ESC: An efficient Error-based Stopping Criterion for Kriging-based reliability analysis methods

Zeyu Wang*, Abdollah Shafieezadeh**
The Ohio State University, Columbus, OH, 43210, United States

Abstract

The ever-increasing complexity of numerical models and associated computational demands have challenged classical reliability analysis methods. Surrogate model-based reliability analysis techniques, and in particular those using Kriging meta-model, have gained considerable attention recently for their ability to achieve high accuracy and computational efficiency. However, existing stopping criteria, which are used to terminate the training of surrogate models, do not directly relate to the error in estimated failure probabilities. This limitation can lead to high computational demands because of unnecessary calls to costly performance functions (e.g. involving Finite Element models) or potentially inaccurate estimates of failure probability due to premature termination of the training process. Here we propose the Error-based Stopping Criterion (*ESC*) to address these limitations. First, it is shown that the total number of wrong sign estimation of the performance function for candidate design samples by Kriging, *S*, follows a Poisson binomial distribution. This finding is subsequently used to estimate the lower and upper bounds of *S* for a given confidence level for sets of candidate design samples classified by Kriging as safe and unsafe. An upper bound of error of the estimated failure probability is subsequently derived according to the probabilistic properties of Poisson binomial distribution. The proposed upper bound is implemented in the Kriging-based reliability analysis method as the stopping criterion. The efficiency and robustness of *ESC* are investigated here using five benchmark reliability analysis problems. Results indicate that the proposed method achieves the set accuracy target and substantially reduces the computational demand, in some cases by over 50%.

Key words: *Reliability analysis; Surrogate model; Stopping Criterion; Adaptive Kriging; Estimation error; Poisson binomial distribution;*

1. Introduction

Reliability analysis is primarily concerned with the vulnerability of components or systems under the impact of factors such as deterioration, normal operation loads, or extreme disturbances. In this context, the failure probability, P_f, is commonly defined as:

$$P_f = P(g(\boldsymbol{x}) \leq 0) = \int_{g(\boldsymbol{x})\leq 0} \rho(\boldsymbol{x}) d\boldsymbol{x} \quad (1)$$

where $\boldsymbol{x}$ is the vector of random variables, $g(\boldsymbol{x})$ is the so-called limit state function or performance function and $\rho(\boldsymbol{x})$ is the joint probability density function (PDF) of $\boldsymbol{x}$. Multiple groups of techniques have been developed for estimating P_f in Eq. (1). These include, but are not limited to, sampling approaches (e.g., Crude Monte-Carlo Simulation [1], [2], Importance Sampling(IS) [3] , Subset Simulation (SS) [4], etc.), optimization approaches (e.g., First or Second Order Reliability Method (FORM & SORM) [5], [6]) and the state-of-the-art surrogate modeling methods (e.g., Response Surface [7], [8], [9], Polynomial Chaos Expansion [10], Support Vector Regression [11], [12], or Kriging [13]–[15]).

Recently, Kriging-based surrogate modeling approaches have gained considerable attention. This is in part because of the ability of Kriging to estimate responses in the form of a distribution with Kriging mean and variance, as opposed to other surrogate models, which provide only the best estimate of the responses. This feature enables adaptive and efficient refinement of the surrogate model in the vicinity of the limit state, $g(\boldsymbol{x}) = 0$. Reviews of Kriging-based reliability analysis algorithms and their advantages compared to other surrogate model-based approaches can be found in [13], [16], [17]. Two representative algorithms called Efficient Global Reliability Analysis (EGRA) proposed by Bichon et al. [18] and Adaptive Kriging with MCS (AK-MCS) proposed by Echard et al. [13] are widely used. Compared to EGRA, AK-MCS uses the coefficient of variation of failure probability COV_{P_f} to evaluate the sufficiency of the considered number of candidate design samples.

A number of recent studies have further improved the performance of Kriging-based reliability analysis. These improvements target sampling strategies, Kriging learning functions and stopping criteria. For sampling strategies, Echard et al. [3], Balesdent et al. [19] and Dubourg et al. [20] used importance sampling techniques alongside the adaptive Kriging model, which facilitates reliability analysis for rare events. Additionally, Subset Simulation techniques are used with Kriging-based reliability analysis in [4],[21] [22]. Wen et al. [23] and Yang et al. [24] proposed the truncated candidate samples region, which cuts off candidate samples with small values of probability density. It is shown that using this approach, the number of calls to performance function can be reduced and acceptable accuracy in failure probability estimates can be achieved [23]. With regard to learning functions, Bichon et al. [15] proposed the expected feasible function (*EFF*), which prioritizes points with high uncertainty and those in the proximity of the limit state, $g(\boldsymbol{x}) = 0$. Echard et al. [13] proposed a learning function called *U* for quantification of the likelihood of wrong sign estimation. This function is widely used in other Kriging-based methodologies [3], [4], [25]. An information entropy-based learning function *H*, which follows a similar strategy to *EFF*, is developed by Lv et al. [26]. Sun et al. [27] proposed the Least Improvement Function (*LIF*), which enhances the learning process by searching next best training points among those that have high probability of wrong sign estimation, are in the vicinity of the limit state, and have high probability density. Xiao et al. [28] combined three learning functions $\boldsymbol{\psi}_d, \boldsymbol{\psi}_\sigma$ and $\boldsymbol{\psi}_m$ for measuring the distance among the limit state, new training points and candidate design samples with high variances to pick the next training point, while avoiding 'clustering' phenomenon. For stopping criteria, most of existing studies set thresholds on values of learning functions. Bichon et al. [18] and Wen et al. [23] adopted the maximum *EFF* smaller than a prescribed threshold (e.g., $max(EFF) \leq 10^{-3}$) as the stopping criterion. On the other hand, the stopping criterion $min(U) \geq 2$ is widely used in many adaptive Kriging-based reliability analysis algorithms [3], [4], [13], [25], [29]. Kriging-based stopping criteria have also been developed and implemented in the field of optimization to achieve sufficient accuracy and avoid unnecessary training of surrogate models [30], [31]. All of these approaches show great merits and improve the performance of Kriging-based reliability analysis. However, there is no systematic approach to establish stopping criteria that can guarantee convergence to the true failure probability and avoid calls to the performance function when the estimated failure probability is close to the true one estimated by MCS.

It is shown that the unnecessary training of the surrogate model in adaptive Kriging-based methods exists for both stopping criteria $max(EFF) \leq 10^{-3}$ and $min(U) \geq 2$ [13], [14], [32], [33]. Gaspar et al. [32] proposed a new stopping criterion based on the stability of the estimated failure probability. Fauriat et al. [14] states that the accuracy of the Kriging model will be sufficiently high if 2% of the candidate design samples satisfy $min(U) \geq 2$. Different from those approaches, Hu et al. [29] derived an estimate of the maximum error for the estimated failure probability assuming that the signs of the limit state function for points satisfying $U \geq 2$ are all correctly estimated and that the true number of failure points satisfying $U < 2$ is no more than the number of points satisfying $U < 2$. However, these assumptions may not be acceptable since there is a possibility of wrong sign estimation even for points that satisfy $U \geq 2$. For example, if there are thousand samples that satisfy $U = 2$ and are categorized into failure domain, the expected number of true failure points is equal to $\overline{N}_f^{U=2} = 1000 - 1000 \times \Phi(-2) = 977$, where $\Phi(\cdot)$ is the cumulative density function (CDF) of the standard normal distribution. The maximum error proposed in [29] fails to capture this crucial point. The lack of direct correspondence of existing stopping criteria with error in estimated failure probabilities can also potentially lead to inaccurate reliability analysis results due to premature termination of the training process.

In this paper, a maximum error $\hat{\epsilon}_{max}$ for the estimated failure probability is derived and an Error-based Stopping Criterion (*ESC*) is proposed. First, it is shown that the number of candidate design samples wrongly assigned to safe and failure domains follows a Poisson binomial distribution. The probabilistic properties of these distributions are subsequently derived. Based on these information, a procedure to estimate the maximum error in the estimated failure probabilities for a given confidence level is proposed. Finally, an error-based stopping criterion is presented by setting a threshold value for

the presented maximum error. The proposed stopping criterion, *ESC*, solves the unnecessary training problem for Kriging-based reliability analyses methods and guarantees convergence to target accuracy levels. Five reliability analysis examples are investigated to show the advantages offered by the proposed *ESC*, especially for high-dimensional problems.

This paper is organized in four sections. The elements of Kriging-based reliability analysis are briefly introduced in Section 2. In Section 3, the derivation of the proposed maximum error $\hat{\boldsymbol{\epsilon}}_{max}$ and *ESC* stopping criterion are presented. In Section 4, five examples are investigated to demonstrate the application of $\hat{\boldsymbol{\epsilon}}_{max}$ and *ESC* in solving reliability problems. Section 5 summarizes and concludes this study.

