BAYESIAN POST-PROCESSING FOR SUBSET SIMULATION FOR DECISION MAKING UNDER RISK

K. M. ZUEV[1] and J. L. BECK[2]

[1] *Department of Mathematics, University of Southern California, USA*
E-mail: kzuev@usc.edu
[2] *Division of Engineering and Applied Science, California Institute of Technology, USA*
E-mail: jimbeck@caltech.edu

Estimation of the failure probability, that is, the probability of unacceptable system performance, is an important and computationally challenging problem in reliability engineering. In cases of practical interest, the failure probability is given by an integral over a high-dimensional uncertain parameter space. Over the past decade, the engineering research community has realized the importance of advanced stochastic simulation methods for solving reliability problems. Subset Simulation, proposed by Au and Beck, provides an efficient algorithm for computing failure probabilities for general high-dimensional reliability problems. Here, a Bayesian post-processor for the original Subset Simulation method is presented that produces the posterior PDF of the failure probability which can be used in risk analyses for life-cycle cost analysis, decision making under risk, etc.

Keywords: Reliability Engineering, Stochastic Simulation Methods, Subset Simulation, Bayesian Approach, Uncertainty Quantification.

1. Introduction

One of the most important and challenging problems in reliability engineering is to estimate the failure probability p_F, that is, the probability of unacceptable system performance. This is usually expressed as an integral over a high-dimensional uncertain parameter space:

$$p_F = \int I_F(\theta)\pi(\theta)d\theta = \mathbb{E}_\pi[I_F(\theta)], \tag{1}$$

where $\theta \in \mathbb{R}^d$ represents the uncertain parameters needed to specify completely the excitation and dynamic model of the system; $\pi(\theta)$ is the joint probability density function (PDF) for θ; $F \subset \mathbb{R}^d$ is the failure domain in the parameter space (i.e. the set of parameters that lead to performance of the system that is considered to be unacceptable); and $I_F(\theta)$ stands for the indicator function, i.e. $I_F(\theta) = 1$ if $\theta \in F$ and $I_F(\theta) = 0$ if $\theta \notin F$.

Over the past decade, the engineering research community has realized the importance of advanced stochastic simulation methods for solving reliability problems because of the inefficiency of ordinary Monte Carlo simulation for highly reliable systems. As a result, many more efficient algorithms have been recently developed, e.g. Subset Simulation [1], Importance Sampling using Elementary Events [2], Line Sampling [9], Auxiliary domain method [8], Spherical Subset Simulation [7], Horseracing Simulation [10], to name but a few.

The usual interpretation of Monte Carlo methods is consistent with a purely *frequentist* approach, meaning that they can be interpreted in terms of the frequentist

Fifth Asian-Pacific Symposium on Structural Reliability and its Applications (5APSSRA).
Edited by Phoon, K. K., Beer, M., Quek, S. T. & Pang, S. D.

ISBN: 981-973-0000-00-0 :: doi: 10.3850/981-973-0000-00-0_BSS-paper-APSSRA12

definition of probability which identifies it with the long-run relative frequency of occurrence of an event. An alternative interpretation can be made based on the *Bayesian* approach which views probability as a measure of the plausibility of a proposition conditional on incomplete information that does not allow us to establish the truth or falsehood of the proposition with certainty [6,3]. Although the Bayesian approach usually leads to high-dimensional integrals that often cannot be evaluated analytically nor numerically by straightforward quadrature, the development of Markov chain Monte Carlo algorithms and increasing computing power have led, over the past few decades, to an explosive growth of Bayesian papers in all research disciplines.

In Section 3 of this paper, a Bayesian post-processor (SS+) for the original Subset Simulation method is developed. In SS+, the uncertain failure probability that one is estimating is modeled as a stochastic variable whose possible values belong to the unit interval. Instead of a single real number as an estimate, SS+ produces the posterior PDF of the failure probability, which takes into account both prior information and the information from the generated samples. This PDF can be used to give a point estimate such as the most probable value based on the available information or, alternatively, can be fully used in risk analyses for life-cycle cost analysis, decision making under risk, etc.

The rest of the paper is organized as follows. In Section 2, the original SS method is described; in Section 3 the Bayesian post-processor SS+ is developed and the relationship between SS and SS+ is discussed. Conclusions are given in Section 5.

