Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Rare event simulation in finite-infinite dimensional space

Siu-Kui Au*, Edoardo Patelli

Institute for Risk and Uncertainty, University of Liverpool, United Kingdom

ARTICLE INFO

Article history:
Received 15 July 2015
Accepted 7 November 2015
Available online 27 November 2015

Keywords:
Curse of dimension
Rare Event
Markov Chain Monte Carlo
Monte Carlo
Subset Simulation

ABSTRACT

Modern engineering systems are becoming increasingly complex. Assessing their risk by simulation is intimately related to the efficient generation of rare failure events. Subset Simulation is an advanced Monte Carlo method for risk assessment and it has been applied in different disciplines. Pivotal to its success is the efficient generation of conditional failure samples, which is generally non-trivial. Conventionally an independent-component Markov Chain Monte Carlo (MCMC) algorithm is used, which is applicable to high dimensional problems (i.e., a large number of random variables) without suffering from 'curse of dimension'. Experience suggests that the algorithm may perform even better for high dimensional problems. Motivated by this, for any given problem we construct an equivalent problem where each random variable is represented by an arbitrary (hence possibly infinite) number of 'hidden' variables. We study analytically the limiting behavior of the algorithm as the number of hidden variables increases indefinitely. This leads to a new algorithm that is more generic and offers greater flexibility and control. It coincides with an algorithm recently suggested by independent researchers, where a joint Gaussian distribution is imposed between the current sample and the candidate. The present work provides theoretical reasoning and insights into the algorithm.

1. Introduction

Modern engineering systems are designed with increasing complexity and expectation of reliable performance. Rare failure events with high consequences are becoming more relevant to risk assessment and management. Unfortunately they are usually not well-understood and can even be out of imagination based on typical experience [1–3]. Studying failure scenarios allows one to gain insights into their cause and consequence, providing information for effective mitigation, contingency planning and improving system resilience. The probability and the consequence of failure events are two basic ingredients for trading off cost and benefit in the design of engineering systems. Assessing risk quantitatively requires proper modeling of the 'input' uncertain parameters by random variables as well as the logical/physical mechanism that predicts the 'output' quantities of interest. While no mathematical model is perfect, useful information can be gained if it is calibrated and interpreted properly, allowing one to make risk-informed decisions.

Let $\mathbf{X} = [X_1, \ldots, X_n]$ be the set of uncertain parameters in the problem, which are modeled by random variables. Without loss of generality $\{X_i\}_{i=1}^{n}$ are assumed to be standard Gaussian (zero mean and unit variance) and i.i.d. (independent and identically distributed). Dependent non-Gaussian random variables can be constructed from Gaussian ones by proper transformation [4]. One important problem in risk assessment is the determination of the failure probability $P(F)$ for a specified failure event F, which can be formulated as an n-dimensional integral or an expectation:

$$P(F) = \int I(\mathbf{x} \in F)\phi(\mathbf{x})d\mathbf{x} = E[I(\mathbf{X} \in F)] \tag{1}$$

where $I(\cdot)$ is the indicator function, equal to 1 if its argument is true and zero otherwise;

$$\phi(\mathbf{x}) = (2\pi)^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \qquad \mathbf{x} = [x_1, \ldots, x_n]^T \tag{2}$$

is the n-dimensional standard Gaussian PDF.

Monte Carlo methods [5–7] provide a robust means for risk assessment of complex systems. Problems of practical significance currently pose three main challenges: small probability, 'high dimension' (i.e., a large number of input random variables) and high complexity (e.g., nonlinearity) in the input–output relationship [8,9]. Small probability renders Monte Carlo method in its direct form computationally expensive or prohibitive. High dimension renders geometric intuitions in low dimensional space inapplicable or misleading [10,11]. High complexity means that

[*] Correspondence to :-Harrison Hughes Building, Brownlow Hill, Liverpool, L69 3GH, UK.

E-mail addresses: siukuiau@liverpool.ac.uk (S.-K. Au), Edoardo.patelli@liverpool.ac.uk (E. Patelli).

http://dx.doi.org/10.1016/j.ress.2015.11.012

the input–output relationship is only implicitly known as a 'black-box'.

1.1. Subset simulation

Advanced Monte Carlo methods generally aim at reducing the variance of estimators beyond direct Monte Carlo method but in doing so they lose application robustness. Subset Simulation is a method that is found to play a balance between efficiency and robustness [12–15]. It has been applied to different disciplines and used for developing algorithms for related problems such as sensitivity [16–18] and design optimization problems [19–24]. There are variants that take advantage of prior knowledge of the problem, e.g., casual dynamical systems [25], transition from linear to nonlinear failure [26], meta-model [27]; or leverage on other computational tools, e.g., delayed rejection [28], Kriging [29] and neural networks [30].

Subset Simulation is based on the idea that a small failure probability can be expressed as the product of larger conditional probabilities of intermediate failure events, thereby potentially converting a rare event simulation problem into a sequence of more frequent ones. A general failure event is represented as $F=\{Y>b\}$, where Y is a suitably defined 'driving response' characterizing failure. In the actual implementation, Subset Simulation produces estimates for the values of b that correspond to fixed failure probabilities, from large to small values. The estimates make use of samples that populate gradually from the frequent to rare failure regions, corresponding to increasing threshold values that are adaptively generated.

