominated by g, that is, $g(\mathbf{x}) = 0 \;\Rightarrow\; H(\mathbf{x})\,f(\mathbf{x}) = 0$. Using the density g we can represent $\mathcal{Z}$ as

$$\mathcal{Z} = \int H(\mathbf{x})\,\frac{f(\mathbf{x})}{g(\mathbf{x})}\,g(\mathbf{x})\,\mathrm{d}\mathbf{x} = \mathbb{E}_g W(\mathbf{X})\,H(\mathbf{X}),$$

where the ratio of densities $W(\mathbf{x}) = f(\mathbf{x})/g(\mathbf{x})$ is called the *likelihood ratio*. Consequently, if $\mathbf{X}_1,\ldots,\mathbf{X}_m \overset{\text{iid}}{\sim} g$, where $\overset{\text{iid}}{\sim} g$ denotes an iid population with pdf g, then an unbiased estimator of $\mathcal{Z}$ is

$$\widehat{\mathcal{Z}} = \frac{1}{m}\sum_{k=1}^{m} Z_k \quad \text{with} \quad Z_k = H(\mathbf{X}_k)\,\frac{f(\mathbf{X}_k)}{g(\mathbf{X}_k)}\,. \tag{2}$$

If $\widehat{\sigma}$ denotes the sample standard deviation of $Z_1,\ldots,Z_m$, then the estimated *relative error* of $\widehat{\mathcal{Z}}$ is $\widehat{\sigma}/(\sqrt{m}\,|\widehat{\mathcal{Z}}|)$ and a normal approximation based $1-\alpha$ confidence interval for $\mathcal{Z}$ is $\left(\widehat{\mathcal{Z}} - z_{1-\alpha/2}\frac{\widehat{\sigma}}{\sqrt{m}},\ \widehat{\mathcal{Z}} + z_{1-\alpha/2}\frac{\widehat{\sigma}}{\sqrt{m}}\right)$, where z_γ denotes the γ-quantile of the $\mathsf{N}(0,1)$ distribution. This estimator is called the *importance sampling estimator* and g is called the *importance sampling density*. The main difficulty in importance sampling is to select an importance sampling density which yields an estimator with small variance. It is well known that a poor choice of g may seriously compromise the estimate and the confidence intervals [2]. The optimal importance sampling density is the one that minimizes the variance of $\widehat{\mathcal{Z}}$, and is therefore the solution to the functional minimization program

$$\min_g \mathrm{Var}_g\left(H(\mathbf{X})\,W(\mathbf{X})\right)\,. \tag{3}$$

It is well-known (see, for example, [29]) that the solution to this program is the minimum variance importance sampling pdf:

$$\pi(\mathbf{x}) = \frac{|H(\mathbf{x})|\,f(\mathbf{x})}{\int |H(\mathbf{x})|\,f(\mathbf{x})\,\mathrm{d}\mathbf{x}}\,. \tag{4}$$

Note that if $H \geqslant 0$, then (4) is the zero variance importance sampling pdf. Unfortunately, $\pi(\mathbf{x})$ depends on the unknown quantity $\breve{\mathcal{Z}} = \int |H(\mathbf{x})|\,f(\mathbf{x})\,\mathrm{d}\mathbf{x} = \mathcal{Z} - 2\int_{H(\mathbf{x})<0} H(\mathbf{x})\,f(\mathbf{x})\,\mathrm{d}\mathbf{x}$ and cannot be used as an importance sampling density. Nevertheless, a "good" importance sampling density g should be "close" to the minimum variance density π. The closeness between two pdfs π and g is frequently measured by the ϕ-divergence distance

$$\int \pi(\mathbf{x})\;\phi\left(\frac{g(\mathbf{x})}{\pi(\mathbf{x})}\right)\mathrm{d}\mathbf{x}\,, \tag{5}$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}$ is twice continuously differentiable, and $\phi(1) = 0$, $\phi''(x) > 0$, for all $x > 0$. The ϕ-divergence subsumes as special cases most of the information-theoretic distances such as the cross-entropy distance, the Hellinger distance, the Havrda-Charvat α-entropy, Burg's divergence, and Pearson's χ^2 distance; see [29].

Suppose that the density f is parameterized by a vector $\boldsymbol{\theta} \in \Theta$, and that $f(\cdot;\boldsymbol{\theta})$ is embedded in the parametric family of pdfs $\{f(\cdot;\boldsymbol{\eta}), \boldsymbol{\eta} \in \Theta\}$. In general, the importance sampling density g can be a member of some parametric family of densities, which may be very different from the family $\{f(\cdot;\boldsymbol{\eta}), \boldsymbol{\eta} \in \Theta\}$. However, it is often convenient to select the importance sampling distribution from the parametric family of f. More precisely, we select the importance sampling pdf $g(\cdot) \equiv f(\cdot;\boldsymbol{\eta}^*)$ that minimizes a suitable ϕ-divergence distance to π:

$$\boldsymbol{\eta}^* = \operatorname*{argmin}_{\boldsymbol{\eta}\in\Theta} \int f(\mathbf{x};\boldsymbol{\theta})\, |H(\mathbf{x})|\, \phi\left(\frac{f(\mathbf{x};\boldsymbol{\eta})}{\pi(\mathbf{x})}\right) \mathrm{d}\mathbf{x}\,. \tag{6}$$

In practice $\boldsymbol{\eta}^*$ is not available, because the ϕ-divergence cannot be easily computed, and it has to be estimated from the stochastic counterpart of (6):

$$\widehat{\boldsymbol{\eta}^*} = \operatorname*{argmin}_{\boldsymbol{\eta}\in\Theta} \frac{1}{n}\sum_{k=1}^{n} \phi\left(\frac{f(\mathbf{X}_k;\boldsymbol{\eta})}{\pi(\mathbf{X}_k)}\right), \quad \mathbf{X}_1,\ldots,\mathbf{X}_n \sim \pi\,. \tag{7}$$

Cross entropy method. The choice $\phi(x) = -\ln(x)$ for the ϕ-divergence gives the popular *Cross Entropy* (CE) method [28] for the optimal selection of the parameter $\boldsymbol{\eta}^*$. In the CE method, the program (7) simplifies to the maximum likelihood estimation program:

$$\widehat{\boldsymbol{\eta}}_{\mathrm{CE}} = \operatorname*{argmax}_{\boldsymbol{\eta}\in\Theta} \frac{1}{n}\sum_{k=1}^{n} \ln(f(\mathbf{X}_k;\boldsymbol{\eta})), \quad \mathbf{X}_1,\ldots,\mathbf{X}_n \sim \pi\,, \tag{8}$$

where (approximate) sampling from π is typically accomplished using MCMC and we assume that the argmax exists. Note that if $f(\cdot;\boldsymbol{\eta})$ is a model from the exponential family, then (8) is a convex optimization program and finding the global maximizer $\widehat{\boldsymbol{\eta}}_{\mathrm{CE}}$ is not difficult; see [20].

Variance minimization method. The choice $\phi(z) = 1/z$ for the ϕ-divergence gives the *Variance Minimization* (VM) method [2,29]. In the VM method the stochastic counterpart (7) simplifies to the nonlinear optimization program

$$\widehat{\boldsymbol{\eta}}_{\mathrm{VM}} = \operatorname*{argmin}_{\boldsymbol{\eta}\in\Theta} \frac{1}{n}\sum_{k=1}^{n} \frac{|H(\mathbf{X}_k)| f(\mathbf{X}_k;\boldsymbol{\theta})}{f(\mathbf{X}_k;\boldsymbol{\eta})}, \; \mathbf{X}_1,\ldots,\mathbf{X}_n \sim \pi. \tag{9}$$

Theoretical analysis has shown that the solutions to the VM and CE programs are qualitatively similar [13], and that an advantage of the CE over the VM approach is that often one can compute $\widehat{\boldsymbol{\eta}}_{\mathrm{CE}}$ analytically (whenever maximum likelihood estimation yields closed form solutions) without resorting to costly nonlinear optimization.

A disadvantage of both the CE and VM methods is that the choice of the parametric family $\{f(\cdot;\boldsymbol{\eta}), \boldsymbol{\eta} \in \Theta\}$ is frequently arbitrary and there is no reason why it should contain a good (or at least a near optimal) importance sampling density. Instead of restricting the search for a good importance sampling density g to a simple parametric form $g(\cdot) \equiv f(\cdot;\boldsymbol{\eta})$, we may optimize (5) over a flexible and complicated class of parametric densities. Unfortunately, expanding the parametric family does not necessarily lead to better efficiency of the importance sampling estimator. There are two reasons for this. First, the larger the family of importance sampling densities, the more difficult it is to estimate $\mathbb{E}_\pi \phi(f(\mathbf{X};\boldsymbol{\eta})/\pi(\mathbf{X}))$ on the right-hand side of the stochastic counterpart (7) (necessitating large sample size n) and the more noisy is the estimator $\widehat{\boldsymbol{\eta}^*}$. This undesirable phenomenon is related to the so-called *curse of dimensionality*, because increasing the complexity of the parametric family typically increases the dimensionality of the parameter vector $\boldsymbol{\eta}^*$. Second, a more flexible and complex parametric family of densities makes the nonlinear optimization in (7) more complex (for example, by increasing the dimension of the search space).

We will show that the MCIS method provides a model-free (nonparametric or semi-parametric) importance sampling density with the advantage that there is no need for a costly nonlinear optimization as in (7) or for tuning a *bandwidth* parameter as in standard kernel density estimation [30].

1.3 MCMC sampling and estimation

Markov chain Monte Carlo (MCMC) was proposed by Metropolis et al. [24] for (approximate) sampling from an arbitrary complex *target* density π. The main idea is to generate a Markov chain whose limiting distribution is equal to the desired distribution. The Gibbs sampler (see, for example, [2,29]) is a special MCMC algorithm for generating d-dimensional random vectors by constructing a Markov chain from a sequence of conditional distributions. Briefly, the Gibbs sampler works as follows. Suppose that we wish to generate a random vector $\mathbf{X} = (X_1,\ldots,X_d)$ with (approximate) density π. Let $\pi(\cdot \,|\, x_1,\ldots,x_{i-1},\, x_{i+1},\ldots,x_d)$ be the conditional pdf of the X_i component, given all the other components $x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d$.

