Structural Safety 61 (2016) 67–77

Contents lists available at ScienceDirect

Structural Safety

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

Computation of failure probability via hierarchical clustering

Chao Yin *, Ahsan Kareem

Nathaz Modeling Laboratory, University of Notre Dame, United States

ARTICLE INFO

Article history:
Received 5 March 2014
Received in revised form 21 August 2015
Accepted 19 September 2015
Available online 2 May 2016

Keywords:
Reliability
Failure probability
Monte Carlo simulation
Subset simulation
Unsupervised learning
Hierarchical clustering

ABSTRACT

Most of the recent reliability analysis methods are based on supervised learning approaches, e.g. neural network, support vector machine, importance sampling (IS), Markov Chain Monte Carlo (MCMC), subset simulation (SS), etc. Among these, SS approach has been shown to outperform in effectiveness and robustness, as it classifies failure samples through multiple levels instead of conventional use of a single level. Inherent to supervised learning approaches is the limitation that these only utilize the evaluated information, i.e. the information from the samples whose performance functions have been computed. To overcome this, a new scheme based on modified hierarchical clustering is proposed. This scheme, referred to here as 'hierarchical failure clustering' (HFC), adds a pre-processing step via an unsupervised learning approach, e.g. various clustering algorithms. The proposed method exploits the statistical structure embedded in the input samples prior to computing the performance functions of these samples. This statistical structure has a tree based on these input samples with nodes as multi-level clusters. Father clusters at a specific level are used to explore an intermediate failure region and the failure portions are sifted out. Child clusters that are attached to the father failure clusters continue to explore a smaller intermediate failure region. Accordingly, the HFC method proceeds recursively until the target failure region is detected. In contrast with the SS and IS, HFC benefits from additional information from unlabeled samples with the help of a pre-built tree. Through a relatively small added premium in terms of computations for pre-processing, the additional information helps to improve the quality of the intermediate failure probability estimates and thereby significantly reduces the overall computational effort. The efficacy of HFC is theoretically demonstrated here and supported by several examples. HFC holds the promise for complex and large-scale systems for which the performance evaluation is computationally intensive.

1. Introduction

Reliability analysis aims to obtain the probability of failure of an event that is defined as

$$P(F) = \int_F p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \int_{\Omega_\Theta} I_F(\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{1}$$

where $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_d] \in \Omega_\Theta \subset \mathbb{R}^d$ represents an input random vector with probability density function (PDF) $p(\boldsymbol{\theta})$; F is the failure region defined on Ω_Θ, constrained by a limit state function $g(\boldsymbol{\theta})$, i.e. $F = \{\boldsymbol{\theta} \in \Omega_\Theta : g(\boldsymbol{\theta}) < 0\}$; I_F is an indicator function, that is $I_F(\boldsymbol{\theta}) = 1$ if $\boldsymbol{\theta} \in F$ and $I_F(\boldsymbol{\theta}) = 0$ otherwise. The estimation of small failure probability poses many challenges in the case of a large-scale engineering application, when each sample requires considerable computational effort. Indeed, if direct sampling-based methods are used, whose popularity is due to their robustness and straightforward implementation, a large number of simulations are needed to ensure the fidelity of the failure probability estimate, and this can be computationally demanding. On the other hand, when a computationally less intensive response surface scheme is adopted to approximate the limit state function using the first-order or second-order approximation FORM/SORM [31,10,4,23], the accuracy in the estimate of the failure probability can be insufficient, especially in the case of limit state functions with large curvatures or multiple design points [6]. Also, non-sampling-based integration methods, such as the sparse-grid quadrature [28,11], can fall short when the failure domain of input random variables is complicated.

In the case of sampling-based methods, one way of reducing their computational effort, especially in the case of small failure probabilities, is given by the use of variance reduction strategies with advanced sampling schemes, which have the aim of limiting the number of simulations without losing accuracy. These methods can be categorized as importance sampling-based methods [26,24], splitting and subset simulation (SS) methods [5,2], and

* Corresponding author.

http://dx.doi.org/10.1016/j.strusafe.2015.09.005
0167-4730/

methods that focus on improving the quality of the samples, such as quasi-Monte Carlo [9,22,18], common random numbers [12], antithetic variates [13,15].

Importance sampling (IS) aims to reduce the redundant samples by shifting the importance density towards the failure region [26,24,27]. The determination of importance density is crucial to IS. To achieve this goal, various IS techniques have been developed, which include (1) establishing the importance density via statistical moments estimation [3] or kernel density estimation (KDE) based on pre-samples [1]; (2) seeking the near-optimal IS density by minimizing the Kullback–Leibler cross entropy (CE) based on pre-samples [25,17]. Undoubtedly, a strategy is more effective when the number of pre-samples needed to identify the optimal importance density is low. However, this is not a trivial task. Uni-mode approximation tends to fail or lose effectiveness where multiple failure regions exist. Multi-mode approximations, such as the Gaussian mixture model (GMM) or KDE, face common concerns of how to determine the choice of number of kernels and how to determine the shape and weight of each kernel, which may cause overfitting or underfitting problems. This issue remains a challenge and is usually addressed manually with experience.