2. Subset Simulation

The basic idea of Subset Simulation [1] is the following: represent a very small failure probability p_F as a product of larger probabilities so $p_F = \prod_{j=1}^{m} p_j$, where the factors p_j are estimated sequentially, $p_j \approx \hat{p}_j$, to obtain an estimate $\hat{p}_F^{SS}$ for p_F as $\hat{p}_F^{SS} = \prod_{j=1}^{m} \hat{p}_j$. To reach this goal, let us consider a decreasing sequence of nested subsets of the parameter space, starting from the entire space and shrinking to the failure domain F:

$$\mathbb{R}^d = F_0 \supset F_1 \supset \ldots \supset F_{m-1} \supset F_m = F. \tag{2}$$

Subsets $F_1, \ldots, F_{m-1}$ are called intermediate failure domains. As a result, the failure probability $p_F = P(F)$ can be rewritten in terms of conditional probabilities as follows:

$$p_F = \prod_{j=1}^{m} P(F_j|F_{j-1}) = \prod_{j=1}^{m} p_j, \tag{3}$$

where $p_j = P(F_j|F_{j-1})$ is the conditional probability at the $(j-1)^{th}$ conditional level. Clearly, by choosing the intermediate failure domains appropriately, all conditional probabilities p_j can be made large. Furthermore, they can be estimated, in principle, by the fraction of independent conditional samples that cause failure at the intermediate level:

$$p_j \approx \hat{p}_j^{MC} = \frac{1}{N} \sum_{i=1}^{N} I_{F_j}(\theta_{j-1}^{(i)}), \qquad \theta_{j-1}^{(i)} \sim \pi(\cdot|F_{j-1}). \tag{4}$$

Hence, the original problem (estimation of the small failure probability p_F) is replaced by a sequence of m intermediate problems (estimation of the larger failure probabilities p_j).

The first probability $p_1 = P(F_1|F_0) = P(F_1)$ is straightforward to estimate by Monte Carlo Simulation (MCS), since (4) requires sampling from $\pi(\cdot)$ that is assumed to be readily sampled. However, if $j \geq 2$, to estimate p_j using (4) one needs to generate independent samples from conditional distribution $\pi(\cdot|F_{j-1})$, which, in general, is not a trivial task. It is not efficient to use MCS for this purpose, especially at higher levels, but it can be done by a specifically tailored Markov chain Monte Carlo technique at the expense of generating dependent samples. In Subset Simulation, the Modified Metropolis algorithm (MMA) [1] is used for sampling from the conditional distributions $\pi(\cdot|F_{j-1})$ for $j \geq 2$. For more details, please see the original paper [1].

3. Bayesian Post-Processor for Subset Simulation

In this section we develop a Bayesian post-processor for the original Subset Simulation algorithm described in Section 2.

Recall that in SS the failure probability p_F is represented as a product of conditional probabilities $p_j = P(F_j|F_{j-1})$, each of which is estimated using (4). Let n_j denote the number of samples $\theta_{j-1}^{(1)}, \ldots, \theta_{j-1}^{(N)}$ that belong to subset F_j. Then

$$\hat{p}_j = \frac{n_j}{N} \quad \text{and} \quad \hat{p}_F^{SS} = \prod_{j=1}^{m} \hat{p}_j = \prod_{j=1}^{m} \frac{n_j}{N}. \tag{5}$$

Note that estimates (5) are purely frequentist. In fact, they are the maximum likelihood estimates for a binomial distribution for the number of failure events ($\theta_{j-1} \in F_j$), given $\theta_{j-1} \in F_{j-1}$. In order to construct a Bayesian post-processor for SS, we replace the maximum likelihood estimates in (5) by their Bayesian analogs. In other words, we treat all $p_1, \ldots, p_m$ and p_F as *stochastic variables* and, following the Bayesian approach, proceed as follows:

(1) Specify prior PDFs $p(p_j)$ for all $p_j = P(F_j|F_{j-1})$;
(2) Update each prior PDF, using new data $\mathcal{D}_{j-1} = \{\theta_{j-1}^{(1)}, \ldots, \theta_{j-1}^{(N)} \sim \pi(\cdot|F_{j-1})\}$, i.e. find the posterior PDFs $p(p_j|\mathcal{D}_{j-1})$ via Bayes' theorem;
(3) Obtain the posterior PDF $p(p_F| \cup_{j=0}^{m-1} \mathcal{D}_j)$ of $p_F = \prod_{j=1}^{m} p_j$ from $p(p_1|\mathcal{D}_0), \ldots, p(p_m|\mathcal{D}_{m-1})$.