Algorithm 11 (Systematic Gibbs Sampler)
Require: An initial state $\mathbf{X}_1$ and sample size n.
for $t = 1,\ldots,n-1$ **do**

Set $\mathbf{Y} = \mathbf{X}_t$. Generate $Y_1 \sim \pi(y_1 \,|\, X_{t,2}, \ldots, X_{t,d})$.
for $i = 2, \ldots, d-1$ **do**
Draw $Y_i \sim \pi(y_i \,|\, Y_1, \ldots, Y_{i-1}, X_{t,i+1}, \ldots, X_{t,d})$.
Draw $Y_d \sim \pi(y_d \,|\, Y_1, \ldots, Y_{d-1})$. Set $\mathbf{X}_{t+1} = \mathbf{Y}$.

The pdf of the transition $\mathbf{x} \to \mathbf{y}$ in the resulting Markov chain is given by

$$\kappa(\mathbf{y} \,|\, \mathbf{x}) = \prod_{i=1}^{d} \pi(y_i \,|\, y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d) \,. \tag{10}$$

It is well known [26] that

$$\int \pi(\mathbf{x})\, \kappa(\mathbf{y} \,|\, \mathbf{x})\, \mathrm{d}\mathbf{y} = \mathbb{E}_\pi[\kappa(\mathbf{y} \,|\, \mathbf{X})] = \pi(\mathbf{y}) \,, \tag{11}$$

from which we can conclude that π is the stationary pdf of the Markov chain $\{\mathbf{X}_t, t = 1, 2, \ldots\}$ with transition density $\kappa(\mathbf{y} \,|\, \mathbf{x})$. In addition, if for every $\mathbf{y} \in \mathbb{R}^d$ the *positivity condition* holds:

$$\prod_{i=1}^{d} \pi_i(y_i) > 0 \quad \text{implies} \quad \pi(\mathbf{y}) > 0 \tag{12}$$

(here π_i is the i-th marginal density of π), then the Markov chain $\{\mathbf{X}_t, t = 1, 2, \ldots\}$ is irreducible and recurrent with limiting pdf π; see [26], where much weaker technical conditions are given as well. As a consequence, to estimate, for example, the ϕ-divergence (5) one can use the estimator $\frac{1}{n}\sum_{t=1}^{n} \phi(g(\mathbf{X}_t)/\pi(\mathbf{X}_t))$.

One of the most widely used MCMC methods for estimation of $\mathcal{Z}$ using the output of a MCMC sampler is Chib's method [15]. For simplicity assume that $H(\mathbf{x})$ in (4) is positive. Then from (4), we have the identity $H(\mathbf{x}^*)f(\mathbf{x}^*)/\pi(\mathbf{x}^*) = \mathcal{Z}$ for any point $\mathbf{x}^*$ in the support of π. It follows that if $\widehat{\pi}(\mathbf{x}^*)$ is an estimator of $\pi(\mathbf{x})$ at $\mathbf{x}^*$, then we may estimate $\mathcal{Z}$ via $\widehat{\mathcal{Z}}_{\text{Chib}} = H(\mathbf{x}^*)f(\mathbf{x}^*)/\widehat{\pi}(\mathbf{x}^*)$, where for numerical accuracy $\mathbf{x}^*$ is a carefully chosen point — typically a mode of π. We now introduce the following notation:

- $\pi_1(x_1)$ denotes the marginal density of X_1 evaluated at x_1;
- $\pi_{2\,|\,1}(x_2 \,|\, x_1)$ denotes the marginal density of X_2 given $X_1 = x_1$;
- $\pi_{3\,|\,1,2}(x_3 \,|\, x_1, x_2)$ denotes the marginal density of X_3 given $X_1 = x_1$ and $X_2 = x_2$;

Note that we use a subscript notation like $\pi_{3\,|\,1,2}$ only for marginal densities, but not for the full conditional densities $\pi(\cdot \,|\, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$, $i = 1, \ldots, d$. Given the identity

$$\pi(\mathbf{x}^*) = \pi_1(x_1^*)\, \pi_{2\,|\,1}(x_2^* \,|\, x_1^*) \times \pi_{3\,|\,1,2}(x_3^* \,|\, x_1^*, x_2^*) \cdots \pi(x_d^* \,|\, x_1^*, \ldots, x_{d-1}^*) \,,$$

Chib proposes to estimate $\pi(\mathbf{x}^*)$ by estimating each of the d terms on the right-hand side of the identity and then computing their product. In particular, the marginal pdf of X_k given $(X_1, \ldots, X_{k-1}) = (x_1^*, \ldots, x_{k-1}^*)$, that is $\pi_{k\,|\,1,\ldots,k-1}(\cdot \,|\, x_1^*, \ldots, x_{k-1}^*)$, is estimated at the point x_k^* via

$$\widehat{\pi}_{k\,|\,1,\ldots,k-1} = \frac{1}{m}\sum_{t=1}^{m} \pi(x_k^* \,|\, x_1^*, \ldots, x_{k-1}^*, X_{t,k+1}, \ldots, X_{t,d}) \,, \tag{13}$$

where

$$(X_{t,k}, X_{t,k+1}, \ldots, X_{t,d}) \overset{\text{approx}}{\sim} \pi(x_k, \ldots, x_d \,|\, x_1^*, \ldots, x_{k-1}^*)$$

for $t = 1, \ldots, m$ are obtained from a Gibbs run (different for each k) in which $(x_1, \ldots, x_{k-1})$ is fixed to $(x_1^*, \ldots, x_{k-1}^*)$, the component x_k is discarded, and the Gibbs sampler runs over $(x_k, \ldots, x_d)$. Hence, the Chib estimator is simply

$$\widehat{\mathcal{Z}}_{\text{Chib}} = \frac{H(\mathbf{x}^*)f(\mathbf{x}^*)}{\pi(x_d^* \,|\, x_1^*, \ldots, x_{d-1}^*)\prod_{k=1}^{d-1}\widehat{\pi}_{k\,|\,1,\ldots,k-1}}, \tag{14}$$

where each $\widehat{\pi}_{k\,|\,1,\ldots,k-1}$ is given in (13) and the conditional density $\pi(\cdot \,|\, x_1, \ldots, x_{d-1})$ is available analytically. Possible problems with this approach include the difficulty in computing empirical and asymptotic error estimates for (14), the bias of the Chib estimator, and the reliance of the estimator on multiple Markov chains (as opposed to a single one). Note also the problem of the somewhat arbitrary location of a suitable high-density point $\mathbf{x}^*$.

2 Markov Chain Importance Sampling

Similar to the cross entropy and variance minimization methods, the MCIS method consists of two stages:

- **Markov Chain (MC)** stage, in which we construct an estimator of the minimum variance importance sampling density (4).
- **Importance Sampling (IS)** stage, in which we use the constructed pdf as an importance sampling density to estimate $\mathcal{Z}$.

Suppose that we have used an MCMC sampler to generate the population

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{\text{approx}}{\sim} \pi(\mathbf{x}), \quad \mathbf{X}_i = (X_{i,1}, \ldots, X_{i,d}) \,. \tag{15}$$

For simplicity and concreteness we may assume that the sample was generated by the Gibbs sampling Algorithm 11. We now present two possible ways (one semi-parametric and the other nonparametric) in which the Markov chain output (15) can be used to construct an importance sampling density (the IS stage).

2.1 Construction of a semi-parametric importance sampling density

Suppose that instead of minimizing (5) over a simple parametric family of densities $\{g(\cdot) \equiv f(\cdot;\boldsymbol{\eta}), \boldsymbol{\eta} \in \Theta\}$, we minimize the ϕ-divergence over all densities of product form:

$$g(\mathbf{y}) = \prod_{i=1}^{d} g_i(y_i),$$

where

$$g_i \in \mathscr{G} \equiv \left\{ g : \mathbb{R} \to [0,\infty),\ \int g(y)\,\mathrm{d}y = 1 \right\}.$$

In other words, instead of solving the parametric optimization program (6), we now solve the functional optimization program:

$$\min_{\substack{g_i \in \mathscr{G} \\ i=1,\ldots,d}} \int \pi(\mathbf{y})\, \phi\left(\frac{\prod_{i=1}^{d} g_i(y_i)}{\pi(\mathbf{y})} \right) \mathrm{d}\mathbf{y}. \tag{16}$$

With $\phi(y) = -\ln(y)$ (giving the cross entropy distance), the solution is $g_i(y_i) = \pi_i(y_i)$ for all i, where π_i is the marginal density of Y_i with $\mathbf{Y} \sim \pi(\mathbf{y})$. To see this, write

$$-\int \pi(\mathbf{y})\, \ln\left(\frac{\prod_{i=1}^{d} g_i(y_i)}{\pi(\mathbf{y})} \right) \mathrm{d}\mathbf{y}$$

$$= \int \pi(\mathbf{y})\, \ln\left(\pi(\mathbf{y})\right) \mathrm{d}\mathbf{y} - \sum_{i=1}^{d} \int \pi_i(y_i) \ln(g_i(y_i))\, \mathrm{d}y_i$$

$$= \sum_{i=1}^{d} \int \pi_i(y_i) \ln\left(\frac{\pi_i(y_i)}{g_i(y_i)} \right) \mathrm{d}y_i + \text{terms without } g_i\ .$$

The functional optimization program (16) is then equivalent to

$$\min_{g_i \in \mathscr{G}} \int \pi_i(y_i) \ln\left(\frac{\pi_i(y_i)}{g_i(y_i)} \right) \mathrm{d}y_i, \quad i = 1,\ldots,d,$$

which we recognize as the cross entropy distances between π_i and g_i for all i. These distances are zero if and only if $g_i \equiv \pi_i$ for all i. Hence, the pdf $g(\mathbf{y}) = \prod_{i=1}^{d} \pi_i(y_i)$, which is the product of the marginal densities of (4), is the best (in the cross entropy sense) importance sampling density of product form. Since the marginals are typically not available in closed form, we use the Markov chain output (15) to estimate each marginal density $\pi_i(y_i)$ via:

$$\widehat{\pi}_i(y_i) = \frac{1}{n} \sum_{k=1}^{n} \pi(y_i \,|\, \mathbf{X}_{k,-i})\ ,$$

where $\mathbf{X}_{k,-i}$ is the same as vector $\mathbf{X}_k$, except that the i-th component is removed. Thus, in the MC stage we can construct the *semi-parametric importance sampling pdf*:

$$\widehat{g}(\mathbf{y}) = \prod_{i=1}^{d} \widehat{\pi}(y_i) = \prod_{i=1}^{d} \frac{1}{n} \sum_{k=1}^{n} \pi(y_i \,|\, \mathbf{X}_{k,-i})\ , \tag{17}$$

which looks similar to Besag's pseudo-likelihood [4]. Note that generating $\mathbf{Y} \sim \widehat{g}(\mathbf{y})$ is straightforward, because each Y_i is generated from the mixture $\widehat{\pi}_i(y_i)$ and, given $\{\mathbf{X}_k\}$, independently from all other components of $\mathbf{Y}$. The estimator $\widehat{g}$ is semi-parametric (as opposed to nonparametric), because $\widehat{g}$ does not converge to π as $n \uparrow \infty$ (unless π is of product form).