SS reduces the number of samples by narrowing down the failure region successively [2]. In SS, the failure probability is a product of conditional probabilities of multi-level intermediate failure events. Each intermediate failure event occurs more frequently than the target rare-event and thus requires fewer samples for evaluation. SS adopts Markov-Chain Monte Carlo (MCMC) [24] in the generation of new samples. However, MCMC still faces the problem of how to choose an appropriate proposal density function. In addition, the samples generated by MCMC are mutually dependent, and hence increase the coefficient of variation (c.o.v.).

A new simulation approach, termed 'hierarchical failure clustering' (HFC), is proposed here to compute the failure probabilities. This approach adopts the basic concept of SS by invoking hierarchical clustering methods [32,21]. This method does not need the importance density function or the proposal density function, therefore it is a good candidate for automation. HFC is straightforward to implement in the following steps: (1) a tree structure is first constructed via hierarchical clustering of the raw samples drawn from the input PDF; (2) higher-level father clusters are used to evaluate more frequent failure events, and the failure portions are sifted out as well as their child clusters, which are used to evaluate more rare events; (3) like SS, the target failure probability is expressed by a product of conditional probabilities of multi-level intermediate failure events. In fact, via the pre-built tree, HFC is able to exploit some useful information from the unlabeled samples, while SS and IS only utilize the information from the labeled samples. In this manner, the quality of the simulations can be enhanced since additional information is involved. Details of this approach are illustrated in the following sections.

2. Subset simulation

Consider the failure probability defined by Eq. (1). The integrand

$$p(\boldsymbol{\theta}|F) \propto I_F(\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{2}$$

represents the distribution of samples falling in the target failure region F. Obviously, if F is a small subset of $\Omega_{\boldsymbol{\Theta}}$, P_F should also be small, and hence a large number of simulations are required to ensure that a sufficient number of samples fall in this region. The subset simulation provides an effective way of reducing the number of simulations. Let $\{F_1 \supset F_2 \cdots \supset F_n = F\}$ be a decreasing sequence of failure events. By definition of conditional probability, the failure probability $P(F)$ is expressed as a product of a sequence of conditional probabilities $\{P(F_i|F_{i-1}):\ i = 2,3,\ldots,n\}$:

$$\begin{aligned} P(F) &= P\left(\bigcap_{i=1}^{n} F_i\right) \\ &= P(F_n|F_{n-1}) \cdot \cdots \cdot P(F_2|F_1)P(F_1) \\ &= P(F_1)\prod_{i=2}^{n} P(F_i|F_{i-1}) \end{aligned} \tag{3}$$

Thus, evaluating the target failure event amounts to evaluating this sequence of conditional failure events, which are given by

$$P(F_i|F_{i-1}) = \int_{\Omega_{\boldsymbol{\Theta}}} I_{F_i}(\boldsymbol{\theta})p(\boldsymbol{\theta}|F_{i-1})\mathrm{d}\boldsymbol{\theta} \tag{4}$$

with

$$p(\boldsymbol{\theta}|F_{i-1}) \propto I_{F_{i-1}}(\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{5}$$

Assuming there are N_{i-1} samples $\{\boldsymbol{\theta}_k,\ k = 1,2,\ldots,N_{i-1}\}$ distributed according to $p(\boldsymbol{\theta}|F_{i-1})$, i.e. $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}|F_{i-1})$, the conditional failure probability $P(F_i|F_{i-1})$ can be estimated by

$$P(F_i|F_{i-1}) \approx \widehat{P}(F_i|F_{i-1}) = \frac{1}{N_{i-1}}\sum_{k=1}^{N_{i-1}} I_{F_i}(\boldsymbol{\theta}_k) \tag{6}$$

Failure samples are generated successively from distributions $\{p(\boldsymbol{\theta}|F_{i-1}),\ i = 2,\ldots,n\}$ in which the number required at each intermediate level is smaller because each intermediate event $F_i|F_{i-1}$ occurs more frequently than the target failure event F. This concept is straightforward but not trivial to implement. For example, the number of samples decreases after sifting at each intermediate level and thus needs to be replenished to continue the subsequent computations. MCMC methods have been used to help SS achieve this goal [2]. As mentioned earlier, MCMC requires the choice of a proposal PDF which significantly affects the computational efficiency. For example, a global proposal PDF may not be efficient, while, for a local proposal PDF, samples may get trapped in a local region and have no opportunity to shift to other regions. In general, the performance of MCMC relies on the density estimation of the failure probabilities at every intermediate level. Another similar approach, referred to as the splitting method [5], uses IS instead of MCMC. A common limitation of SS and IS is that the input PDF reduces to a small number of independent samples and a significant amount of information is not exploited. The information gained after each level of evaluation is only from labeled samples. A possible improvement consists therefore in exploiting the additional information provided by the unlabeled samples. To this aim, it is here proposed to construct a tree-based data structure for raw samples drawn from the input PDF. The number of the raw samples is large but only a small portion needs to be evaluated. The unlabeled samples also contribute with some information to the analysis, albeit in an implicit manner. Obviously, methods that can utilize more information without additional computational cost have the potential to perform better. The central concept of the proposed approach is presented in the following section.

3. The method of hierarchical failure clustering

3.1. Basic concept

The proposed 'hierarchical failure clustering' (HFC) for computing failure probabilities embodies the basic idea detailed in the following. Let us suppose that the direct MC requires N samples to compute a target failure probability. With HFC, the same number of samples are used but reorganized as a tree structure in advance. This tree structure is built in such a manner that the most similar