To choose the prior distribution for each p_j, we use the *Principle of Maximum Entropy* (PME), introduced by Jaynes [5]. The PME postulates that, subject to specified constraints, the prior PDF p which should be taken to represent the current state of knowledge is the one that gives the largest measure of uncertainty, i.e. maximizes Shannon's entropy $H(p) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx$. Since the set of all possible values for each stochastic variable p_j is the unit interval, we impose this as the only constraint for $p(p_j)$, i.e. supp $p(p_j) = [0, 1]$. It is well known that the uniform distribution is the maximum entropy distribution among all continuous distributions on $[0, 1]$, so

$$p(p_j) = 1, \quad 0 \leq p_j \leq 1. \tag{6}$$

Since initial samples $\theta_0^{(1)}, \ldots, \theta_0^{(N)}$ are i.i.d. according to π, the sequence of zeros and ones, $I_{F_1}(\theta_0^{(1)}), \ldots, I_{F_1}(\theta_0^{(N)})$, can be considered as Bernoulli trials and, therefore, the likelihood function $p(\mathcal{D}_0|p_1)$ is a binomial distribution where $\mathcal{D}_0$ consists of the

number of F_1-failure samples $n_1 = \sum_{k=1}^{N} I_{F_1}(\theta_0^{(k)})$ and the total number of samples is N. Hence, the posterior distribution of p_1 is the beta distribution $\mathcal{B}e(n_1+1, N-n_1+1)$ with parameters (n_1+1) and $(N-n_1+1)$, i.e.

$$p(p_1|\mathcal{D}_0) = \frac{p_1^{n_1}(1-p_1)^{N-n_1}}{\mathrm{B}(n_1+1, N-n_1+1)}, \tag{7}$$

The beta function B in (7) is a normalizing constant. If $j \geq 2$, all MCMC samples $\theta_{j-1}^{(1)}, \ldots, \theta_{j-1}^{(N)}$ are distributed according to $\pi(\cdot|F_{j-1})$, however, they are not independent. Nevertheless, we can use an expression similar to (7) as a good approximation for the posterior PDF $p(p_j|\mathcal{D}_{j-1})$ for $j \geq 2$, so:

$$p(p_j|\mathcal{D}_{j-1}) \approx \frac{p_j^{n_j}(1-p_j)^{N-n_j}}{\mathrm{B}(n_j+1, N-n_j+1)}, \quad j \geq 1, \tag{8}$$

where $n_j = \sum_{k=1}^{N} I_{F_j}(\theta_{j-1}^{(k)})$ is the number of F_j-failure samples. Note that the MCMC samples $\theta_{j-1}^{(1)}, \ldots, \theta_{j-1}^{(N)}$ consist of the states of multiple Markov chains with different initial seeds obtained from previous conditional levels. This makes the approximation (8) more accurate in comparison with the case of a single chain.

The last step is to find the PDF of the product of stochastic variables $p_F = \prod_{j=1}^{m} p_j$, given the distributions of all factors p_j by (8).

In general, the product of beta variables does not follow the beta distribution, nevertheless, it can be accurately approximated by a beta variable.

Theorem 3.1 (Fan [4]). *Let $X_1, \ldots, X_m$ be independent beta variables, $X_j \sim \mathcal{B}e(a_j, b_j)$, and $Y = X_1 X_2 \ldots X_m$. Then Y is approximately distributed as $\tilde{Y} \sim \mathcal{B}e(a, b)$, where*

$$a = \mu_1 \frac{\mu_1 - \mu_2}{\mu_2 - \mu_1^2}, \quad b = (1-\mu_1)\frac{\mu_1 - \mu_2}{\mu_2 - \mu_1^2}, \tag{9}$$

and

$$\mu_1 = \mathbb{E}[Y] = \prod_{j=1}^{m} \frac{a_j}{a_j + b_j}, \quad \mu_2 = \mathbb{E}[Y^2] = \prod_{j=1}^{m} \frac{a_j(a_j+1)}{(a_j+b_j)(a_j+b_j+1)}. \tag{10}$$

It is easy to check that if $\tilde{Y} \sim \mathcal{B}e(a, b)$ with a and b given by (9), then the first two moments of stochastic variables Y and $\tilde{Y}$ coincide, i.e. $\mathbb{E}[\tilde{Y}] = \mathbb{E}[Y]$ and $\mathbb{E}[\tilde{Y}^2] = \mathbb{E}[Y^2]$. The accuracy of approximation $Y \dot{\sim} \mathcal{B}e(a, b)$ is discussed by Fan [4].