In the IS stage of the MCIS method we generate the iid population $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ from the pdf $\widehat{g}$ and then deliver the estimator:

$$\widehat{\mathscr{Z}} = \frac{1}{m} \sum_{k=1}^{m} \frac{|H(\mathbf{Y}_k)|\, f(\mathbf{Y}_k)}{\widehat{g}(\mathbf{Y}_k)}\ . \tag{18}$$

We defer giving an example to the next section and we first explain how we can use the Markov chain sample (15) to construct a fully nonparametric estimator of π.

2.2 Construction of nonparametric importance sampling density

From the global balance equation (11) and the availability of the transition density (10) in closed form, we can construct the *nonparametric importance sampling pdf*:

$$\widehat{\pi}(\mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \kappa(\mathbf{y} \,|\, \mathbf{X}_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} \pi(y_j \,|\, y_1, \ldots, y_{j-1}, X_{i,j+1}, \ldots, X_{i,d})\ . \tag{19}$$

Note that, while the normalization constant of π itself is not available, the availability of the normalization constants of the conditional densities of π in (19) is what makes $\widehat{\pi}$ practicable as an importance sampling density. The advantage of the importance sampling density (19) is that we are no longer restricted by rigid parametric models such as those used in (8) and (9). Unlike in the CE and VM methods, we have $\mathbb{P}\left(\lim_{n\to\infty} \widehat{\pi}(\mathbf{y}) = \pi(\mathbf{y})\right) = 1$ for every $\mathbf{y}$, see, for example, [26, Page 240]. It is in this sense that we classify $\widehat{\pi}$ as a nonparametric estimator—we can recover the true π as $n \uparrow \infty$, at least in principle. We also classify the

product model (17) as a semi-parametric model, because, while in general not converging to π as $n \uparrow \infty$, it involves the estimation of the infinite dimensional marginal densities $\{\pi_i\}$.

Another advantage of the MCIS construction of the importance sampling density is that here we need not solve any nonlinear optimization programs such as (8) and (9). A disadvantage of the nonparametric estimator (19) is that it may suffer from the curse of dimensionality — the variance of the estimator $\widehat{\pi}$ of π may deteriorate with increasing d. In this sense, the semi-parametric importance sampling density (17) is a compromise between the nonparametric model (19) and the parametric models of the CE and VM methods. On the one hand, the semi-parametric model does not require any nonlinear optimization and is less sensitive to the curse of dimensionality, because it estimates densities on subspaces of $\mathbb{R}^d$. On the other hand, the product form of the semi-parametric model may fail to capture some important interdependence between the components of $\mathbf{Y} \sim \pi(\mathbf{y})$.

Note that sampling from $\widehat{\pi}$ is straightforward using the composition method: Sample a random index J uniformly from the set $\{1, \ldots, n\}$ and then generate $\mathbf{Y} \sim \kappa(\mathbf{y} \,|\, \mathbf{X}_J)$. The IS stage of the MCIS method simply consists of generating the iid population $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ using the composition method above, and then computing the estimator:

$$\widehat{\mathcal{Z}} = \frac{1}{m} \sum_{k=1}^{m} \frac{|H(\mathbf{Y}_k)|\, f(\mathbf{Y}_k)}{\widehat{\pi}(\mathbf{Y}_k)} \,. \tag{20}$$

Note that if the positivity condition (12) holds, then $\widehat{\pi}$ *dominates* π, that is $\widehat{\pi}(\mathbf{y}) = 0 \Rightarrow |H(\mathbf{y})| f(\mathbf{y}) = 0$ for all $\mathbf{y}$. In Section 3 we discuss problems for which (12) does not hold.

2.3 Generic MCIS algorithm

In summary, a generic version of the MCIS algorithm reads as follows.

Algorithm 21 (MCIS for the estimation of $\mathcal{Z}$)

1. **MC step.** *Run any MCMC sampler with stationary density* (4) *to generate the population* $\mathbf{X}_1, \ldots, \mathbf{X}_n$.
2. **IS step.** *Generate an iid population* $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ *from the semi-parametric importance sampling pdf* (17) *(or from the nonparametric importance sampling pdf* (19)*) and deliver the importance sampling estimator* (18) *(or the importance sampling estimator* (20)*). The estimated relative error of the estimator is* $\widehat{\sigma}/(\sqrt{m}\,|\widehat{\mathcal{Z}}|)$*, where* $\widehat{\sigma}$ *is the sample standard deviation of the population* $\{|H(\mathbf{Y}_k)|\, f(\mathbf{Y}_k)/\widehat{g}(\mathbf{Y}_k)\}$ *(or the sample standard deviation of the population* $\{|H(\mathbf{Y}_k)|\, f(\mathbf{Y}_k)/\widehat{\pi}(\mathbf{Y}_k)\}$*).*

Remark 1 (Different transition densities) In principle the MC step can be executed using a Markov transition density different from the κ used in the IS step. While the IS step requires the transition density to have a simple form and satisfy the positivity condition (12), there are no such requirements for the MC step (the positivity condition is not required for ergodicity [26]). Thus, for the MC step we may use, for example, an MCMC sampler based on the generalized splitting method [6].

Remark 2 (Dependence of error on Markov chain) Whether the estimated relative error $\widehat{\sigma}/(\sqrt{m}|\widehat{\mathcal{Z}}|)$ in the IS step of Algorithm 21 is small or large depends in part on the mixing speed of the Markov chain in the MC step. This is because the quality of the nonparametric or semi-parametric approximation to the optimal importance sampling density π depends on the MCMC output $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

2.4 Numerical example

We illustrate the effectiveness of the nonparametric version of the MCIS method (that is, using (19) as importance sampling pdf) as a Monte Carlo variance reduction method on a simple bridge network test problem borrowed from [20] and depicted on Figure 1.

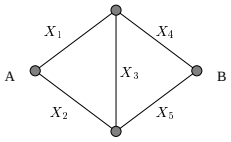

Fig. 1 A bridge network with four nodes and five links. The i-th link has a random length given by X_i.

The problem is to compute the expected length, say $\mathcal{Z}$, of the shortest path between the nodes A and B, where the five links have lengths given by the random variables $X_1, \ldots, X_5$. We assume that the lengths are independent and $X_i \sim \mathsf{U}(0, a_i)$ for all i with $(a_1, \ldots, a_5) = (1, 2, 3, 1, 2)$. In other words, we wish to compute $\mathcal{Z} = \mathbb{E}_f H(\mathbf{X})$, where f is the uniform density

$$f(\mathbf{x}) = \prod_{i=1}^{5} \frac{\mathrm{I}\{0 < x_i < a_i\}}{a_i}, \quad \mathbf{x} \in \mathbb{R}^5 \tag{21}$$

and the function H is defined as follows ($a \wedge b \stackrel{\text{def}}{=} \min\{a, b\}$)

$$H(\mathbf{x}) = (x_1 + x_4) \wedge (x_2 + x_5) \wedge (x_1 + x_3 + x_5) \wedge (x_2 + x_3 + x_4) \,.$$

To apply the MCIS Algorithm 21 we first derive the conditional densities of (4). Defining $x^+ \stackrel{\text{def}}{=} \max\{0, x\}$, we can, after some straightforward manipulations, write the conditional densities as

$$\pi(y_i \,|\, \mathbf{x}_{-i}) \propto \begin{cases} \alpha_i, & \eta_i < y_i < a_i \\ y_i + \beta_i, & 0 < y_i < \eta_i \end{cases}, \quad \eta_i = a_i \wedge (\alpha_i - \beta_i)^+,$$

where

$$\begin{aligned} &\alpha_1 = (x_2 + x_3 + x_4) \wedge (x_2 + x_5), \beta_1 = x_4 \wedge (x_3 + x_5) \\ &\alpha_2 = (x_1 + x_3 + x_5) \wedge (x_1 + x_4), \beta_2 = x_5 \wedge (x_3 + x_4) \\ &\alpha_3 = (x_1 + x_4) \wedge (x_2 + x_5), \qquad \beta_3 = (x_1 + x_5) \wedge (x_2 + x_4) \\ &\alpha_4 = (x_1 + x_3 + x_5) \wedge (x_2 + x_5), \beta_4 = x_1 \wedge (x_2 + x_3) \\ &\alpha_5 = (x_1 + x_4) \wedge (x_2 + x_3 + x_4), \beta_5 = x_2 \wedge (x_1 + x_3), \end{aligned} \tag{22}$$

and the constant of proportionality is $\delta_i = \alpha_i(a_i - \eta_i) + \frac{1}{2}\eta_i^2 + \beta_i\eta_i$. In other words, the conditional densities can be written as a mixture of two uniform densities (on disjoint intervals) and the density

$$\frac{y_i \, \mathrm{I}\{0 < y_i < \eta_i\}}{\eta_i^2/2} \,,$$

that is,

$$\pi(y_i \,|\, \mathbf{x}_{-i}) = \frac{\alpha_i(a_i - \eta_i)}{\delta_i} \frac{\mathrm{I}\{\eta_i < y_i < a_i\}}{a_i - \eta_i} + \frac{\eta_i^2}{2\delta_i} \frac{y_i \, \mathrm{I}\{0 < y_i < \eta_i\}}{\eta_i^2/2} + \frac{\beta_i \, \eta_i}{\delta_i} \frac{\mathrm{I}\{0 < y_i < \eta_i\}}{\eta_i} \,.$$

Given these conditional densities, for the Markov chain step of the MCIS algorithm (Step 1) we apply the Gibbs Algorithm 11 with $n = 100$ and $\mathbf{X}_1 = (a_1, \ldots, a_5)/2$. For the IS step (Step 2) we use a sample size of $m = 10^4$ with the importance sampling density given in (19).