2. Kriging-Based Reliability Analysis

2.1 Kriging elements

Kriging, also called Gaussian Process Regression, is an interpolation-based regression method [34]. It assumes that the estimated responses $K(\boldsymbol{x})$ for unknown observations and the true responses for known observations $\boldsymbol{Y}(\boldsymbol{x})$ follow a joint Gaussian distribution [34], [35]. In this section, a brief overview of Kriging is presented. More information about this method can be found in [3], [4], [13], [25]. The Kriging model $K(\boldsymbol{x})$ can be described as follows:

$$K(\boldsymbol{x}) = F(\boldsymbol{\beta}, \boldsymbol{x}) + Z(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{f}(\boldsymbol{x}) + Z(\boldsymbol{x}) \tag{2}$$

where $F(\boldsymbol{\beta}, \boldsymbol{x})$ is the deterministic regression part representing the Kriging trend and $Z(\boldsymbol{x})$ is the stochastic interpolation part with Gaussian distribution assumption. Expanding $F(\boldsymbol{\beta}, \boldsymbol{x})$, $\boldsymbol{f}(\boldsymbol{x})$ is the Kriging basis and $\boldsymbol{\beta}$ is the regression coefficient of $\boldsymbol{f}(\boldsymbol{x})$. $\boldsymbol{\beta}^T f(\boldsymbol{x})$ often takes ordinary (β_0), linear ($\beta_0 + \sum_{i=1}^{N} \beta_i x_i$) or quadratic ($\beta_0 + \sum_{i=1}^{N} \beta_i x_i + \sum_{i=1}^{N} \sum_{j=i}^{N} \beta_{ij} x_i x_j$) forms, where N is the dimension of the random input vector $\boldsymbol{x}$. In this paper, we use the ordinary Kriging model. Moreover, $Z(\boldsymbol{x})$ follows a stationary normal Gaussian process with zero mean and covariance matrix as shown below:

$$\mathrm{COV}\left(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)\right) = \sigma^2 R(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}) \tag{3}$$

where σ^2 is the process variance from the regression part (e.g., generalized mean square error), $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are two observations, and $R(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$ is the correlation function or the so-called kernel function, which represents the correlation function of the process with hyper-parameter $\boldsymbol{\theta}$. Several forms have been used for the correlation function in Kriging approach; these include linear, exponential, Gaussian, and Matérn functions, among others. In this paper, the Gaussian kernel function is used with the following form:

$$R(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}) = \prod_{k=1}^{N} \exp\left(-\theta^k \left(x_i^k - x_j^k\right)^2\right) \tag{4}$$

where N is the dimension of the random input vector. The hyper-parameter $\boldsymbol{\theta}$ can be estimated via Maximum Likelihood Estimation (MLE) or Cross-Validation (CV) [34]. It is shown that $\boldsymbol{\theta}$ has a significant impact on the performance of Kriging [16], [23], [36]. To keep the consistency with previous studies for comparison purposes, here, an optimization toolbox called DACE [37], [38] that uses MLE is used to search for optimal θ^k in $(0,10)$. The Maximum Likelihood Estimation can be represented as:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \left(\left| \boldsymbol{R}(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}) \right|^{\frac{1}{m}} \sigma^2 \right) \tag{5}$$

Accordingly, the regression coefficient $\boldsymbol{\beta}$, and Kriging estimated mean and variance can be determined as follows [34]:

$$\boldsymbol{\beta} = (\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{Y}$$
$$\mu_K(\boldsymbol{x}) = \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{r}^T(\boldsymbol{x})\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta})$$
$$\sigma_K^2(\boldsymbol{x}) = \sigma^2(1 - \boldsymbol{r}^T(\boldsymbol{x})\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x}) + (\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}))^T(\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}(\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}))) \quad (6)$$

where $\boldsymbol{F}$ is the matrix of basis function $\boldsymbol{f}(\boldsymbol{x})$ evaluated at known training points, i.e. $F_{ij} = f_j(\boldsymbol{x}_i)$, $i = 1, 2, \dots, m$; $j = 1,2, \dots, p$, $\boldsymbol{r}(\boldsymbol{x})$ is the vector of correlation between known training points $\boldsymbol{x}_i$ and an unknown point $\boldsymbol{x}$: $r_i = \boldsymbol{R}(\boldsymbol{x}, \boldsymbol{x}_i, \boldsymbol{\theta})$, $i = 1,2 \dots m$, and $\boldsymbol{R}$ is the autocorrelation matrix for known training points: $R_{ij} = \boldsymbol{R}(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{\theta})$, $i = 1,2, \dots, m; j = 1,2, \dots, m$. Due to the prior assumption of Kriging model, the responses from Kriging follow a normal distribution with Kriging mean $\mu_K(\boldsymbol{x})$ and Kriging variance $\sigma_K^2(\boldsymbol{x})$:

$$K(\boldsymbol{x}) \sim N\left(\mu_K(\boldsymbol{x}), \sigma_K^2(\boldsymbol{x})\right) \quad (7)$$

Compared with the points that are further away from the training points, responses of points close to the training points are expected to have higher confidence. The Kriging-based reliability analyses apply 'learning functions' that use information of Kriging mean and variance to strategically pick points from the candidate design samples. Two learning functions, i.e. the *EFF* and *U* are briefly introduced in the following subsection.

2.2 Learning function & stopping criterion

Learning functions play an important role in Kriging-based reliability analysis. Learning functions facilitate searching for points in the set of candidate design samples that lead to highest gains for failure probability estimation. As stated in the introduction of this paper, multiple learning functions have been proposed. Two popular learning functions including Expected Learning Function (EFF) and U function, are considered in this paper. In *EFF*, the proximity of points to the limit state $g(\boldsymbol{x}) = a$ and their variance are the two key factors. The mathematical expression of EFF is presented below:

$$EFF(\boldsymbol{x}) = \int_{a-\delta(\boldsymbol{x})}^{a+\delta(\boldsymbol{x})} [\delta(\boldsymbol{x}) - |a - h|]\phi\big(h; \mu_K(\boldsymbol{x}), \sigma_K(\boldsymbol{x})\big)dh$$
$$= (\mu_K(\boldsymbol{x}) - a)\left[2\Phi\left(\frac{a - \mu_K(\boldsymbol{x})}{\sigma_K(\boldsymbol{x})}\right) - \Phi\left(\frac{a^- - \mu_K(\boldsymbol{x})}{\sigma_K(\boldsymbol{x})}\right) - \Phi\left(\frac{a^+ - \mu_K(\boldsymbol{x})}{\sigma_K(\boldsymbol{x})}\right)\right]$$
$$-\sigma_K(\boldsymbol{x})\left[2\phi\left(\frac{a - \mu_K(\boldsymbol{x})}{\sigma_K(\boldsymbol{x})}\right) - \phi\left(\frac{a^- - \mu_K(\boldsymbol{x})}{\sigma_K(\boldsymbol{x})}\right) - \phi\left(\frac{a^+ - \mu_K(\boldsymbol{x})}{\sigma_K(\boldsymbol{x})}\right)\right]$$
$$+2\sigma_K(\boldsymbol{x})\left[\Phi\left(\frac{a^+ - \mu_K(\boldsymbol{x})}{\sigma_K(\boldsymbol{x})}\right) - \Phi\left(\frac{a^- - \mu_K(\boldsymbol{x})}{\sigma_K(\boldsymbol{x})}\right)\right] \quad (8)$$

where $\phi(\cdot)$ is the standard normal probability density function. Here, $a = 0$, $\delta(\boldsymbol{x}) = 2\sigma_K(\boldsymbol{x})$, $a^+ = a + \delta(\boldsymbol{x})$ and $a^- = a - \delta(\boldsymbol{x})$. The term $[\delta(\boldsymbol{x}) - |a - h|]$ in Eq. (8) measures the proximity of the target point, and is weighted by the term $\phi(h; \mu_K(\boldsymbol{x}), \sigma_K(\boldsymbol{x}))$. The point that maximizes the EFF response is chosen as the next point to refine the Kriging model. The conventional stopping criterion based on this learning function is expressed as $max\big(EFF(\boldsymbol{x})\big) \leq 10^{-3}$.

Another widely accepted learning function is *U* learning function, which represents the uncertainties in the sign ($\pm$) estimation by $\hat{g}(\boldsymbol{x})$=0. This learning function is also investigated in this paper. *U* takes the probabilistic distribution of estimated responses into consideration, and quantifies the probability of making a wrong sign estimation for $\hat{g}(\boldsymbol{x})$. The formulation of U is:

$$U(\boldsymbol{x}) = \frac{|\mu_K(\boldsymbol{x})|}{\sigma_K(\boldsymbol{x})} \tag{9}$$

The point that minimizes the response of *U* learning function is selected as the next training point. The conventional stopping criterion based on U is defined as $min\big(U(\boldsymbol{x})\big) \geq 2$, which is interpreted as that the probability of making wrong sign estimation should not exceed 0.023. It is shown that both *EFF* and *U* learning functions are efficient for the selection of appropriate points in the adaptive Kriging process [13]. Moreover, *EFF* tends to converge faster than *U* in achieving true probability of failure P_f, while *U* learning function converges faster to its own stopping criterion ($min\big(U(\boldsymbol{x})\big) \geq 2$) than *EFF* ($max\big(EFF(\boldsymbol{x})\big) \leq 10^{-3}$). As EFF and U do not directly correspond to the error in the failure probability that is estimated at each iteration of the adaptive Kriging process, the corresponding stopping criteria are set to be overly conservative to ensure that the true error in the estimates of failure probability is confidently acceptable for reliability problems. The lack of direct correspondence to the error, therefore, leads to a large number of required costly simulations. Thus, it is necessary to derive a stopping criterion that is based on the accuracy of the estimates of the failure probability. To address this challenge, we present a maximum error for estimated failure probabilities within the adaptive Kriging reliability analysis process. Furthermore, based on this accuracy measure, a new stopping criterion called *ESC* is proposed. The performance of this measure is compared with ($min\big(U(\boldsymbol{x})\big) \geq 2$) and $\big(max\big(EFF(\boldsymbol{x})\big) \leq 10^{-3}\big)$ strategies.

3. ESC: An Error-Based Stopping Criterion

Since the true failure probability, denoted as P_f, is unavailable, the failure probability estimated via Monte Carlo Simulations, P_f^{MCS}, can be regarded as the benchmark for measuring the accuracy, assuming that sufficiently large number of simulations are used in the calculation of P_f^{MCS}. Thus, the estimated failure probability via Kriging-based reliability analysis with MCS can be denoted as $\hat{P}_f^{MCS}$. The relative error of the $\hat{P}_f^{MCS}$ with respect to P_f^{MCS} can be defined as:

$$\epsilon = \frac{\left|\hat{P}_f^{MCS} - P_f^{MCS}\right|}{P_f^{MCS}} = \left|\frac{\hat{P}_f^{MCS}}{P_f^{MCS}} - 1\right| \tag{10}$$

For consistency in the comparison, the set of candidate design samples for crude MCS and Kriging-based MCS should be the same. This way, a reliability analysis algorithm Ψ is optimal, if it requires the least number of calls to performance function, N_{call}, while maintaining the error below a prescribed threshold $\epsilon_{thr} \in (0,1)$:

$$\Psi^* = \underset{\Psi \in \pi,\ \epsilon \leq \epsilon_{thr}}{\arg\min}\ N_{call} \tag{11}$$

where π is the set of all feasible reliability analysis algorithms. As the true error ϵ for MCS-based reliability approaches is not known, our objective here is to mathematically derive the maximum error $\hat{\epsilon}_{max}$ for a given confidence level by leveraging the statistical information available via Kriging. Subsequently, a stopping criterion named *ESC* is proposed to demonstrate the ability of this approach in solving the unnecessary training problem.