Using this theorem, we can therefore approximate the posterior distribution p^{SS+} of stochastic variable p_F by the beta distribution as follows:

$$p^{SS+}(p_F| \cup_{j=0}^{m-1} \mathcal{D}_j) \approx \tilde{p}^{SS+}(p_F| \cup_{j=0}^{m-1} \mathcal{D}_j) = \mathcal{B}e(p_F|a, b), \tag{11}$$

i.e. $p_F \dot{\sim} \mathcal{B}e(a, b)$, where

$$a = \frac{\prod_{j=1}^{m} \frac{n_j+1}{N+2}\left(1 - \prod_{j=1}^{m} \frac{n_j+2}{N+3}\right)}{\prod_{j=1}^{m} \frac{n_j+2}{N+3} - \prod_{j=1}^{m} \frac{n_j+1}{N+2}}, \quad b = \frac{\left(1 - \prod_{j=1}^{m} \frac{n_j+1}{N+2}\right)\left(1 - \prod_{j=1}^{m} \frac{n_j+2}{N+3}\right)}{\prod_{j=1}^{m} \frac{n_j+2}{N+3} - \prod_{j=1}^{m} \frac{n_j+1}{N+2}}. \tag{12}$$

Since the first two moments of p^{SS+} and $\tilde{p}^{SS+}$ are equal, we have:

$$\mathbb{E}_{\tilde{p}^{SS+}}[p_F] = \mathbb{E}_{p^{SS+}}[p_F] = \prod_{j=1}^{m} \mathbb{E}_{p(p_j|\mathcal{D}_{j-1})}[p_j] = \prod_{j=1}^{m} \frac{n_j+1}{N+2}, \tag{13}$$

and therefore

$$\lim_{N\to\infty} \mathbb{E}_{\tilde{p}^{SS+}}[p_F] = \lim_{N\to\infty} \hat{p}_F^{SS} \quad \text{so} \quad \mathbb{E}_{\tilde{p}^{SS+}}[p_F] \approx \hat{p}_F^{SS}, \text{ when } N \text{ is large.} \tag{14}$$

Therefore, the mean value of the approximation $\tilde{p}^{SS+}$ to the posterior PDF p^{SS+} is accurately approximated by the frequentist point estimate $\hat{p}_F^{SS}$ of the failure probability.

From the algorithmic point of view, SS+ differs from SS only in the produced output. Instead of a single real number as an estimate as in SS, SS+ produces $\tilde{p}^{SS+}(p_F|\cup_{j=0}^{m-1}\mathcal{D}_j)$, an approximation of the posterior PDF of p_F, which takes into account both prior information and the sampled data $\cup_{j=0}^{m-1}\mathcal{D}_j$. The frequentist estimate $\hat{p}_F^{SS}$ of failure probability obtained in the original SS algorithm is accurately approximated by the mean of $\tilde{p}^{SS+}$ as in (14).

Note that one can use the full PDF $\tilde{p}^{SS+}$ for risk analyses, including life-cycle cost analysis and decision making under risk. For instance, a performance loss function $\mathcal{L}$ often depends on the failure probability. In this case one can calculate an expected loss given by the following integral:

$$\mathbb{E}[\mathcal{L}(p_F)] = \int_0^1 \mathcal{L}(p_F)\tilde{p}^{SS+}(p_F)dp_F, \tag{15}$$

which takes into account the uncertainty in the value of the failure probability.