To compare the MCIS method with another Markov chain based method for estimation, we implemented Chib's estimator (14) as follows. We used a Gibbs sample size of $m = 10^4$ for the estimation of all components (13) with $\mathbf{x}^* = (a_1, \ldots, a_5)/2$. To estimate the relative error of (14) we repeated the simulation ten independent times. The result is given in Table 1, which also shows the comparative performance of the Cross Entropy and Variance Minimization methods. For both the VM and CE methods we used the importance sampling estimator (2) with $m = 10^4$ and importance sampling density

$$g(\mathbf{x}) \equiv f(\mathbf{x}; \boldsymbol{\eta}) = \prod_{i=1}^{5} \frac{\eta_i}{a_i} \left(\frac{x_i}{a_i}\right)^{\eta_i - 1}, \quad x_i \in (0, a_i) \,. \tag{23}$$

Note that $f(\mathbf{x}; \mathbf{1})$ yields the original uniform distribution. For the VM method we used the estimated parameter $\widehat{\boldsymbol{\eta}}_{\text{VM}} = (1.26, 1.08, 1.01, 1.23, 1.06)$, computed from (9) with $n = 10^3$. For the CE method we used the estimated CE parameter $\widehat{\boldsymbol{\eta}}_{\text{CE}} = (1.27, 1.12, 1.00, 1.32, 1.07)$, computed from (8) with $n = 10^3$. For all methods we used a Monte Carlo sample size of $m = 10^4$ in the importance sampling stage, and for each method we recorded the resulting estimate together with the estimated relative error and the corresponding variance reduction factor (relative to crude Monte Carlo). For example, the VM approach gives an estimator with variance approximately 3 times smaller than the variance of the crude Monte Carlo estimator.

From the table we can see that for this particular example the MCIS method gives the smallest relative error and the largest variance reduction factor of 810. In our implementation the computational time for the MCIS method was roughly the same as that for the Chib estimator.

3 MCIS for rare-event probability estimation

An important class of estimation problems of the form (1) is rare-event probability estimation, in which $H(\mathbf{x}) = \mathrm{I}\{S(\mathbf{x}) \geqslant \gamma\}$ for some function $S(\mathbf{x})$ and a level or threshold parameter γ. In this case $\mathcal{Z} = \mathbb{E}_f H(\mathbf{X}) = \mathbb{P}_f(S(\mathbf{X}) \geqslant \gamma)$. In many interesting problems $\mathcal{Z}$ happens to be a very small probability, say, smaller than 10^{-4}, and in such cases the event $\{S(\mathbf{X}) \geqslant \gamma\}$ is called a *rare-event* and $\mathcal{Z}$ is called a *rare-event probability* [2, 5, 27].

Suppose for the moment that $f(\mathbf{x}) = \prod_{j=1}^{d} f_j(x_j)$, that is, the components of the vector $\mathbf{X}$ are independent and we wish to estimate the rare-event probability $\mathcal{Z} = \mathbb{P}_f(S(\mathbf{X}) \geqslant \gamma)$ using Algorithm 21. Then, a straightforward calculation shows that the estimator (20) simplifies to $\widehat{\mathcal{Z}} = \frac{n}{m} \sum_{k=1}^{m} \left(\sum_{i=1}^{n} \prod_{j=1}^{d} \mathrm{I}_{k,i,j} / c_{k,i,j} \right)^{-1}$, where the conditional probability $c_{k,i,j}$ is given by

$$c_{k,i,j} = \mathbb{P}(S(Y_{k,1}, \ldots, Y_{k,j}, X_{i,j+1}, \ldots, X_{i,d}) \geqslant \gamma \,| \text{ given everything except } Y_{k,j}) \,,$$

and $\mathrm{I}_{k,i,j}$ is the indicator of the event:

$$\{S(Y_{k,1}, \ldots, Y_{k,j}, X_{i,j+1}, \ldots, X_{i,d}) \geqslant \gamma\}.$$

Note that in the rare-event setting the positivity condition (12) does not hold and for a finite n it is possible that $\sum_i \prod_j \mathrm{I}_{k,i,j} = 0$ for at least one k, making the estimator $\widehat{\mathcal{Z}}$ invalid. The reason for this is that for a valid importance sampling estimator (20) we need to ensure

Table 1 Empirical performance of widely used importance sampling and MCMC integration methods on the bridge network.

	method of integration	estimate	relative error	variance reduction factor
5	crude Monte Carlo	0.93	0.43%	1.00
4	cross entropy method	0.9289	0.25%	3
3	variance minimization	0.9295	0.24%	3
2	Chib's method	0.9296	0.09%	23
1	MCIS method	0.92978	0.015%	810

that the importance sampling density $\widehat{\pi}$ dominates π. However, while

$$\mathbb{P}(\lim_{n\to\infty} \widehat{\pi}(\mathbf{y}) = \pi(\mathbf{y})) = 1$$

for all $\mathbf{y}$, for finite n there is no simple way to ensure that $\widehat{\pi}$ will have the same support as the density $\pi(\mathbf{y}) \propto f(\mathbf{y})\mathrm{I}\{S(\mathbf{y}) \geqslant \gamma\}$ (unless the transition density (10) dominates π). One way to avoid this problem is to use the fact that the product of the marginal densities $\prod_{i=1}^{d} \pi_i(x_i)$ (approximated by $\widehat{g}(\mathbf{x})$) has equal or larger support than the joint density $\pi(\mathbf{x})$. Thus, in rare-event settings we may use the following mixture density as an importance sampling pdf:

$$\widehat{\pi}_w(\mathbf{x}) = w\,\widehat{g}(\mathbf{x}) + (1-w)\,\widehat{\pi}(\mathbf{x}), \quad w \in (0,1)\,, \tag{24}$$

where $\widehat{g}$ is the density in (17). In this way, if we can ensure that $\widehat{g}$ dominates π, then $\widehat{\pi}_w$ will dominate π for any $w \in (0,1)$, and we have the valid importance sampling estimator ($\mathbf{Y}_1, \ldots, \mathbf{Y}_m \overset{\text{iid}}{\sim} \widehat{\pi}_w(\mathbf{y})$)

$$\widehat{\mathcal{Z}}_w = \frac{1}{m}\sum_{k=1}^{m} \frac{f(\mathbf{Y}_k)\mathrm{I}\{S(\mathbf{Y}_k) \geqslant \gamma\}}{w\,\widehat{g}(\mathbf{Y}_k) + (1-w)\widehat{\pi}(\mathbf{Y}_k)}\,. \tag{25}$$

Typically, it is easier to ensure that $\widehat{g}$ dominates π than to ensure that $\widehat{\pi}$ dominates π. For example, (17) simplifies to

$$\widehat{g}(\mathbf{x}) = \frac{1}{n^d}\prod_{i=1}^{d}\sum_{k=1}^{n} \frac{f(x_i \,|\, \mathbf{X}_{k,-i})}{\mathbb{P}(S(\mathbf{X}_k) \geqslant \gamma \,|\, \mathbf{X}_{k,-i})} \times$$

$$\mathrm{I}\{S(X_{k,1}, \ldots, X_{k,i-1}, x_i, X_{k,i+1}, \ldots, X_{k,d}) \geqslant \gamma\}\,,$$

which dominates π if

$$S(X_{k,1}, \ldots, X_{k,i-1}, x_i, X_{k,i+1}, \ldots, X_{k,d}) \geqslant \gamma$$

for at least one k, for all i. In other words, for each problem we verify that for each i there is at least one vector out of the n samples (15) such that

$$S(X_{k,1}, \ldots, X_{k,i-1}, x_i, X_{k,i+1}, \ldots, X_{k,d}) \geqslant \gamma\,.$$

If this condition is not true, then we must increase n until it is satisfied.

Example 1 (Tail probabilities for sums of correlated log-normals)

Consider the estimation of the rare-event probability

$$\mathcal{Z} = \mathbb{P}(\mathrm{e}^{X_1} + \cdots + \mathrm{e}^{X_d} \geqslant \gamma) = \int f(\mathbf{x})\,\mathrm{I}\{S(\mathbf{x}) \geqslant \gamma\}\,\mathrm{d}\mathbf{x}\,,$$

where $\mathbf{x} = (x_1, \ldots, x_d)$, $\mathbf{X} = (X_1, \ldots, X_d) \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$, $S(\mathbf{x}) = \mathrm{e}^{x_1} + \cdots + \mathrm{e}^{x_d}$, and f is the density of the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma = (\Sigma_{i,j})$ with associated precision matrix $\Lambda = (\Lambda_{i,j}) = \Sigma^{-1}$. Such rare-event probabilities are of significant interest in financial engineering [18]. We now estimate $\mathcal{Z}$ via the importance sampling estimator (25). To compute the nonparametric (17) and semi-parametric (19) densities we derive the conditional densities of the optimal importance sampling pdf $\pi(\mathbf{y}) = f(\mathbf{y})\mathrm{I}\{S(\mathbf{x}) \geqslant \gamma\}/\mathcal{Z}$:

$$\pi(y_i \,|\, \mathbf{y}_{-i}) \propto$$

$$\begin{cases} f(y_i \,|\, \mathbf{y}_{-i}) & \text{if } \sum_{j\neq i} \mathrm{e}^{y_j} \geqslant \gamma \\ f(y_i \,|\, \mathbf{y}_{-i})\,\mathrm{I}\Big\{y_i \geqslant \ln(\gamma - \sum_{j\neq i} \mathrm{e}^{y_j})\Big\} & \text{if } \sum_{j\neq i} \mathrm{e}^{y_j} < \gamma \end{cases},$$

where via standard calculations (see, for example, [20, Page 146]) $f(y_i \,|\, \mathbf{y}_{-i})$ is the univariate normal density with mean

$$\mu_i + \frac{1}{\Lambda_{i,i}}\sum_{j\neq i} \Lambda_{i,j}(\mu_j - y_j) \quad \text{and variance} \quad \frac{1}{\Lambda_{i,i}}\,.$$

Thus, depending on $\sum_{j\neq i} \mathrm{e}^{y_j}$, the distribution of $Y_i \sim \pi(y_i \,|\, \mathbf{y}_{-i})$ is either a normal or a truncated normal density with the above mean and variance.