3.1 Derivation of the maximum error

Let us define the true and estimated probabilities of failure as:

$$P_f^{MCS} = \frac{N_f}{N_{MCS}} \tag{12}$$

$$\hat{P}_f^{MCS} = \frac{\widehat{N}_f}{N_{MCS}} \tag{13}$$

where N_{MCS} is the number of candidate design samples for MCS, N_f is the number of failures determined using the true performance function $g(\boldsymbol{x})$, and $\widehat{N}_f$ denotes the number of points from the candidate design samples Ω determined by the Kriging model to indicate failure. The true domains of failure and survival within Ω are denoted as Ω_f and Ω_s, respectively, while the Kriging-estimated failure and survival domains are represented by $\widehat{\Omega}_f$ and $\widehat{\Omega}_s$, respectively. Therefore, $\widehat{N}_f$ is the size of $\widehat{\Omega}_f$. Subsequently, the relative error in Eq. (10) can be rewritten as,

$$\epsilon = \left|\frac{\hat{P}_f^{MCS}}{P_f^{MCS}} - 1\right| = \left|\frac{\widehat{N}_f}{N_f} - 1\right| \tag{14}$$

In Kriging surrogate models, the true error ϵ is not known, since N_f is unknown. Let us denote the total number of candidate design points in $\widehat{\Omega}_f$ that belong to Ω_s as $\hat{S}_f$ and those in $\widehat{\Omega}_s$ that belong to Ω_f as $\hat{S}_s$. Thus, N_f can be determined as:

$$N_f = \widehat{N}_f + \hat{S}_s - \hat{S}_f \tag{15}$$

The probabilistic distribution of $\hat{S}_s$and $\hat{S}_f$ are explored in the next subsection. With confidence level α, a confidence interval of N_f can be represented as:

$$N_f \in \left[\widehat{N}_f - \hat{S}_f^u, \quad \widehat{N}_f + \hat{S}_s^u\right] \tag{16}$$

where $\hat{S}_f^u$ and $\hat{S}_s^u$ are the upper bound of the confidence interval of $\hat{S}_s$ and $\hat{S}_f$, respectively. Accordingly, the maximum error can be calculated as,

$$\epsilon = \left|\frac{\widehat{N}_f}{N_f} - 1\right| \leq max\left(\left|\frac{\widehat{N}_f}{\widehat{N}_f - \hat{S}_f^u} - 1\right|, \left|\frac{\widehat{N}_f}{\widehat{N}_f + \hat{S}_s^u} - 1\right|\right) = \hat{\epsilon}_{max} \tag{17}$$

where $\hat{S}_f^u$and $\hat{S}_s^u$ are unknown. In the following section, it is shown that both $\hat{S}_s$ and $\hat{S}_f$ follow a Poisson binomial distribution, and the confidence intervals can be determined accordingly.

3.2 Probability distribution of $\widehat{S}_s$ and $\hat{S}_f$

For each candidate design point $\boldsymbol{x}_i$ in $\widehat{\Omega}_f$ or $\widehat{\Omega}_s$, let's define an indicator function I_i that takes one when $K(\boldsymbol{x}_i)$ makes a wrong estimation of the sign of $g(\boldsymbol{x}_i)$, and zero when the sign estimation is correct. Subsequently, $\hat{S}_f = \sum_{i=1}^{\widehat{N}_f} I_i\,, \boldsymbol{x}_i \in \widehat{\Omega}_f$ and $\hat{S}_s = \sum_{i=1}^{\widehat{N}_s} I_i\,, \boldsymbol{x}_i \in \widehat{\Omega}_s$, where $\widehat{N}_s$ denotes the number of points from the candidate design samples Ω that are determined by the Kriging model to be safe. Since the output of the Kriging model follows a normal distribution with mean $\hat{y}_{\boldsymbol{x}_i}$ and standard deviation $\hat{\sigma}_{\boldsymbol{x}_i}$, the probability of the event that the sign estimate of $\boldsymbol{x}_i$ is wrong is [13]:

$$P\left(I_i = 1 \mid \boldsymbol{x}_i \in \widehat{\Omega}_f \cup \widehat{\Omega}_s\right) = P_i^{wse} = \Phi\left(-\left|\frac{\hat{y}_{\boldsymbol{x}_i}}{\hat{\sigma}_{\boldsymbol{x}_i}}\right|\right) \tag{18}$$

where P_i^{wse} denotes the probability of wrong sign estimation for $\boldsymbol{x}_i$. This derivation is illustrated in Fig. 1. Thus, it is evident that I_i follows a Bernoulli distribution with the following mean and variance:

$$\mathrm{E}\left[I_i \middle| \boldsymbol{x}_i \in \widehat{\Omega}_f \cup \widehat{\Omega}_s\right] = P_i^{wse} \tag{19}$$

$$\mathrm{Var}\left[I_i \middle| \boldsymbol{x}_i \in \widehat{\Omega}_f \cup \widehat{\Omega}_s\right] = P_i^{wse}(1 - P_i^{wse}) \tag{20}$$

It can be shown that the sum of independent Bernoulli trials follows a Poisson binomial distribution [39]. Consequently, $\hat{S}_f$ and $\hat{S}_s$ follow Poisson binomial distributions as follows:

$$\hat{S}_s \sim PB\left(\mu_{\hat{S}_s}, \sigma_{\hat{S}_s}^2\right), x_i \in \widehat{\Omega}_s, \mu_{\hat{S}_s} = \sum_{i=1}^{\widehat{N}_s} P_i^{wse}\,, \sigma_{\hat{S}_s}^2 = \sum_{i=1}^{\widehat{N}_s} P_i^{wse}\,(1 - P_i^{wse}) \tag{21}$$

$$\hat{S}_f \sim PB\left(\mu_{\hat{S}_f}, \sigma_{\hat{S}_f}^2\right), x_i \in \widehat{\Omega}_f, \mu_{\hat{S}_f} = \sum_{i=1}^{\widehat{N}_f} P_i^{wse}\,, \sigma_{\hat{S}_f}^2 = \sum_{i=1}^{\widehat{N}_f} P_i^{wse}\,(1 - P_i^{wse}) \tag{22}$$

Fig. 1. An illustration of the probability of wrong sign estimation in Eq.(18) considering (a) $\hat{y}_{x_i} \geq 0$ or (b) $\hat{y}_{x_i} \leq 0$.

Based on these distributions, the confidence intervals (CIs) of $\hat{S}_s$ and $\hat{S}_f$ are given by:

$$\hat{S}_s \in \left(\boldsymbol{\Theta}_{\hat{S}_s}^{-1}\left(\frac{\alpha}{2}\right), \boldsymbol{\Theta}_{\hat{S}_s}^{-1}\left(1-\frac{\alpha}{2}\right)\right) \tag{23}$$

$$\hat{S}_f \in \left(\boldsymbol{\Theta}_{\hat{S}_f}^{-1}\left(\frac{\alpha}{2}\right), \boldsymbol{\Theta}_{\hat{S}_f}^{-1}\left(1-\frac{\alpha}{2}\right)\right) \tag{24}$$

where $\boldsymbol{\Theta}_{\hat{S}_s}^{-1}(\cdot)$ and $\boldsymbol{\Theta}_{\hat{S}_f}^{-1}(\cdot)$ are the inverse CDF of the Poisson binomial distribution of $\hat{S}_s$ and $\hat{S}_f$, respectively, and α is the confidence level. Analytical solutions for the above confidence intervals are typically not available, instead, numerical approaches or approximate analytical methods can be pursued. Sampling techniques can be used to numerically determine the inverse CDF of Poisson binomial distributions. However, in most cases, the confidence interval of $\hat{S}_s$ can be approximately determined using the Central Limit theorem. In cases where the probability of failure is small, $\hat{S}_f$ can be approximately obtained using Poisson distribution. These two cases are presented below in the form of corollaries:

Corollary 1. *$\hat{S}_s$ in distribution converges to a normal distribution for sufficiently large $\widehat{N}_s$, and the confidence interval of $\hat{S}_s$ can be obtained accordingly.*

Proof: Given that

$$\lim_{\widehat{N}_s \to \infty}\left(\max_{i=1,\dots,\widehat{N}_s} \frac{Var[I_i]}{Var[\hat{S}_s]}\right) = 0, \quad x_i \in \widehat{\Omega}_s \tag{25}$$

Lindeberg's condition for the Central Limit theorem for the sum of independent, not identically distributed random variables is satisfied [40]. Subsequently, for sufficiently large $\widehat{N}_s$, $\hat{S}_s$ in distribution converges to a normal distribution:

$$\hat{S}_s \sim N\left(\mu_{\hat{S}_s}, \sigma_{\hat{S}_s}^2\right), \qquad x_i \in \widehat{\Omega}_s \tag{26}$$

The CI of $\hat{S}_s$ can then be obtained as:

$$\hat{S}_s \in \left[\mu_{\hat{S}_s} - \gamma_{ci}\sigma_{\hat{S}_s},\ \mu_{\hat{S}_s} + \gamma_{ci}\sigma_{\hat{S}_s}\right], \quad x_i \in \widehat{\Omega}_s \tag{27}$$

where $\gamma_{ci} = 1.96$ for the confidence level $\alpha = 0.05$. As $\widehat{N}_s$ is large in Kriging-based reliability analysis problems, the above confidence bounds for $\hat{S}_s$ are accurate. It should be noted that because the event of failure is often rare, $\widehat{N}_f$ is not sufficiently large for the distribution of $\hat{S}_f$ to converge to a normal distribution and therefore to use corollary 1.