4. Illustrative Example

As an example, consider a linear failure domain. Let $d = 10^3$ be the dimension of the linear problem and suppose $p_F = 10^{-3}$ is the exact failure probability. The failure domain F is defined as

$$F = \{\theta \in \mathbb{R}^d : \langle \theta, e\rangle \geq \beta\}, \tag{16}$$

where e is a unit vector drawn from a uniform distribution over the surface of a unit sphere, and $\beta = \Phi^{-1}(1-p_F) \approx 3.09$ is the reliability index. This example is one where FORM gives the exact failure probability in terms of β. Note that $\theta^* = \beta e$ is the design point of the failure domain F. In the application of Subset Simulation, two implementation scenarios are considered: $N = 250$ and $N = 1000$ samples are simulated at each conditional level. The failure probability estimates $\hat{p}_F^{SS}$ obtained by SS for these scenarios and the approximation of the corresponding posterior PDFs $\tilde{p}^{SS+}$ obtained by SS+ are given in Fig. 1. Observe that the more samples used (i.e. the more information about the system that is extracted), the more narrowly $\tilde{p}^{SS+}$ is focused around $\hat{p}_F^{SS}$, as expected. The coefficients of variation are $\delta_{\tilde{p}^{SS+}} = 0.32$ and $\delta_{\tilde{p}^{SS+}} = 0.16$ for $N = 250$ and $N = 1000$, respectively. In SS+, the COV $\delta_{\tilde{p}^{SS+}}$ can be considered as a measure of uncertainty, based on the generated samples.

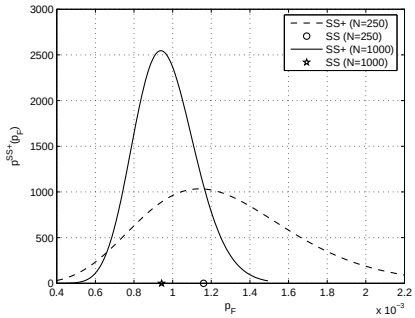

Fig. 1. The failure probability point estimates $\hat{p}_F^{SS}$ obtained by SS and $\tilde{p}^{SS+}$ obtained by SS+.

5. Conclusions

In this paper a Bayesian post-processor (SS+) for the original Subset Simulation method [1] is developed. In SS+, the uncertain failure probability that one is estimating is modeled as a stochastic variable whose possible values belong to the unit interval. Instead of a single real number as an estimate, SS+ produces an accurate approximation $\tilde{p}^{SS+}$ of the posterior PDF of the failure probability, which takes into account both prior information and the sampled data. This PDF can be fully used in risk analyses, including life-cycle cost analysis and decision making under risk.

Acknowledgments

This work was supported by the National Science Foundation, under award number EAR-0941374. This support is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation.

References

1. S.K. Au and J.L. Beck, "Estimation of small failure probabilities in high dimensions by subset simulation," *Probabilistic Engineering Mechanics*, 16(4), p. 263-277, 2001.
2. S.K. Au and J.L. Beck, "First-excursion probabilities for linear systems by very efficient importance sampling," *Probabilistic Engineering Mechanics*, 16(3), p. 193-207, 2001.
3. J.L. Beck, "Bayesian system identification based on probability logic," *Structural Control and Health Monitoring*, 17, p. 825-847, 2010.
4. D.-Y. Fan, "The distribution of the product of independent beta variables," *Communications in Statistics - Theory and Methods*, 20(12), p. 4043-4052, 1991.
5. E.T. Jaynes, "Information Theory and Statistical Mechanics", *Phys. Rev.*, 106, p. 620-630, 1957.
6. E.T. Jaynes, *Probabiity Theory: The Logic of Science*, Cambridge University Press, 2003.
7. L.S. Katafygiotis and S.H. Cheung, "Application of spherical subset simulation method and auxiliary domain method on a benchmark reliability study," *Str. Safety*, 29, p. 194-207, 2007.
8. L.S. Katafygiotis, T. Moan and S.H. Cheung, "Auxiliary domain method for solving multi-objective dynamic reliability problems for nonlinear structures," *Structural Engineering and Mechanics*, 25(3), p. 347-363, 2007.
9. P. Koutsourelakis, H.J. Pradlwarter, and G.I Schuëller, "Reliability of structures in high dimensions, part I: algorithms and applications," *Probabilistic Engineering Mechanics*, 19(4), p. 409-417, 2004.
10. K.M. Zuev and L.S. Katafygiotis, "Horseracing Simulation algorithm for evaluation of small failure probabilities," *Probabilistic Engineering Mechanics*, 26(2), p. 157-164, 2011.