As a numerical example consider Table 2, where we compute $\mathcal{Z}$ for various values of the common correlation coefficient

$$\frac{\Sigma_{i,j}}{\sqrt{\Sigma_{i,i}\Sigma_{j,j}}} = \varrho, \quad \text{for all } i \neq j.$$

The rest of the parameters of the density f are set to: $d = 10$, $\mu_i = i - 10$, $\sigma_i^2 = i$ $(i = 1, \ldots, d)$, $\gamma = 5 \times 10^5$. We used the estimator (25) with $m = 5 \times 10^5$, $w = 0.01$ and a Markov chain sample of size $n = 80$ obtained using the splitting Algorithm A1 given in Appendix A.

We compare the MCIS results with two recommended importance sampling schemes given in [1] — the importance sampling vanishing relative error estimator (ISVE) and the cross entropy vanishing relative error estimator (CEVE) with a sample size of 5×10^5. Both of these estimators rely on the decomposition of the probability $\mathcal{Z}(\gamma)$ into two parts:

$$\mathcal{Z}(\gamma) = \mathbb{P}(\max_i \mathrm{e}^{X_i} \geqslant \gamma) + \mathbb{P}(S(\mathbf{X}) \geqslant \gamma,\ \max_i \mathrm{e}^{X_i} < \gamma),$$

where the asymptotically dominant term $\mathbb{P}(\max_i \mathrm{e}^{X_i} \geqslant \gamma)$ is estimated via an importance sampling estimator and (the so called residual term) $\mathbb{P}(S(\mathbf{X}) \geqslant \gamma,\ \max_i \mathrm{e}^{X_i} < \gamma)$ is estimated by a second and different importance sampling estimator in a way that guarantees the strong efficiency of the sum of the two estimators. The first term is asymptotically dominant in the sense that

$$\lim_{\gamma\to\infty} \frac{\mathbb{P}(S(\mathbf{X}) \geqslant \gamma,\ \max_i \mathrm{e}^{X_i} < \gamma)}{\mathbb{P}(\max_i \mathrm{e}^{X_i} \geqslant \gamma)} = 0\ .$$

For more details, see [1].

From the table we can see that for medium and high correlation (and $\mathcal{Z}$ of the order 10^{-5}) the MCIS estimator outperforms the ISVE and CEVE importance sampling estimators, both of which enjoy the property of vanishing relative error [1] as $\mathcal{Z} \downarrow 0$. Note that the ISVE and CEVE estimators deteriorate as the correlation coefficient ϱ becomes larger, and the MCIS estimator performs significantly better for large values of ϱ.

Table 3 illustrates that the ISVE estimator enjoys the property of vanishing relative error when $\gamma \uparrow \infty$ (so that $\mathcal{Z} \downarrow 0$). The table was generated using the same algorithmic and problem parameters as the ones used for Table 2, except that $\varrho = 0.9$ in all cases. The third and fourth columns of the table show the relative error achieved by MCIS and ISVE methods using the same sample size of $m = 5 \times 10^5$. Although the estimates for the relative error of both MCIS and ISVE methods are noisy, from the table we can see the trend that the MCIS method is more accurate in estimating probabilities larger than about 10^{-14}, and the ISVE method eventually becomes more accurate for probabilities smaller than 10^{-14}. In addition, although we have not been able to prove it, the numerical results suggests that the MCIS estimator might enjoy bounded relative error properties. The fifth and sixth columns of the table show the estimated *work normalized relative variance* (WNRV) used here as a performance measure that takes into account the simulation time. For an estimator $\widehat{\mathcal{Z}}$ computed in τ seconds, this performance measure is defined as

$$\mathrm{WNRV} = \frac{\tau \mathrm{Var}(\widehat{\mathcal{Z}})}{\mathcal{Z}^2}\ .$$

Using this measure the conclusions remain the same — the MCIS estimator is more efficient than the ISVE estimator for probabilities larger than about 10^{-14}.

Example 2 (Large portfolio losses modeled via Student's t copula) Suppose we have a portfolio of loans consisting of n^* obligors each of whom may default with probability $p_i = \mathbb{P}(X_i \geqslant x_i),\ i = 1, \ldots, n^*$, where each X_i is a continuous random variable describing the underlying financial liabilities of the i-th obligor; see [3,14]. Usually each X_i is not directly observable by the lender and is thus a latent random variable. The lender can only observe the default event $\{X_i \geqslant x_i\}$, where x_i is some critical level of financial liability beyond which the i-th obligor is bankrupt. The total loss incurred from the defaults is given by

$$L(\mathbf{X}) = \sum_{i=1}^{n^*} e_i \mathrm{I}\{X_i \geqslant x_i\}, \quad \mathbf{X} = (X_1, \ldots, X_{n^*})\ , \quad (26)$$

where each e_i represents the size of the loan to the i-th obligor. We wish to estimate the probability of a large loss, $\mathcal{Z} = \mathbb{P}(L(\mathbf{X}) \geqslant \gamma)$, where the random vector $\mathbf{X}$ is specified by the t copula model:

$$X_i = \Big(\varrho Z + \sqrt{1-\varrho}\,\eta_i\Big)\Big/\sqrt{V}, \quad i = 1, \ldots, n^*\ ,$$

with $Z \sim \mathsf{N}(0,1),\ \eta_1, \ldots, \eta_{n^*} \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2),\ 0 < \varrho < 1$, and $V \sim \mathsf{Gamma}(\nu/2, \nu/2)$. We denote random variables in upper case font ($\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{n^*})$) and their realizations in smaller case font ($\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{n^*})$). Note that each X_i has marginally a Student's t distribution with $\nu > 0$ degrees of freedom. Since $\mathbf{X}$ is a function of $\boldsymbol{\eta}, Z, V$, we can write $\mathcal{Z} = \mathbb{P}(S(\boldsymbol{\eta}, Z, V) \geqslant \gamma)$, where $S(\boldsymbol{\eta}, z, v) = L(\mathbf{x})$, and hence the zero-variance importance sampling density can be written as

$$\pi(\boldsymbol{\eta}, z, v) \propto$$

$$\exp\Big(-\frac{\|\boldsymbol{\eta}/\sigma\|^2 + z^2 + \nu\, v}{2}\Big) v^{\nu/2-1}\, \mathrm{I}\{S(\boldsymbol{\eta}, z, v) \geqslant \gamma\}\ .$$

We now estimate $\mathcal{Z}$ via the importance sampling estimator (25) with $w = 1$. Using the estimator (25) with say, $w = 0.5$, did not yield any dramatic improvement in efficiency for this example. To construct the semiparametric importance sampling density (19) we now derive the conditional densities of $\pi(\boldsymbol{\eta}, z, v)$.

The density $\pi(z \,|\, V, \boldsymbol{\eta})$.

Define $Q_k = x_k V^{1/2} - \sqrt{1-\varrho^2}\,\eta_k$ (note that x_k are fixed thresholds in (26) and not realizations of X_k) and let $(p_1, \ldots, p_{n^*})$ be the permutation corresponding to the order statistics: $Q_{p_1} < \cdots < Q_{p_{n^*}}$. In other words,

Table 2 Empirical performance of MCIS compared to importance sampling schemes with vanishing relative error properties.

		relative error %		
ϱ	MCIS est. $\widehat{\mathscr{Z}}$	MCIS	CEVE	ISVE
0	1.7950×10^{-5}	0.0092	0.0063	0.0069
0.4	1.8077×10^{-5}	0.093	0.17	0.23
0.7	1.9014×10^{-5}	0.04	0.65	2.85
0.9	2.0735×10^{-5}	0.068	0.63	2.80
0.93	2.0997×10^{-5}	0.17	3.5	4.59
0.95	2.1412×10^{-5}	0.11	6	8.09
0.99	2.1882×10^{-5}	0.29	15	4.35

Table 3 Empirical performance of MCIS and ISVE algorithms for various values of the threshold parameter $\gamma = 5 \times 10^{c+3}$, $c = 1, \ldots, 14$.

		relative error %		WNRV	
γ	ISVE estimate	MCIS	ISVE	MCIS	ISVE
5×10^{4}	3.9865×10^{-4}	0.049	1.6	25	1500
5×10^{5}	2.0802×10^{-5}	0.067	4.3	75	10000
5×10^{6}	6.4385×10^{-7}	0.077	5.5	64	22000
5×10^{7}	1.2039×10^{-8}	0.041	3.0	21	5000
5×10^{8}	1.3468×10^{-10}	0.034	6.1	12	20000
5×10^{9}	8.9791×10^{-13}	0.036	7.6	17	30000
5×10^{10}	3.5899×10^{-15}	0.043	0.014	27	0.11
5×10^{11}	8.5302×10^{-18}	0.079	0.029	81	0.50
5×10^{12}	1.2082×10^{-20}	0.025	0.024	7	0.32
5×10^{13}	1.0148×10^{-23}	0.042	0.00064	28	0.00021
5×10^{14}	5.0428×10^{-27}	0.042	0.00030	27	4.6×10^{-5}
5×10^{15}	1.4898×10^{-30}	0.024	0.00014	7	1.0×10^{-5}
5×10^{16}	2.5961×10^{-34}	0.023	3.87×10^{-15}	8.2	7.7×10^{-27}
5×10^{17}	2.6754×10^{-38}	0.012	4.22×10^{-13}	2.6	9.1×10^{-23}

$Q_{p_k} = Q_{(k)}$ for all k. Let $k = \min\{j : \gamma < \sum_{i=1}^{j} e_{p_i}\}$, then $L(\mathbf{X}) \geqslant \gamma$ if and only if $\varrho Z \geqslant Q_{(k)}$. It follows that the conditional density of Z given $\boldsymbol{\eta}$ and V is a truncated normal density:

$$\pi(z \,|\, V, \boldsymbol{\eta}) = \frac{\exp\big(-z^2/2\big)\, \mathrm{I}\{z \geqslant Q_{(k)}/\varrho\}}{\Phi(-Q_{(k)}/\varrho)} \,.$$

The density $\pi(v \,|\, Z, \boldsymbol{\eta})$.