Corollary 2. *The distribution of $\hat{S}_f$ can be approximately represented using a Poisson distribution and the confidence interval of $\hat{S}_f$ can be obtained accordingly.*

Proof: It is shown that $\hat{S}_f$ follows a Poisson binomial distribution. According to Le cam's theorem [39], [41]:

$$\sum_{k=0}^{\infty}\left|\Pr\left(\hat{S}_f=k\right)-\frac{\mu_{\hat{S}_f}^{k}e^{-\mu_{\hat{S}_f}}}{k!}\right|<2\sum_{i=1}^{\hat{N}_f}(P_i^{wse})^2 \tag{28}$$

which indicates that the distribution of $\hat{S}_f$ can be approximately represented as a Poisson distribution:

$$\Pr\left(\hat{S}_f=k\right)\approx\frac{\mu_{\hat{S}_f}^{k}e^{-\mu_{\hat{S}_f}}}{k!},k=0,1,\dots,\hat{N}_f \tag{29}$$

The CI of $\hat{S}_f$ can be determined as:

$$\hat{S}_f\in\left[\boldsymbol{\Gamma}_{\hat{S}_f}^{-1}\left(\frac{\alpha}{2}\right),\boldsymbol{\Gamma}_{\hat{S}_f}^{-1}\left(1-\frac{\alpha}{2}\right)\right] \tag{30}$$

where $\boldsymbol{\Gamma}_{\hat{S}_f}^{-1}(\cdot)$ is the inverse CDF of the Poisson distribution with both mean and variance equal to $\mu_{\hat{S}_f}$ defined in Eq. (22), and α is the confidence level. It is recommended that $\hat{S}_s$ be estimated as normal distribution since $\hat{N}_s$ is large in most cases and $\hat{S}_f$ be estimated as Poisson distribution due to the fact $\hat{N}_f$ is relatively small.

3.3 ESC: Error-based Stopping Criterion

In adaptive procedures of reliability analysis, the stopping criterion is very important. However, it is shown that traditional stopping criteria do not have direct correspondence with the extent of error, thus they are set to be very strict to ensure that the estimation error is acceptable. To address this limitation, *ESC* includes an upper bound for the error, $\hat{\epsilon}_{max}$, determined using Eq. (17) to ensure that the extent of error in failure probability estimation does not exceed a prescribed error threshold ϵ_{thr}. The updating process of the Kriging model stops when the following condition is reached:

$$\hat{\epsilon}_{max}\leq\epsilon_{thr} \tag{31}$$

Accordingly, it is expected that the true error (denoted as ϵ), with the confidence level α, should be smaller than $\hat{\epsilon}_{max}$. The relationship between ϵ, $\hat{\epsilon}_{max}$, and ϵ_{thr} can be presented as:

$$\epsilon\leq\hat{\epsilon}_{max}\leq\epsilon_{thr} \tag{32}$$

3.4 Implementation of ESC

The procedure to implement *ESC* for reliability analysis is illustrated in the flowchart presented in Fig. 2. The primary steps for the implementation of the proposed method are described below:

- **Step 1**: *Generation of initial candidate design samples.* Generate N_{MCS} candidate design samples using Latin Hypercube Sampling (*LHS*). These samples are denoted as *S*.

- **Step 2**: *Selection of initial training points*. Randomly select from *S* an initial set of training points denoted as $\boldsymbol{x}_{tr}$ for Kriging construction, and evaluate their responses $g(\boldsymbol{x}_{tr})$.

- **Step 3**: *Kriging construction.* Construct the Kriging model using $\boldsymbol{x}_{tr}$. Denote the Kriging model as $K(\boldsymbol{x})$. For the construction of the Kriging model, MATLAB toolbox DACE [37] is used here with ordinary Kriging basis and Gaussian correlation function.

- **Step 4:** *Kriging estimation.* Obtain the current Kriging responses including the mean $\mu_K(\boldsymbol{x})$ and variance $\sigma_K^2(\boldsymbol{x})$, and subsequently estimate the $\hat{P}_f^{MCS}$ on *S*.

- **Step 5**: *Identification of the next training point.* Select the next most valuable point for training according to $\boldsymbol{x}^*=\underset{\boldsymbol{x}\in S}{\mathrm{Max}}(EFF)$ or $\boldsymbol{x}^*=\underset{\boldsymbol{x}\in S}{\mathrm{Min}}(U)$.

- **Step 6**: *Updating the set of training points.* Add the identified point $\boldsymbol{x}^*$ to the set of training points $\boldsymbol{x}_{tr}$.

- **Step 7**: *Estimation of the maximum error.* Determine the maximum error $\hat{\epsilon}_{max}$ using Eq. (16), (17), (23), (24), (27), (30).

- **Step 8**: *Evaluation of the stopping criterion.* Check the stopping criterion ($\hat{\epsilon}_{max} \leq \epsilon_{thr}$). If the stopping criterion is not satisfied, then go to step 3, otherwise, go to step 9.

- **Step 9**: *Evaluation of the sufficiency of initial design sample set.* Determine the coefficient of variation of $\hat{P}_f$ using:

$$COV_{P_f} = \sqrt{\frac{1-\hat{P}_f}{\hat{P}_f N_{MCS}}} \leq COV_{\text{thr}} \tag{33}$$

  where COV_{thr} is the threshold for the coefficient of variation of $\hat{P}_f$, and is usually assigned 0.05 [13]. If Eq. (33) is satisfied then go to step 10. If not, it means that the number of candidate design samples N_{MCS} is not sufficient, and an additional number N_{Δ_S} of candidate design samples Δ_S should be added to *S.* Then go back to step 4.
- **Step 10**: *End.* Report $\hat{P}_f$.

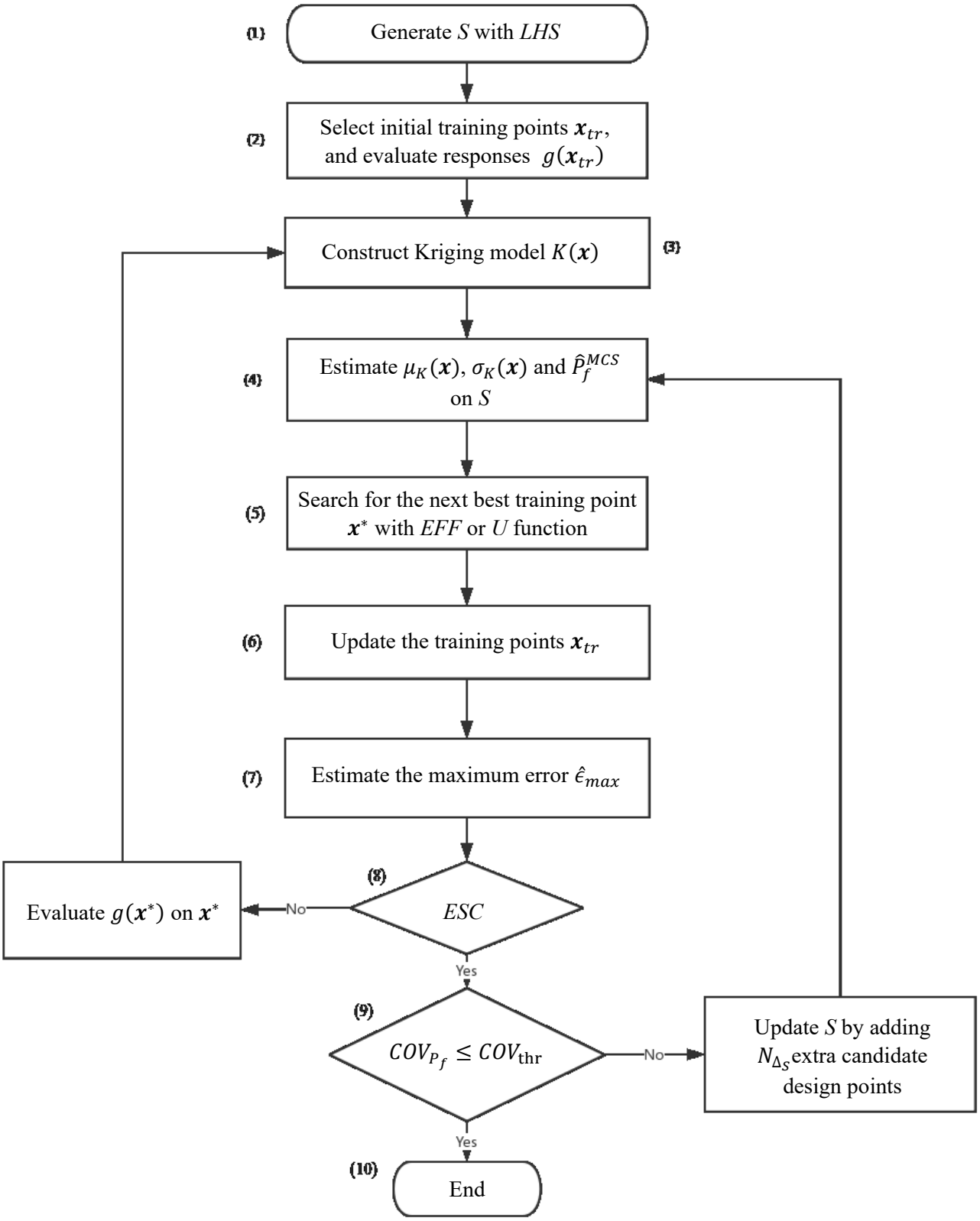

Fig 2. Flowchart of the proposed adaptive Kriging-based reliability analysis method using ESC

4. Numerical Investigations

In this section, five examples with different forms of complexities are investigated. These examples have one or more of the following properties: highly nonlinear, non-differentiable, and high-dimensional.