A typical Subset Simulation run starts with 'simulation level' 0, where N samples of $\mathbf{X}$ are generated according to the parameter PDF $\phi(\mathbf{x})$, i.e., direct Monte Carlo. The values of the response Y are then calculated and sorted. The p_0N+1 largest value is taken as the threshold level b_1 for simulation level 1, where p_0 is the 'level probability' chosen by the user (conventional choice is 0.1). The top p_0N samples of $\mathbf{X}$ are used as seeds for generating additional samples conditional on $Y>b_1$, to make up a population of N conditional samples at level 1. The p_0N+1 largest value of Y among these samples is taken as the threshold level b_2 for simulation level 2. Samples for level 2 are generated and the procedure is repeated for higher threshold levels until the level of interest is covered.

1.2. Generation of conditional samples

The efficient generation of conditional failure samples, i.e., samples that are conditional on intermediate failure events, is pivotal to Subset Simulation. This is conventionally performed using an independent-component Markov Chain Monte Carlo (MCMC) algorithm [12,31,7], which is applicable for high dimensional problems and makes the algorithm robust to applications. For each X_i, let $p_i^*(\cdot\,;\,\cdot)$ be the proposal PDF assumed to be symmetric, i.e., Metropolis random walk. Suppose we are given a sample $\mathbf{X}^{(1)}=[X_1^{(1)},\dots,X_n^{(1)}]$ distributed as the target conditional distribution, i.e.,

$$\phi(\mathbf{x}|F)=P(F)^{-1}I(x\in F)\phi(\mathbf{x}) \tag{3}$$

According to the algorithm the next sample $\mathbf{X}^{(2)}=[X_1^{(2)},\dots,X_n^{(2)}]$ that is also distributed as $\phi(\mathbf{x}|F)$ is generated as follow:

```
Algorithm I (independent-component MCMC)
Step I. Generate $\mathbf{X}'=\{X_i'\}_{i=1}^n$
For $i=1,\dots,n$
  1. Generate $\xi_i$ from the proposal PDF $p_i^*(\cdot\,;X_i^{(1)})$ and $U_i$ uniformly on [0, 1].
  2. Calculate $r_i=\phi(\xi_i)/\phi(X_i^{(1)})$.
  Set $X_i'=\xi_i$ if $U_i\le r_i$. Otherwise set $X_i'=X_i^{(1)}$.
End $i$
Step II (Check failure)
Set $\mathbf{X}^{(2)}=\mathbf{X}'$ if $\mathbf{X}'\in F$ (accept). Otherwise set $\mathbf{X}^{(2)}=\mathbf{X}^{(1)}$ (reject).
```

In the above, $\phi(x)=(2\pi)^{-1/2}\exp(-x^2/2)$ denotes the one-dimensional standard Gaussian PDF. The correlation among the conditional samples is an important factor influencing the efficiency of Subset Simulation. It is high (hence low efficiency) if $\mathbf{X}'$ is rejected too often in either Step I (MCMC mechanism) or Step II (not lying in the failure region); or when $\{\xi_i\}_{i=1}^n$ is of close proximity to $\mathbf{X}$ (governed by the proposal PDF).

1.3. Objectives and key findings

Theoretical arguments and numerical experience reveal that as the number of variables increases the rejection of the candidate $\mathbf{X}'$ tends to be governed by Step II; the efficiency of Subset Simulation is insensitive to the type of proposal PDF and may even be higher [12,15]. Motivated by this, for any given problem (generally finite dimensional) we consider an equivalent problem with an arbitrary number of random variables and investigate the limiting behavior of the algorithm as the number increases indefinitely. Specifically, each Gaussian variable X_i can be represented by an arbitrary (hence possibly infinite) number of 'hidden' Gaussian variables. As the key result of this work, we show that applying Algorithm I to the equivalent problem results in the following 'limiting algorithm' as the number of hidden variables is infinite:

```
Algorithm II (Limiting algorithm)
Step I. Generate $\mathbf{X}'=\{X_i'\}_{i=1}^n$
Generate $\mathbf{X}'=[X_1',\dots,X_n']$ as a Gaussian vector with independent components, with mean vector $[a_1X_n^{(1)},\dots,a_nX_n^{(1)}]$ and variances $[s_1^2,\dots,s_n^2]$.
Step II (Check failure)
Set $\mathbf{X}^{(2)}=\mathbf{X}'$ if $\mathbf{X}'\in F$ (accept). Otherwise set $\mathbf{X}^{(2)}=\mathbf{X}^{(1)}$ (reject).
```

Algorithm II differs from Algorithm I only in Step I. Here, $0\le s_i\le 1$ is the standard deviation of the candidate X_i' from the current sample and $a_i=\sqrt{1-s_i^2}$. It is related to the proposal PDF but which is no longer relevant because the algorithm is now controlled directly through $\{a_i\}_{i=1}^n$ or equivalently $\{s_i\}_{i=1}^n$. This algorithm is remarkably simple and MCMC rejection no longer appears explicitly. As the algorithm does not depend on any details of the hidden variables, the infinite-dimensional equivalent problem is only involved at a conceptual level to arrive at the limiting result.

The limiting algorithm shows that it is possible to generate the candidate in Step I simply as a Gaussian vector whose statistics depend on the current sample. In fact the same algorithm has been recently proposed by independent researchers [32] who ingeniously imposed this condition and verified this possibility. The present work provides a theoretical reasoning leading to the algorithm via a completely different route.

This paper is organized as follow. We first describe in Section 2 the equivalent problem with hidden variables that links the original problem and the conceptual infinite-dimensional problem. For ease of reading, the limiting behavior of the candidate and hence the MCMC algorithm is summarized in Section 3. Examples are then given in Section 4 to illustrate the results. The remaining sections provide the derivations for the limiting behavior and the results in Section 3.