Next define $E_k = \Big(\sqrt{1-\varrho^2}\,\eta_k + Z\Big)\Big/ x_k$ and let the vector $(p_1, \ldots, p_{n^*})$ be the permutation corresponding to the ordering: $E_{p_1} > \cdots > E_{p_{n^*}}$. Let $k = \min\{j : \gamma < \sum_{i=1}^{j} e_{p_i}\}$, then $L(\mathbf{X}) \geqslant \gamma$ if and only if $V < E_{p_k}^2$. Therefore, the conditional density of V given Z and $\boldsymbol{\eta}$ is a right-truncated gamma density:

$$\pi(v \,|\, Z, \boldsymbol{\eta}) = \frac{\exp(-\nu\, v/2)\, v^{\nu/2-1}\, \mathrm{I}\{V < E_{p_k}^2\}}{\mathsf{P}(\nu/2, E_{p_k}^2 \nu/2)} \,,$$

where $\mathsf{P}(\nu, x)$ is the incomplete gamma function [20, Page 716]. Note that efficient random variable generation from a right-truncated gamma distribution is accomplished via the accept-reject algorithm of Philippe [25].

The density $\pi(\boldsymbol{\eta} \,|\, Z, V)$.

Finally, the conditional density of $\boldsymbol{\eta}$ given Z and V is the multivariate truncated normal density:

$$\pi(\boldsymbol{\eta} \,|\, Z, V) \propto \exp\big(-\|\boldsymbol{\eta}/\sigma\|^2/2\big)\, \mathrm{I}\Big\{\sum_i e_i\, \mathrm{I}\{\eta_i \geqslant t_i\} \geqslant \gamma\Big\}$$

with $t_i \stackrel{\text{def}}{=} \frac{x_i\sqrt{V} - \varrho\, Z}{\sqrt{1-\varrho^2}}$, from whence the conditional density of η_i is straightforward to derive:

$$\pi(\eta_i \,|\, Z, V, \boldsymbol{\eta}_{-i}) \propto$$

$$\begin{cases} \exp\big(-\eta_i^2/2\sigma^2\big) & \text{if } \sum_{j\neq i} e_j\, \mathrm{I}\{\eta_j \geqslant t_j\} \geqslant \gamma \\ \exp\big(-\eta_i^2/2\sigma^2\big)\, \mathrm{I}\{\eta_i \geqslant t_i\} & \text{otherwise.} \end{cases}$$

In other words, the distribution of η_i given Z and V is either $\mathsf{N}(0, \sigma^2)$ or a truncated version of it.

For a numerical example we consider the same set of parameters used in Bassamboo et al. [3]: $\sigma^2 = 9$, $n^* = 250$, $\gamma = n^*/4$, $x_1 = \cdots = x_{n^*} = \sqrt{n^*}/2$, $e_1 = \cdots = e_{n^*} = 1$. Table 4 shows the performance of the (semi-parametric) MCIS method for various values of the correlation parameter ϱ and degrees of freedom ν. We used the estimator (25) with $n = 80$ and $m = 5 \times 10^5$. The table also includes the importance sampling schemes proposed and recommended in [3], namely the one based

Table 4 Estimated relative error of MCIS compared to exponential and hazard rate twisting. In the left panel the correlation parameter is $\varrho = 0.25$ and in the right panel the degree of freedom is $\nu = 12$.

ν	$\widehat{\mathcal{Z}}_1$	MCIS	ECM	HRT
8	2.39×10^{-4}	0.79	0.9	1.8
12	1.06×10^{-5}	1.05	1.7	2.6
16	6.11×10^{-7}	1.4	2.8	3.6
20	4.43×10^{-8}	1.9	3.7	5.4

ϱ	$\widehat{\mathcal{Z}}_1$	MCIS	ECM	HRT
0.1	8.536×10^{-6}	0.79	0.9	1.8
0.2	9.76×10^{-6}	0.98	1.2	2.3
0.3	1.18×10^{-5}	1.1	1.7	3.2
0.4	1.42×10^{-5}	1.1	3.1	4

on exponential change of measure (ECM) and hazard rate twisting (HRT). Both ECM and HRT estimators are the average of $m = 5 \times 10^5$ replications.

From the table we can see that the MCIS estimator yields smaller relative error than the exponential change of measure and the hazard rate twisting methods. Note that while the ECM and HRT methods are applicable only to the case where V has a light tailed distribution [3], the MCIS recipe remains unchanged regardless of the distribution of V.

Example 3 (A random walk hitting a non-convex set) Consider the problem of estimating the probability

$$\mathbb{P}(\{S_d > \gamma\} \cup \{S_d < -(\gamma + \varepsilon)\}), \quad \varepsilon, \gamma > 0 \tag{27}$$

where $S_d = (X_1 + \cdots + X_d)/d$ and $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$ for all i, independently. Standard importance sampling methods such as exponential or hazard rate twisting do not work in this case [19].

The purpose of this toy example is not only to show how the MCIS method can be used to tackle nonstandard rare-event settings, but, more importantly, to explore some of its limitations due to the effects of the curse of dimensionality. Note that while the methods proposed in [19] and more recently in [9] to estimate (27) explicitly exploit the decomposition of the rare-event into two disjoint events, the MCIS method has the much more difficult task of learning about this decomposition automatically and without any prior knowledge of it.

As in the previous examples, the first step is to derive the conditional densities of the optimal importance sampling pdf $\pi(\mathbf{x})$. Here the conditional densities are all mixtures of two truncated Gaussian densities ($i = 1, \ldots, d$):

$$\pi(x_i \,|\, \mathbf{X}_{-i}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right) \times \frac{\mathrm{I}\{x_i < a\} + \mathrm{I}\{x_i > b\}}{\Phi((a-\mu_i)/\sigma_i) + \Phi(-(b-\mu_i)/\sigma_i)},$$

where $a \stackrel{\text{def}}{=} -d(\gamma + \varepsilon + S_d) + X_i$, $b \stackrel{\text{def}}{=} d(\gamma - S_d) + X_i$, and $\Phi(\cdot)$ is the cdf of the $\mathsf{N}(0,1)$ distribution. Simulation from $\pi(x_i \,|\, \mathbf{X}_{-i})$ entails sampling from a mixture: generate a Bernoulli variable B with success probability $\Phi((a-\mu_i)/\sigma_i)/(\Phi((a-\mu_i)/\sigma_i) + \Phi(-(b-\mu_i)/\sigma_i))$. If $B = 1$, then generate X_i from a distribution $\mathsf{N}(\mu_i, \sigma_i^2)$, truncated to $(-\infty, a]$; otherwise generate X_i from a distribution $\mathsf{N}(\mu_i, \sigma_i^2)$, truncated to $[b, \infty)$. For simplicity here we generate the population $\mathbf{X}_1, \ldots, \mathbf{X}_n$ in (15) without any approximation error using the accept-reject method. In this way any influence of the mixing of the Gibbs Markov chain is eliminated from the ensuing analysis.

As a particular example consider the problem with parameters $\gamma = 1.3$, $\varepsilon = .01$, $d = 6$, and $\mu_i = 0$, $\sigma_i = 1$ for all i, which gives the exact probability of $\mathcal{Z} = 0.0013918....$

Note that the estimator (25) requires that we estimate both the marginal densities of all d components of $\mathbf{X}$ via the semi-parametric estimator (17), and the joint density of $\mathbf{X}$ via the nonparametric estimator (19). We now consider an alternative to (25):

$$\widetilde{\mathcal{Z}}_w = \frac{1}{m} \sum_{k=1}^{m} \frac{f(\mathbf{Y}_k)\mathrm{I}\{S(\mathbf{Y}_k) \geqslant \gamma\}}{w\,\widehat{g}(\mathbf{Y}_k) + (1-w)\widetilde{\pi}(\mathbf{Y}_k)}\,, \tag{28}$$

where $\widetilde{\pi}_w(\mathbf{y}) = w\widehat{g}(\mathbf{y}) + (1-w)\widetilde{\pi}(\mathbf{y})$, $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \overset{\text{iid}}{\sim} \widetilde{\pi}_w$, and $\widetilde{\pi}$ is an estimator of ($\pi_{i,j}$ denotes the marginal density of the vector (X_i, X_j))

$$\pi_{1,2}(x_1, x_2)\pi_{3,4}(x_3, x_4)\pi_{5,6}(x_5, x_6)$$

as opposed to $\pi(x_1, \ldots, x_6)$. In other words, $\widetilde{\pi}(\mathbf{y}) = \widehat{\pi}_{1,2}(y_1, y_2)\widehat{\pi}_{3,4}(y_3, y_4)\widehat{\pi}_{5,6}(y_5, y_6)$, where

$$\widehat{\pi}_{1,2}(y_1, y_2) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \pi(y_1 \,|\, X_{i,2}, \ldots, X_{i,6})\, \pi(y_2 \,|\, y_1, X_{i,3}, \ldots, X_{i,6})$$

and similarly for $\widehat{\pi}_{3,4}$ and $\widehat{\pi}_{5,6}$. Thus, instead of estimating the full joint density model $\pi(\mathbf{x})$ via (19), we consider estimating the joint densities of three blocks of $\mathbf{X}$, namely, $(X_1, X_2), (X_3, X_4)$, (X_5, X_6), and then take the product of these three densities. We can also consider other models such as

$$\widetilde{\pi}(\mathbf{y}) = \widehat{\pi}_{1,2,3}(y_1, y_2, y_3)\widehat{\pi}_{4,5,6}(y_4, y_5, y_6), \quad \text{where}$$

$$\widehat{\pi}_{1,2,3}(y_1, y_2, y_3) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \pi(y_1 \,|\, X_{i,2}, \ldots, X_{i,6})\, \pi(y_2 \,|\, y_1, X_{i,3}, \ldots, X_{i,6}) \times \pi(y_3 \,|\, y_1, y_2, X_{i,4}, X_{i,5}, X_{i,6})\,.$$

Table 5 shows the results of using estimator (28) with $m = 10^5$, $w = 0.01$, $n = 100$, and different models for $\widetilde{\pi}$. The second column of the table shows the proportion of wasted samples for which the rare-event $\{S(\mathbf{Y}) < \gamma\}$ has not occurred. In the table Model 1 corresponds to the case where $\widetilde{\pi} \equiv \widehat{\pi}$ and estimators (28) and (25) coincide. Model 4 corresponds to the case where $\widetilde{\pi} \equiv \widehat{g}$ and hence $\widetilde{\mathcal{Z}}_w = \widehat{\mathcal{Z}}_1$.