4.1 Four-boundary series system

The first analytical example is a series-system with four boundaries [13], [23], [27], [32]. The performance function includes two independent identically distributed standard normal random variables (e.g. mean of 0 and standard deviation of 1) x_1 and x_2 as follows:

$$g(x_1, x_2) = min \begin{cases} 3 + 0.1(x_1 - x_2)^2 - \dfrac{(x_1 + x_2)}{\sqrt{2}} \\ 3 + 0.1(x_1 - x_2)^2 + \dfrac{(x_1 + x_2)}{\sqrt{2}} \\ (x_1 - x_2) + \dfrac{6}{\sqrt{2}} \\ -(x_1 - x_2) + \dfrac{6}{\sqrt{2}} \end{cases} \tag{34}$$

This reliability analysis is performed using MCS and adaptive Kriging with *EFF* and *U* learning functions for different stopping criteria. The proposed stopping criterion *ESC* is compared with the conventional stopping criteria $Max(EFF) \leq 10^{-3}$ and $Min(U) \geq 2$. The number of initial training points affects the quality of the initial Kriging model and the computational demand of the reliability analysis. For the same or very similar problems to those considered in this paper, the study in [13] indicated that 12 initial training points are adequate. This value is adopted in this research. The performance of these methods is compared in terms of the number of calls to performance function, N_{call}, estimated probability of failure, $\hat{P}_f$, coefficient of variation of estimated probability of failure, COV_{P_f}, estimated maximum error, $\hat{\epsilon}_{max}$, and the true error, ϵ. Table 1 presents reliability analysis results for *ESC + EFF*, *AK-MCS + EFF*, *ESC + U* and *AK-MCS + U* methods for different ϵ_{thr} values. The threshold COV_{thr} for COV_{P_f} is set as 0.05, and the initial number of candidate design points for each simulation is $N_S = 10^4$ with $N_{\Delta_S} = 10^4$. In this comparison, the same set of candidate design samples and initial training points are used for all methods in order to remain consistent in failure probability estimation.

Table 1. Reliability analysis results for *MCS*, *ESC + EFF, AK-MCS + EFF, ESC + U* and *AK-MCS + U* for different ϵ_{thr} ($COV_{\text{thr}} = 0.05$, $N_S = 10^4$, and $N_{\Delta_S} = 10^4$).

ϵ_{thr}	Methodology	N_{call}	$\hat{P}_f (COV_{\hat{P}_f})$	$\hat{\epsilon}_{max}$	ϵ
-	Monte Carlo	10^5	4.520×10^{-3} (4.7%)	-	-
	AK-MCS + EFF	12 + 93	4.520×10^{-3}	No estimation	0
	AK-MCS + U	12 + 68	4.510×10^{-3}	No estimation	0.0022
0.05	*ESC + EFF*	12 + 40	4.700×10^{-3}	0.0444	0.0398
	ESC + U	12 + 41	4.530×10^{-3}	0.0362	0.0022
0.03	*ESC + EFF*	12 + 60	4.520×10^{-3}	0.0216	0
	ESC + U	12 + 43	4.530×10^{-3}	0.0279	0.0022
0.01	*ESC + EFF*	12 + 67	4.540×10^{-3}	0.0089	0.0044
	ESC + U	12 + 58	4.500×10^{-3}	0.0090	0.0044

As shown in Table 1, the conventional stopping criterion $Max(EFF) \leq 10^{-3}$ requires 105 calls to the performance function, while the proposed *ESC + EFF* for $\epsilon_{thr} = 0.05$, $\epsilon_{thr} = 0.03$ and $\epsilon_{thr} = 0.01$ requires N_{call} of 52, 72 and 79, respectively. Furthermore, the true error for the case of $\epsilon_{thr} = 0.05$ is only 3.98%, which is considered accurate for engineering applications. The true error reduces significantly to 0 and 0.44% for $\epsilon_{thr} = 0.03$ and $\epsilon_{thr} = 0.01$, respectively. In the case of *U* learning function, *ESC* yields even a better performance. The N_{call} for *ESC + U* approach is 53, 55 and 70 for

$\epsilon_{thr} = 0.05$, $\epsilon_{thr} = 0.03$ and $\epsilon_{thr} = 0.01$, respectively. The limit states estimated for different levels of error threshold are plotted in the Fig. 3 and Fig. 4 for *EFF* and *U* learning functions. It is shown that the proposed stopping criterion can achieve the target errors for estimated failure probability with limited number of performance function evaluations. In both figures, it is observed that as the target threshold for error decreases, the accuracy of the Kriging-based estimate of the limit state function increases.

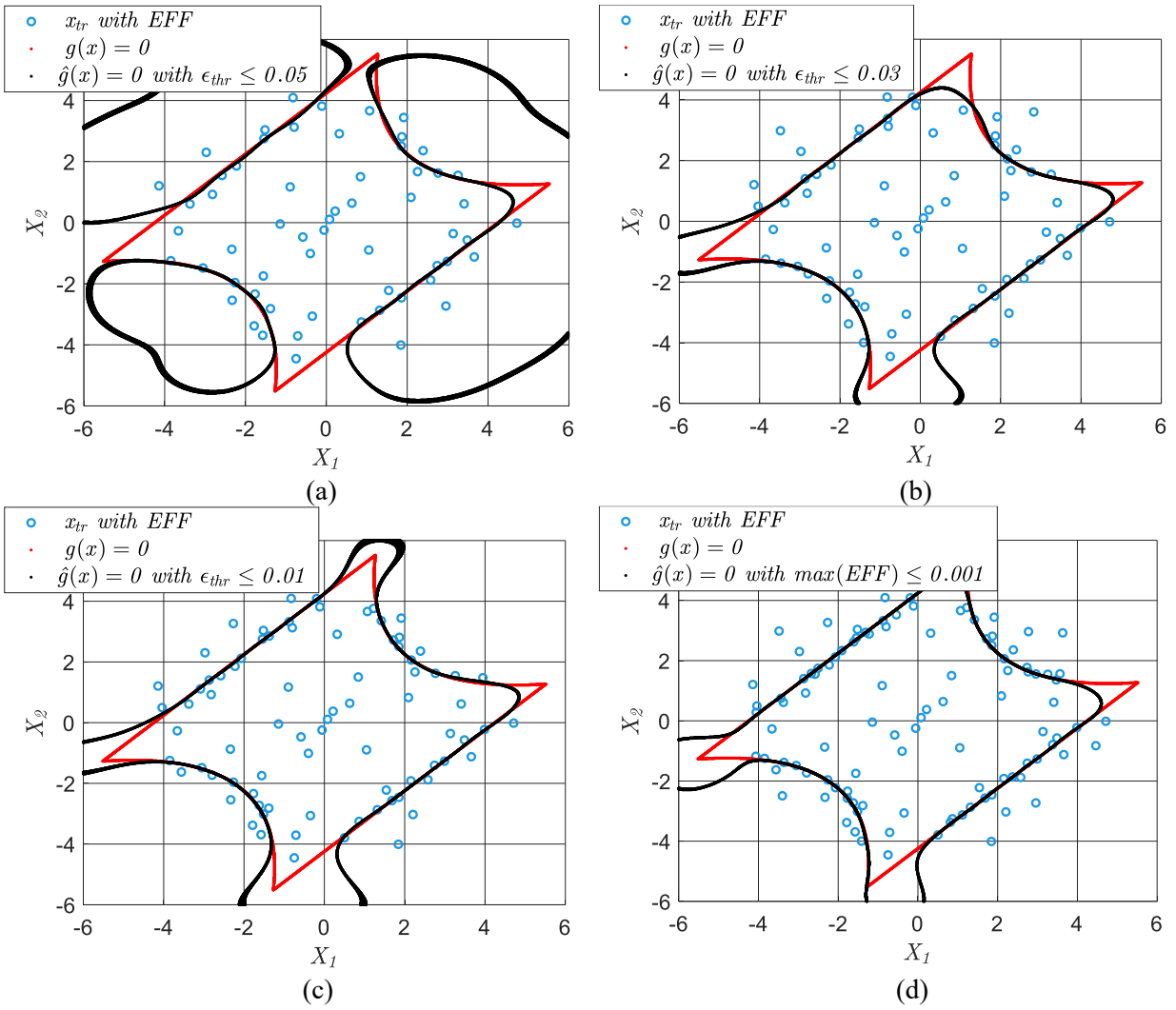

Fig. 3. The limit state with different levels of thresholds by *EFF* learning function in Table 1 (a) $\epsilon_{thr} \leq 0.05$, (b) $\epsilon_{thr} \leq 0.03$, (c) $\epsilon_{thr} \leq 0.01$, and (d) $\max(EFF) \leq 10^{-3}$.