From the table we can see that for this example estimation of the joint density of $x_1, \ldots, x_6$ (Model 1) gives the smallest relative error. However, this does not imply that we should always aim to estimate the joint density of $\mathbf{X}$ (as opposed to any sub-blocks of $\mathbf{X}$), because we expect that estimating the full joint density $\pi(\mathbf{x})$ via (19) will become inefficient as the dimension of $\mathbf{x}$ increases. This is indeed confirmed by the numerical results given in Table 6, where $d = 12$. The best performance is achieved by Model 3 and not Model 1. Even though Model 1 generates 99% of the time samples for which $S(\mathbf{Y}) \geqslant \gamma$, the estimation of the full joint pdf $\widehat{\pi}(\mathbf{x})$ is not good enough and the simpler model of case 3 (which wastes 84% of the generated samples) yields a smaller relative error.

In conclusion, the numerical results suggest that it is better to have a good estimator of an approximation to π (like the product of marginals density $\prod_{i=1}^{d} \pi_i(x_i)$ or Model 3 in Table 6), than have a poor estimator of π itself.

Ultimately estimation of the optimal π is less efficient in high dimensions and we may need to devise simpler semi-parametric models such as models 2 through 6 given in the first column of Table 6. Of course, there is a vast number of possible semi-parametric models. For example, we may choose the model

$$\widetilde{\pi}(\mathbf{x}) = \widehat{\pi}_{1}(x_1)\, \widehat{\pi}_{2}(x_2)\, \widehat{\pi}_{3,4,12}(x_3, x_4, x_{12}) \times \widehat{\pi}_{5,6,7,8,9,10,11}(x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}) \,.$$

Which model proves to be good depends on the particular problem at hand. This freedom of choice is a mixed blessing, because it is not different from the infinite number of ways that one can devise an MCMC sampler.

Example 4 (Network reliability) Consider the static network reliability problem, in which we want to estimate the probability $\mathcal{Z}$ that the nodes 1 and 20 of the dodecahedron network on Figure 2 are *not* connected, given that each of the 30 edges (or links) fails independently with probability ε. This problem is considered by many authors; see [7,11,12,17] and the references therein.

More specifically, suppose each edge $e \in \{1, \ldots, d\}$ is assigned a random weight X_e and $X_1, \ldots, X_d$ are independent $\mathsf{N}(0, \sigma^2)$ random variables with $\sigma = -1/\Phi^{-1}(\varepsilon)$ (Φ^{-1} is the inverse of the cdf of the standard normal distribution). Note that $\varepsilon = \mathbb{P}(X_e > 1)$ for all edges and the event $\{X_e > 1\}$ is equivalent to the event that edge e has failed. Let $\mathscr{P} = \{\mathcal{P}_j\}$ denote the set of all paths connecting nodes 1 and 20 (here each $\mathcal{P}_j$ represents a sequence of edges connecting nodes 1 and 20). For example, Figure 2 shows three distinct paths connecting nodes 1 and 20 (the edges belonging to the paths are in bold). The failure probability $\mathcal{Z}$ can be expressed as

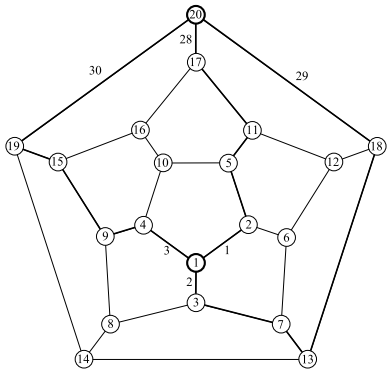

Fig. 2 A dodecahedron network with 20 nodes and 30 links. The nodes 1 and 20 are shown in bold circles. Links $1, 2, 3$ and $28, 29, 30$ are indicated.

$$\mathcal{Z} = \mathbb{P}_f(S(\mathbf{X}) > 1),$$

where $\mathbf{X} = (X_1, \ldots, X_d) \sim f$, the pdf f is the density of the multivariate normal distribution $\mathsf{N}(\mathbf{0}, \sigma^2 I)$, and S is defined via

$$S(\mathbf{X}) = \min_{\mathcal{P}_j \in \mathscr{P}} \max_{e \in \mathcal{P}_j} X_e \,. \tag{29}$$

When ε is small, say, smaller than 10^{-3}, the probability $\mathcal{Z}$ is typically a rare-event probability. To estimate this probability we now explain how to apply the MCIS method.

In the first (MC) stage, we use the adaptive splitting Algorithm A1 (see appendix) to generate a Markov chain sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with approximate density π.

In the second (IS) stage, we use the mixture (24) as the importance sampling density, where the transition density κ is given by (10). Here, the conditional pdfs of $\pi(\mathbf{x})$ are [7,8]:

$$\pi(x_e \,|\, \mathbf{X}_{-e}) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left(-x_e^2/(2\sigma^2)\right) & \text{if } S_e > 1 \\ \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left(-x_e^2/(2\sigma^2)\right) \dfrac{\mathrm{I}\{x_e > 1\}}{\varepsilon} & \text{otherwise} \,, \end{cases}$$

for $e = 1, \ldots, d$, where

$$\{S_e > 1\} \equiv \{S(X_1, \ldots, X_{e-1}, 0, X_{e+1}, \ldots, X_d) > 1\}$$

Table 5 Effect of the curse of dimensionality with $d = 6$ dimensions on the estimator $\widetilde{\mathcal{Z}}_w$. Here $\gamma = 1.3, \varepsilon = 0.01$ and $m = 10^5, n = 100, w = 0.01$.

Model	$\widetilde{\pi}(\mathbf{x}) =$	$\frac{1}{m}\sum_{k=1}^{m} \mathrm{I}\{S(\mathbf{Y}_k) < \gamma\}$	relative error %
1	$\widehat{\pi}(x_1, x_2, x_3, x_4, x_5, x_6)$	0.0099	1.62
2	$\widehat{\pi}_{1,2,3}(x_1, x_2, x_3)\widehat{\pi}_{4,5,6}(x_4, x_5, x_6)$	0.69	5.40
3	$\widehat{\pi}_{1,2}(x_1, x_2)\widehat{\pi}_{3,4}(x_3, x_4)\widehat{\pi}_{5,6}(x_5, x_6)$	0.84	2.21
4	$\prod_{i=1}^{6} \widehat{\pi}_i(x_i)$	0.95	3.63

Table 6 Effect of the curse of dimensionality with $d = 12$ dimensions on the estimator $\widetilde{\mathcal{Z}}_w$. Here $\gamma = 1.3, \varepsilon = 0.01, \mathcal{Z} \approx 6.184686 \times 10^{-6}$ and $m = 10^5, n = 10^3, w = 0.01$.

Model	$\widetilde{\pi} \equiv$	$\frac{1}{m}\sum_{k=1}^{m} \mathrm{I}\{S(\mathbf{Y}_k) < \gamma\}$	relative error %
1	$\widehat{\pi}$	0.099	4.00
2	$\widehat{\pi}_{1,2,3,4,5,6}\ \widehat{\pi}_{7,8,9,10,11,12}$	0.69	2.27
3	$\widehat{\pi}_{1,2,3,4}\ \widehat{\pi}_{5,6,7,8}\ \widehat{\pi}_{9,10,11,12}$	0.84	2.00
4	$\widehat{\pi}_{1,2,3}\ \widehat{\pi}_{4,5,6}\ \widehat{\pi}_{7,8,9}\ \widehat{\pi}_{10,11,12}$	0.92	2.69
5	$\widehat{\pi}_{1,2}\ \widehat{\pi}_{3,4}\ \widehat{\pi}_{5,6}\ \widehat{\pi}_{7,8}\ \widehat{\pi}_{9,10}\ \widehat{\pi}_{11,12}$	0.96	4.20
6	$\prod_{i=1}^{12} \widehat{\pi}_i$	0.99	11.76

Table 7 Empirical performance of the MCIS and merge process methods on the dodecahedron network.

		relative error %		WNRV	
ε	MCIS estimate	MCIS	merge	MCIS	merge
10^{-2}	2.03×10^{-6}	0.82	1.33	270	730
10^{-4}	2.03×10^{-12}	1.11	1.36	500	770
10^{-6}	2.01×10^{-18}	0.32	1.39	60	800
10^{-8}	2.00×10^{-24}	0.38	1.38	87	790
10^{-10}	1.97×10^{-30}	0.91	1.39	470	790
10^{-12}	2.01×10^{-36}	0.59	1.38	220	800
10^{-14}	2.01×10^{-42}	0.52	1.38	200	900
10^{-16}	2.73×10^{-48}	0.46	1.37	160	1000

is the event that the nodes are not connected given that edge e is forced to work (which is the same as $X_e < 1$). That is, if adding link e does not make the network operational, no change of measure is applied when sampling this link, otherwise the distribution is truncated to $[1, \infty)$. Note that in the importance sampling step it is sufficient to sample the Bernoulli events $\{X_e > 1\}$, instead of the random variables X_e.

Table 7 shows the MCIS performance for the failure probabilities $\varepsilon = 10^{-c}$, $c = 2, 4, \ldots, 16$ obtained with $w = 0.01$, $m = 10^4$, and $n = 10^3$. For comparison we also show the performance of the *merge process* algorithm for network reliability estimation. The merge process algorithm is specifically designed for network reliability estimation and is one of the most efficient and widely used algorithms; see [11,17]. The numerical results suggests that for this particular graph the MCIS performs well compared to the merge process algorithm.

4 Conclusions

In this paper we have presented a novel Monte method that combines importance sampling with Markov chain Monte Carlo methodology. An advantage of the MCIS method over other adaptive importance sampling methods is that our construction of the importance sampling density does not require the specification of a rigid parametric model and the concomitant estimation of any parameters via nonlinear optimization. The MCIS method provides a nonparametric density that directly aims to estimate the minimum variance importance sampling density. Unlike Chib's estimator, the MCIS estimator is unbiased and provides a straightforward empirical estimate of the relative error.

The numerical results show that for the problems we consider the proposed method is efficient. We show that for two light- and heavy-tailed rare-event problems, the proposed approach can handle successfully dependence specified by a Gaussian copula model. In particular, when the tail probability of sums of correlated log-normals is not too small, the MCIS method compares favorably to the currently recommended vanishing relative error estimator.

As future research we intend to examine the rare-event robustness properties of the MCIS estimator as the probability of the rare-event decreases to zero.

A Adaptive splitting sampler

Here we briefly present the MCMC sampling algorithm that is used to generate an approximate sample from the zero-variance density π in the rare-event setting in our examples. The main idea is to use the splitting method [6,7]. Suppose we are given an integer $s \geqslant 2$, called the splitting factor. Initially, we generate $n \times s$ independent states $\mathbf{X}$ from density f, and determine a threshold parameter γ_1 so that exactly n of them have $S(\mathbf{X}) \geqslant \gamma_1$. Then at each step t, for $t = 2, 3, \ldots$, we run for s steps a Markov chain with stationary density $f(\mathbf{x})\mathrm{I}\{S(\mathbf{x}) \geqslant \gamma_{t-1}\}/\mathbb{P}(S(\mathbf{x}) \geqslant \gamma_{t-1})$, from each of those n states $\mathbf{X}$ for which $S(\mathbf{X}) \geqslant \gamma_{t-1}$. We denote the transition kernel density of this Markov chain by $\tilde{\kappa}_{t-1}$. This gives another $n \times s$ states and we select a parameter γ_t so that exactly n of them have $S(\mathbf{X}) \geqslant \gamma_t$. This is done until $\gamma_t \geq 1$ for some t. This iterative procedure is summarized in the following algorithm.

Algorithm A1 (Adaptive splitting sampler) ***Require:*** *an integer* $s \geq 2$
$q \leftarrow n \times s - n$
$\mathscr{X}_1 \leftarrow \emptyset$
for $i = 1$ *to* $n \times s$ ***do***
    generate a vector $\mathbf{Y}$ *from density* f *and add it to* $\mathscr{X}_1$
sort the elements of $\mathscr{X}_1$ *by increasing order of* $S(\mathbf{X})$*, say* $\mathbf{X}_{(1)}, \ldots, \mathbf{X}_{(n\times s)}$
$\gamma_1 \leftarrow [S(\mathbf{X}_{(q)}) + S(\mathbf{X}_{(q+1)})]/2$
$t \leftarrow 1$
while $\gamma_t \leqslant 1$ ***do***
    $t \leftarrow t + 1$
    $\mathscr{X}_{t-1} \leftarrow \{\mathbf{X}_{(q)}, \ldots, \mathbf{X}_{(n\times s)}\}$ {*retain only the best* n *elements from* $\mathscr{X}_{t-1}$}
    $\mathscr{X}_t \leftarrow \emptyset$
    for all $\mathbf{X}_0 \in \mathscr{X}_{t-1}$ ***do***
        for $j = 1$ *to* s ***do***
            sample $\mathbf{X}_j$ *from the density* $\tilde{\kappa}_{t-1}(\cdot \mid \mathbf{X}_{j-1})$ *and add it to* $\mathscr{X}_t$
    sort the elements of $\mathscr{X}_t$ *by increasing order of* $S(\mathbf{X})$*, say* $\mathbf{X}_{(1)}, \ldots, \mathbf{X}_{(n\times s)}$
    $\gamma_t \leftarrow \min\{[S(\mathbf{X}_{(q)}) + S(\mathbf{X}_{(q+1)})]/2, 1\}$
return $\mathbf{X}_1 \ldots, \mathbf{X}_n$*, for which* $S(\mathbf{X}) \geqslant 1$*, as a sample with approximate density* $\pi(\mathbf{x}) = f(\mathbf{x})\mathrm{I}\{S(\mathbf{x}) \geqslant 1\}/\mathscr{Z}$

In this algorithm, $\mathscr{X}_t$ denotes a set of vectors $\mathbf{X}$ for which $S(\mathbf{X}) > \gamma_{t-1}$. When this set contains $n \times s$ elements, we sort it to retain the n vectors having the largest value of $S(\mathbf{X})$, and we remove the other vectors from this set. The threshold parameter γ_t is placed midway between the n-th and the $(n+1)$-th largest values of $S(\mathbf{X})$. In this paper we set $s = 2$ and select the transition density $\tilde{\kappa}_{t-1}$ to be the transition density of the corresponding systematic Gibbs sampler with stationary density $f(\mathbf{x})\mathrm{I}\{S(\mathbf{x}) \geqslant \gamma_{t-1}\}/\mathbb{P}(S(\mathbf{x}) \geqslant \gamma_{t-1})$.

References

1. Asmussen, S., Blanchet, J., Juneja, S., Rojas-Nandayapa, L.: Efficient simulation of tail probabilities of sums of correlated lognormals. Annals of Operations Research (2009). DOI: 10.1007/s10479-009-0658-5
2. Asmussen, S., Glynn, P.W.: Stochastic Simulation: Algorithms and Analysis. Springer-Verlag, New York (2007)
3. Bassamboo, A., Juneja, S., Zeevi, A.: Portfolio credit risk with extremal dependence: Asymptotic analysis and efficient simulation. Operations Research **56**(3), 593–606 (2008)
4. Besag, J.: Statistical analysis of non-lattice data. The Statistician **24**(3), 179195 (1975)
5. Blanchet, J.H., Glynn, P.W.: Efficient rare-event simulation for the maximum of heavy-tailed random walks. Annals of Applied Probability **18**, 1351–1378 (2008)
6. Botev, Z.I., Kroese, D.P.: Efficient Monte Carlo simulation via the generalized splitting method. Statistics and Computing (2010). DOI:10.1007/s11222-010-9201-4
7. Botev, Z.I., L'Ecuyer, P., Rubino, G., Simard, R., Tuffin, B.: Static network reliability estimation via generalized splitting. INFORMS Journal on Computing (2012). To appear
8. Botev, Z.I., L'Ecuyer, P., Tuffin, B.: Importance sampling method based on a one-step look-ahead density from a Markov chain. In: Proceedings of the 2011 Winter Simulation Conference, Phoenix, Arizona (2011)
9. Brereton, T.J., Chan, J.C.C., Kroese, D.P.: Fitting mixture importance sampling distributions via improved cross-entropy. In: Proceedings of the 2011 Winter Simulation Conference, Phoenix, Arizona (2011)
10. Bucklew, J.: Introduction to rare-event simulation. Springer, New York (2004)
11. Cancela, H., El Khadiri, M., Rubino, G.: Rare event analysis by Monte Carlo techniques in static models. In: G. Rubino, B. Tuffin (eds.) Rare Event Simulation Using Monte Carlo Methods, pp. 145–170. Wiley (2009). Chapter 7
12. Cancela, H., L'Ecuyer, P., Lee, M., Rubino, G., Tuffin, B.: Analysis and improvements of path-based methods for Monte Carlo reliability evaluation of static models. In: J. Faulin, A.A. Juan, S. Martorell, E. Ramirez-Marquez (eds.) Simulation Methods for Reliability and Availability of Complex Systems, pp. 65–84. Springer Verlag (2009)
13. Chan, J.C.C., Glynn, P.W., Kroese, D.P.: A comparison of cross-entropy and variance minimization strategies. Journal of Applied Probability (2011). Forthcoming
14. Chan, J.C.C., Kroese, D.P.: Improved cross-entropy method for estimation. manuscript (2011)
15. Chib, S.: Marginal likelihood from the Gibbs output. Journal of the American Statistical Association **90**(432), 1313–1321 (1995)
16. Dupuis, P., Leder, K., Wang, H.: Importance sampling for sums of random variables with regularly varying tails. ACM TOMACS **17**(3) (2006). Article 14
17. Gertsbakh, I.B., Shpungin, Y.: Models of Network Reliability. CRC Press, Boca Raton, FL (2010)
18. Glasserman, P.: Monte Carlo Methods in Financial Engineering. Springer-Verlag, New York (2004)
19. Glasserman, P., Wang, Y.: Counterexamples in importance sampling for large deviations probabilities. The Annals of Applied Probability **7**(3), 731–746 (1997)
20. Kroese, D.P., Taimre, T., Botev, Z.I.: Handbook of Monte Carlo methods. John Wiley & Sons, New York (2011)
21. L'Ecuyer, P., Blanchet, J.H., Tuffin, B., Glynn, P.W.: Asymptotic robustness of estimators in rare-event simulation. ACM Transactions on Modeling and Computer Simulation **20**(1), Article 6 (2010)
22. L'Ecuyer, P., Tuffin, B.: Approximate zero-variance simulation. In: Proceedings of the 2008 Winter Simulation Conference, pp. 170–181. IEEE Press (2008)
23. Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer-Verlag, New York (2001)
24. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equations of state calculations by fast computing machines. Journal of Chemical Physics **21**(6), 1087–1092 (1953)

25. Philippe, A.: Simulation of right and left truncated gamma distribution by mixtures. Statistics and Computing **7**, 173–181 (1997)
26. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods, second edn. Springer-Verlag, New York (2004)
27. Rubino, G., (editors), B.T.: Rare Event Simulation. John Wiley & Sons, New York (2009)
28. Rubinstein, R.Y., Kroese, D.P.: The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer-Verlag, New York (2004)
29. Rubinstein, R.Y., Kroese, D.P.: Simulation and the Monte Carlo Method, second edn. John Wiley & Sons, New York (2007)
30. Zhang, P.: Nonparametric importance sampling. Journal of the American Statistical Association **91**(435), 1245–1253 (1996)