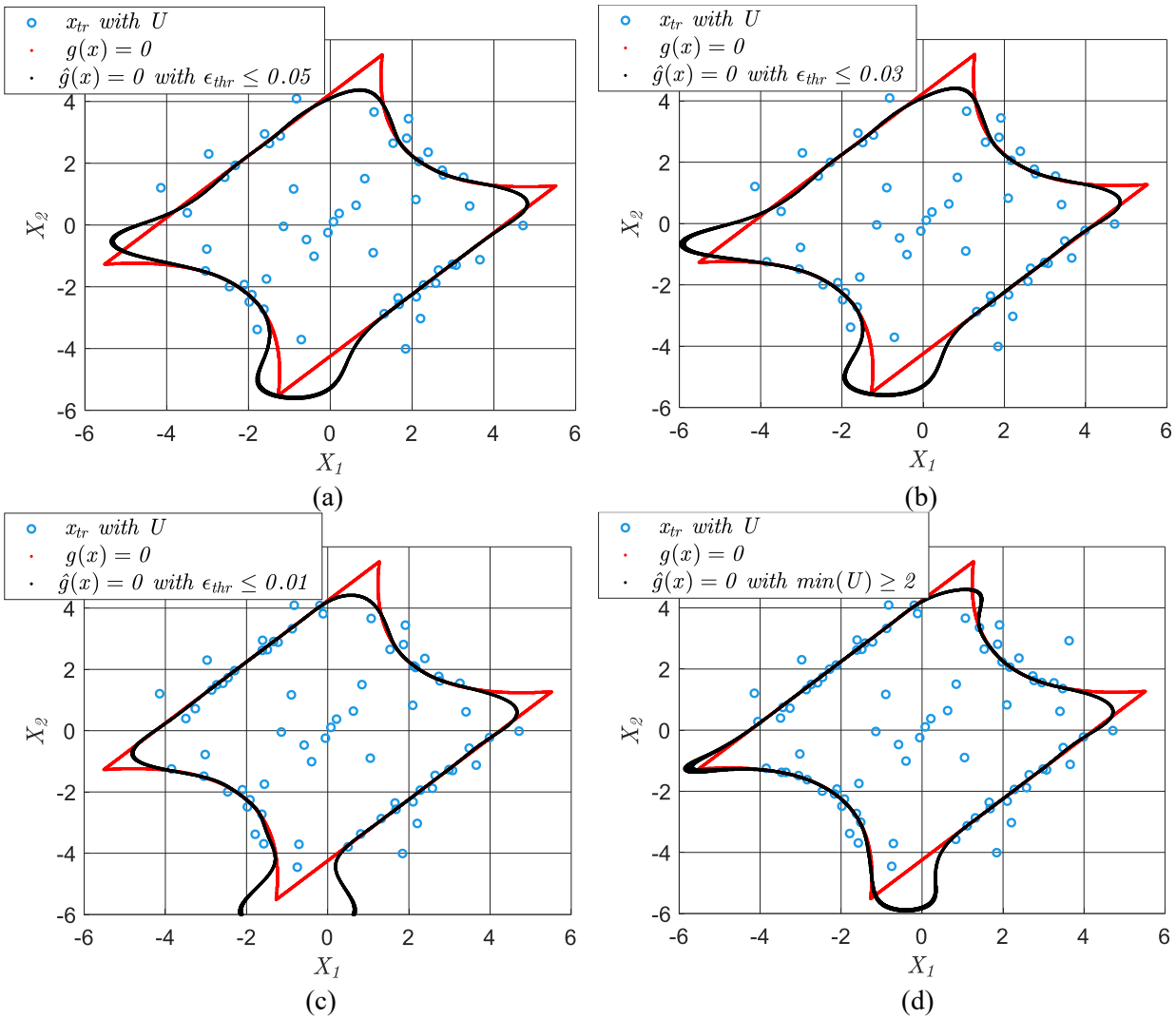

Fig. 4. The limit state with different levels of thresholds by U learning function in Table 1 (a) $\epsilon_{thr} \leq 0.05$, (b) $\epsilon_{thr} \leq 0.03$, (c) $\epsilon_{thr} \leq 0.01$, and (d) $\min(U) \geq 2$.

The convergence of the estimated failure probability by the conventional and proposed methods to true failure probability is presented in Fig. 5. It is seen that the case of $\epsilon_{thr} = 0.05$ yields fastest convergence, as expected. Even in the strict case of $\epsilon_{thr} = 0.01$, the convergence of the failure probability is faster than the traditional stopping criteria. Moreover, one should note that at about $N_{call} = 60$ in Fig. 5 (a) and $N_{call} = 50$ in Fig. 5 (b), $N_{\Delta_S} = 10^4$ extra candidate design samples are added nine times to satisfy the requirement of $COV_{\hat{P}_f} \leq COV_{thr}$ (see step 9 in Section 3.4). With every addition of new candidate design sample sets, the estimated error rate changes. However, all simulations finally converge to the red solid line, i.e. the estimate of the failure probability by MCS.

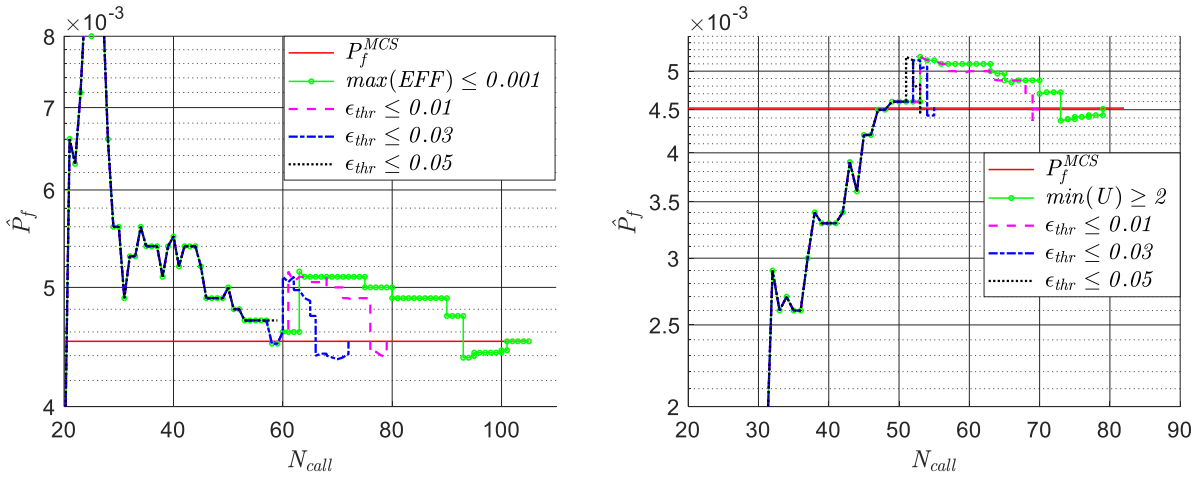

Fig. 5. The convergence performance in Table 1 with: (a) P_f vs N_{call} in *ESC* + *EFF* and *AK-MCS* + *EFF*. (b) P_f vs N_{call} in *ESC* + *U* and *AK-MCS* + *U*.

The performance of *ESC* in avoiding unnecessary training of the surrogate model is more significant for the case of $COV_{thr} = 0.015$, as presented in Table 2. As this threshold for the variation in estimated probability of failure is stricter compared to $COV_{thr} = 0.05$ in Table 1, a larger set of candidate design samples is required. For this reason, the initial number of candidate design points for each simulation is $N_S = 10^5$ with $N_{\Delta_S} = 10^5$. To remain consistent in the comparisons, the set of candidate design samples and the set of initial training points are kept the same for failure probability estimation. According to the results in Table 2, N_{call} for *AK-MCS* + *EFF* and *AK-MCS* + *U* is 120 and 117, respectively, while this quantity for *ESC* + *EFF* and *ESC* + *U* cases for $\epsilon_{thr} = 0.05$ are 53 and 58, respectively, representing over 50% reduction in the computational demand. Therefore, the proposed method can help researchers reduce the number of calls to complex computational codes according to their accuracy requirement.

Table 2. Reliability analysis results for *MCS*, *ESC* + *EFF*, *AK-MCS* + *EFF*, *ESC* + *U* and *AK-MCS* + *U* for different ϵ_{thr} ($COV_{\text{thr}} = 0.015$ [13], $N_S = 10^5$, and $N_{\Delta_S} = 10^5$).

ϵ_{thr}	Methodology	N_{call}	$\hat{P}_f(COV_{\hat{P}_f})$	$\hat{\epsilon}_{max}$	ϵ
-	Monte Carlo	10^6	4.434×10^{-3} (1.5%)	-	-
	AK-MCS + *EFF*	12 + 108	4.437×10^{-3}	No estimation	0.0007
	AK-MCS + *U*	12 + 105	4.434×10^{-3}	No estimation	0
0.05	*ESC* + *EFF*	12 + 41	4.424×10^{-3}	0.0434	0.0023
	ESC + *U*	12 + 46	4.390×10^{-3}	0.0218	0.0099
0.03	*ESC* + *EFF*	12 + 49	4.468×10^{-3}	0.0253	0.0077
	ESC + *U*	12 + 46	4.390×10^{-3}	0.0218	0.0099
0.01	*ESC* + *EFF*	12 + 61	4.429×10^{-3}	0.0080	0.0011
	ESC + *U*	12 + 54	4.419×10^{-3}	0.0068	0.0034

In simulation-based reliability analysis methods, the estimate of the failure probability and the performance of the analysis techniques may vary with the sample set used. To capture these variations and ensure the robustness of the proposed method, the variation in the performance for 100 simulations is investigated via boxplot for the considered set of error thresholds. Considered performance measures include the number of calls to performance function N_{call} and the difference between the estimated maximum error and true error, $\epsilon_{max} - \epsilon$. The $COV_{\hat{P}_f}$ is chosen as 0.05 and the total number of candidate

design points is 10^5 with initial number of samples $N_S = 10^4$ and $N_{\Delta_S} = 10^4$. Results of these analyses are presented in Fig. 6 in terms of (a) N_{call} vs ϵ_{thr} for *ESC +EFF*, (b) $\epsilon_{max} - \epsilon$ vs ϵ_{thr} for *ESC +EFF*, (c) N_{call} vs ϵ_{thr} for *ESC + U*, and (d) $\epsilon_{max} - \epsilon$ vs ϵ_{thr} for *ESC + U*, (e) N_{call} for *AK-MCS + EFF* and *AK-MCS + U*. Results indicate that:

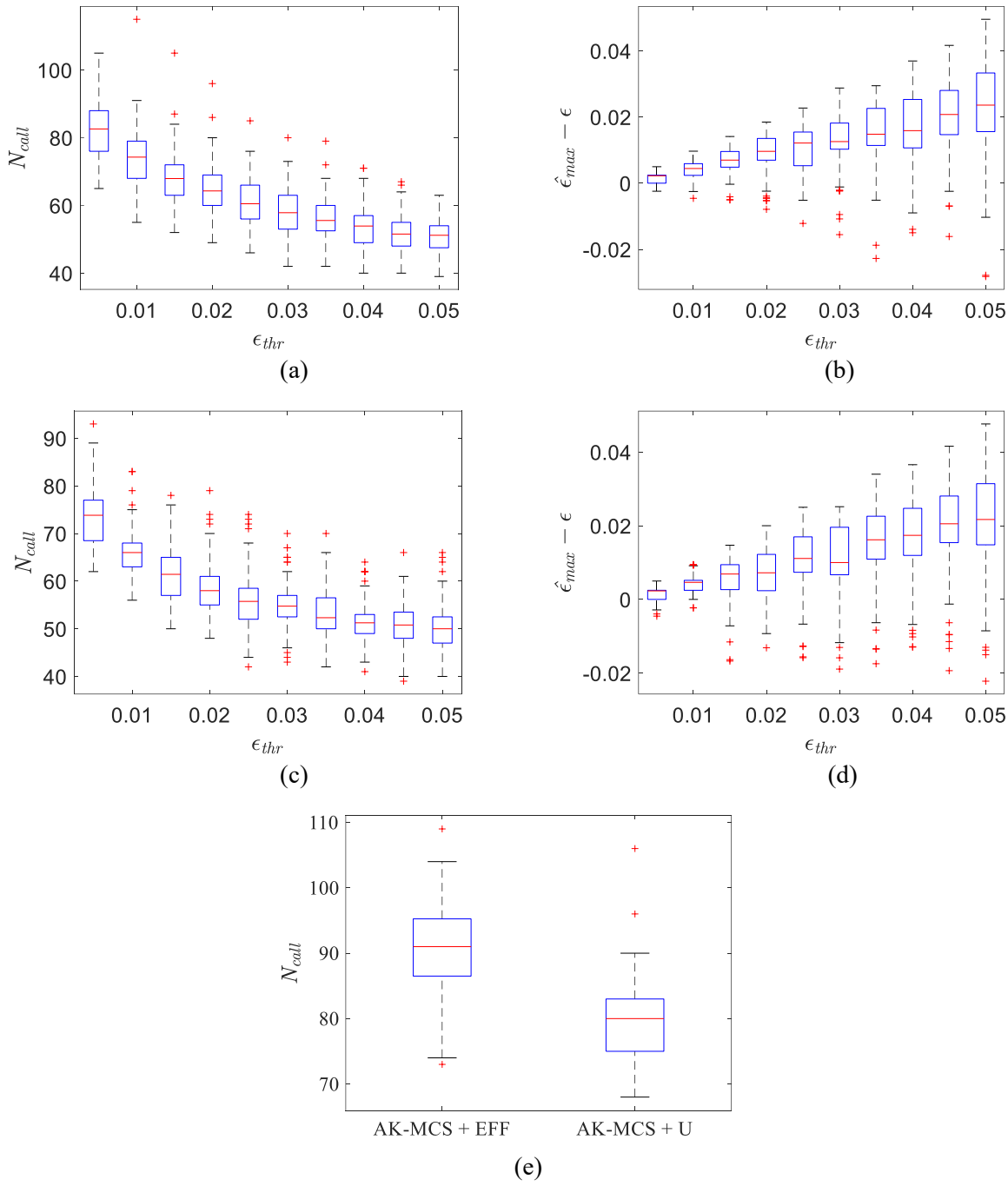

Fig. 6. Boxplots of (a) N_{call} vs ϵ_{thr} for *ESC +EFF*, (b) $\epsilon_{max} - \epsilon$ vs ϵ_{thr} for *ESC +EFF*, (c) N_{call} vs ϵ_{thr} for *ESC + U*, and (d) $\epsilon_{max} - \epsilon$ vs ϵ_{thr} for *ESC + U*, (e) N_{call} for *AK-MCS + EFF* or *AK-MCS + U*

- The average number of calls to performance function for *AK-MCS + EFF*, *AK-MCS + U*, *ESC + EFF* and *ESC + U* methods are 92, 80, 52 and 50 as shown in Fig. 6 (a), (c) and (e), respectively, for the case when the error threshold ϵ_{thr} is 0.05. This shows that the proposed method based on ESC reduces $\overline{N}_{call}$ considerably compared to *AK-MCS*. Even for the higher accuracy of ϵ_{thr}= 0.01, the average number of calls to performance function for the ESC-based *EFF* and *U* approaches are 82 and 74, respectively, which are smaller than 92 and 80 for *AK-MCS + EFF* and *AK-MCS + U*.

- Generally, as the threshold for error ϵ_{thr} increases, the average number of calls to performance function $\overline{N}_{call}$ decreases. This point is evident in Fig. 6 (a) and (c).

- The estimated maximum error ϵ_{max} is greater than the true error ϵ in most cases as seen in Fig. 6 (b) and (d), which follows the principle in Eq. (32). However, there indeed exists some cases that ϵ is slightly greater than ϵ_{max}. This is attributed to the fact that the derivation of ϵ_{max} is based on a prescribed confidence level, here 95% with $\alpha = 0.05$. Therefore, it is expected that for a limited number of cases, ϵ will be larger than ϵ_{max}.

4.2 Modified Rastrigin function

The second example considered here is the modified Rastrigin function, which is a highly nonlinear limit state function and thus requires a large number of calls to performance function to refine surrogate models [13], [27], [42]. This performance function is defined as:

$$g(x_1, x_2) = 10 - \sum_{i=1}^{2} \left(x_i^2 - 5cos(2\pi x_i) \right) \tag{49}$$

where x_is are independent standard normal random variables (e.g. mean of 0 and standard deviation of 1). Reliability analysis results for three levels of error are summarized in Table 3. In these analyses, the threshold COV_{thr} for COV_{P_f} is set as 0.037, and the initial number of candidate design points for each simulation is $N_S = 10^3$ with $N_{\Delta_S} = 10^3$. For consistency in the comparisons, the set of candidate design samples and the set of initial training points are kept the same.

Table 3. Reliability analysis results for *MCS*, *ESC + EFF*, *AK-MCS + EFF*, *ESC + U* and *AK-MCS + U* for different ϵ_{thr} ($COV_{\text{thr}} = 0.037$, $N_S = 10^3$, and $N_{\Delta_S} = 10^3$)

ϵ_{thr}	Methodology	N_{call}	$\hat{P}_f(COV_{\hat{P}_f})$	$\hat{\epsilon}_{max}$	ϵ
-	Monte Carlo	10^4	7.120×10^{-2} (<3.7%)	-	-
	AK-MCS + EFF	12 + 370	7.120×10^{-2}	No estimation	0
	AK-MCS + U	12 + 335	7.120×10^{-2}	No estimation	0
0.05	*ESC + EFF*	12 + 220	7.010×10^{-2}	0.0494	0.0154
	ESC + U	12 + 189	6.84×10^{-2}	0.0487	0.0393
0.03	*ESC + EFF*	12 + 260	7.120×10^{-2}	0.0289	0
	ESC + U	12 + 211	7.04×10^{-2}	0.0246	0.0112
0.01	*ESC + EFF*	12 + 313	7.090×10^{-2}	0.0098	0.0042
	ESC + U	12 + 287	7.060×10^{-2}	0.0098	0.0084

It is evident that *ESC* is computationally very efficient compared to the conventional *AK-MCS* approach. When ϵ_{thr} is 0.05, *ESC + EFF* and *ESC + U* estimate the failure probability with N_{call} of 232 (ϵ = 1.54%) and 201 (ϵ = 3.93%), respectively, compared to N_{call} of 382 and 347 for *AK-MCS + EFF* and *AK-MCS + U* methods. Even in the strict case of $\epsilon_{thr} = 0.01$, the N_{call} using *ESC* is still lower than the traditional stopping criteria. The complex form of the limit state function of this problem is illustrated in Fig. 7 and Fig.8 for *EFF* and *U* learning functions, respectively. The unnecessary training in the conventional *AK-MCS* is clearly seen here as it refines the Kriging model in regions with weak

probability density. However, in the proposed stopping criterion, these 'overfitting' problems are solved by setting a threshold to control the maximum error; as such, the Kriging model is refined in regions that contribute the most to the failure probability estimation until the set error threshold is met.

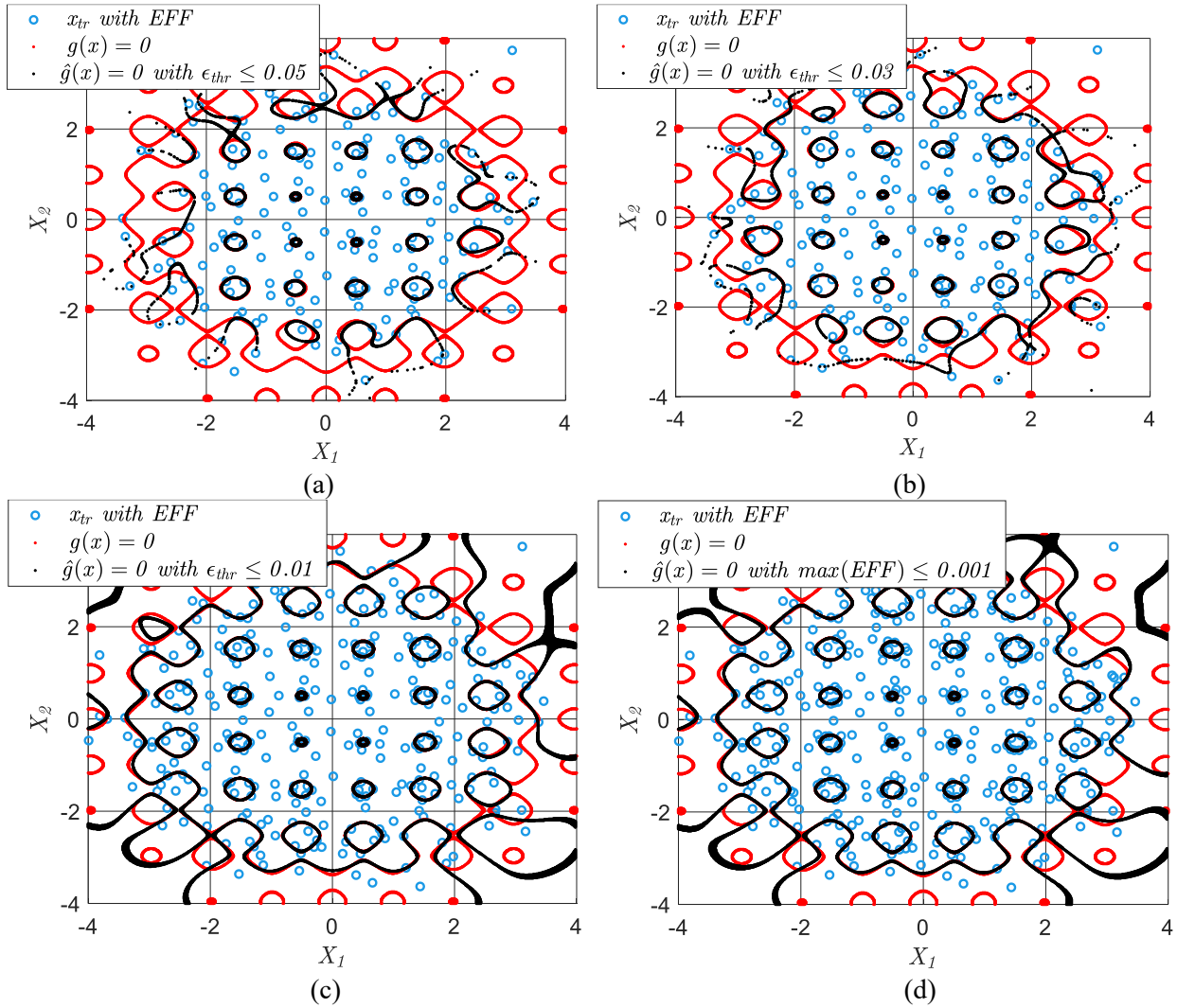

Fig. 7. The limit state with different levels of thresholds by *EFF* learning function in Table 3 (a) $\epsilon_{thr} \leq 0.05$, (b) $\epsilon_{thr} \leq 0.03$, (c) $\epsilon_{thr} \leq 0.01$, and (d) $\max(EFF) \leq 10^{-3}$.

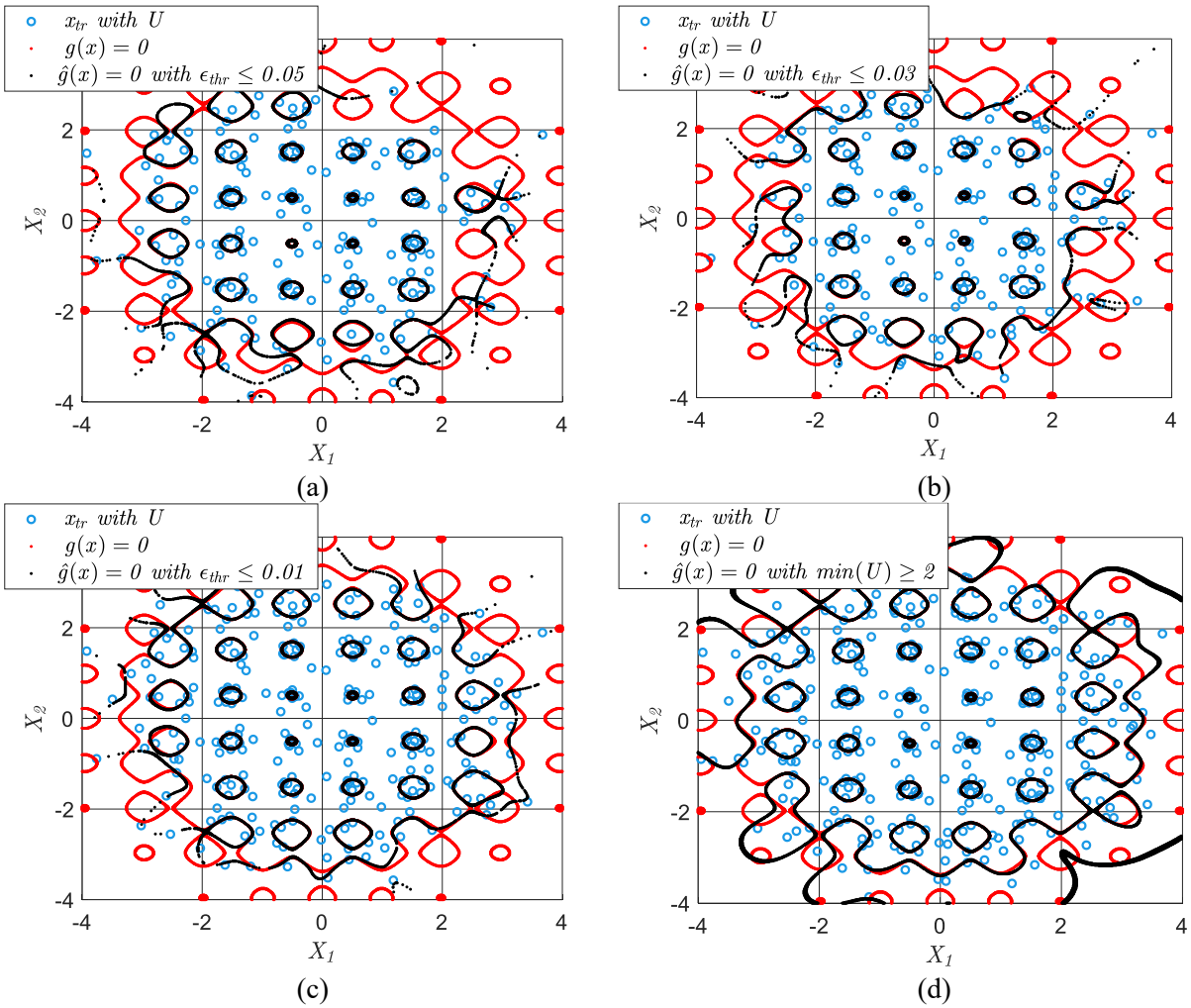

Fig. 8. The limit state with different levels of thresholds by U learning function in Table 3 (a) $\epsilon_{thr} \leq 0.05$, (b) $\epsilon_{thr} \leq 0.03$, (c) $\epsilon_{thr} \leq 0.01$, and (d) $min(U) \geq 2$.

4.3 Nonlinear oscillator

The third example is a nonlinear and un-damped single degree of freedom (Fig. 9) with six random variables. The details of this model can be found in [13], [42]–[45]. The performance function is described below:

$$g(c_1, c_2, m, r, t_1, F_1) = 3r - \left| \frac{2F_1}{m\omega_0^2} sin\left(\frac{\omega_0 t_1}{2}\right) \right| \tag{50}$$

where $\omega_0 = \sqrt{\frac{c_1+c_2}{m}}$ is the system frequency. The probabilistic description of the six random variables is presented in Table 4, and the results of reliability analyses are summarized in Table 5.

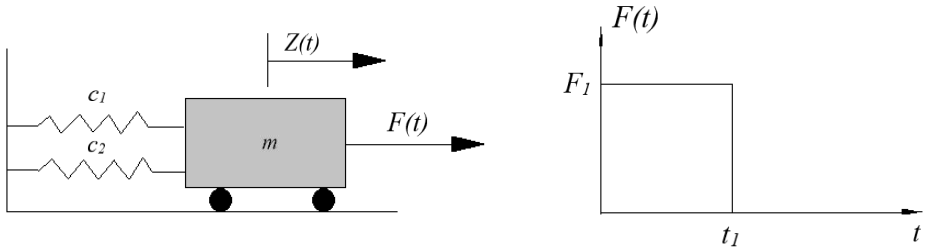

Fig. 9 Example 3, nonlinear oscillator

Table 4. Random variables in example 3.

Random variable	Distribution type	Mean	Standard Deviation
m	Normal	1	0.05
c_1	Normal	1	0.1
c_2	Normal	0.1	0.01
r	Normal	0.5	0.05
F_1	Normal	1	0.2
t_1	Normal	1	0.2

For this example, the threshold COV_{thr} is 0.022, the initial number of candidate design points is $N_S = 10^4$ with $N_{\Delta_S} = 10^4$. Consistent with other examples, results in Table 5 indicate that ESC is very efficient for this moderate dimensional problem. As seen, the use of *ESC* averts the unnecessary training points while accurately estimating the error. These are key features that are not available in AK-MCS. The number of calls to performance function when the error threshold is 0.05 is only 25 for *ESC + EFF* and 24 for *ESC + U* compared with 53 and 73 for *AK-MCS + U* and *AK-MCS + EFF*, respectively. Furthermore, N_{call} slightly increases to 27 for *ESC + EFF* and *ESC + U*, when the error threshold is $\epsilon_{thr} = 0.03$. For the strict error threshold of 0.01, the number of calls to performance function is 29 for both *ESC + EFF* and *ESC + U* compared with 53 in *AK-MCS + EFF* and 73 in *AK-MCS + U*. As seen in Table 5, the accuracy reaches a high level for $\epsilon_{thr} = 0.03$. In the case of $\epsilon_{thr} = 0.01$, the Kriging model achieved a high accuracy for failure probability estimation after 29 number of calls (including the 12 initial points) to performance function. The additional 24 points for *EFF* or 44 for *U* that are used to further refine the model may not be necessary for engineering applications. However, *ESC* stopping criterion successfully avoids unnecessary training by setting the target accuracy.

Table 5. Reliability analysis results for *MCS, ESC + EFF, AK-MCS + EFF, ESC + U* and *AK-MCS + U* for different ϵ_{thr} (COV_{thr} = 0.022, $N_S = 10^4$, and $N_{\Delta_S} = 10^4$).

ϵ_{thr}	Methodology	N_{call}	$\hat{P}_f(COV_{\hat{P}_f})$	$\hat{\epsilon}_{max}$	ϵ
-	Monte Carlo	7×10^4	2.800×10^{-2} (<2.2%)	-	-
	AK-MCS + EFF	12 + 41	2.809×10^{-2}	No estimation	0.0031
	AK-MCS + U	12 + 61	2.800×10^{-2}	No estimation	0
0.05	*ESC + EFF*	12 + 13	2.781×10^{-2}	0.0304	0.0067
	ESC + U	12 + 12	2.794×10^{-2}	0.0355	0.0020
0.03	*ESC + EFF*	12 + 15	2.841×10^{-2}	0.0187	0.0148
	ESC + U	12 + 15	2.807×10^{-2}	0.0195	0.0026
0.01	*ESC + EFF*	12 + 17	2.823×10^{-2}	0.0093	0.0082
	ESC + U	12 + 17	2.807×10^{-2}	0.0072	0.0026

4.4 Ten-dimensional analytical example

The last example is an analytical example presented in [4], [12], [13] with the following form:

$$g(x_1, \dots, x_n) = \left(n + 3\sigma\sqrt{n}\right) - \sum_{i=1}^{n} x_i \tag{51}$$

The random variables, x_is, follow lognormal distribution with mean of 1 and standard deviation $\sigma = 0.2$. To investigate the efficiency of the considered reliability analysis methods for a high dimensional problem, here $n = 10$. Table 6 presents reliability analysis results for three levels of error threshold. Here COV_{thr} = 0.022, the initial number of candidate design points for each simulation is $N_S = 10^5$ and $N_{\Delta_S} = 10^5$.

For this ten-dimensional problem, the proposed method based on *ESC* outperforms the conventional techniques. With 23 number of calls to performance function, the proposed method reaches the true error of 0.07%, while AK-MCS requires more than 60 number of calls to performance function. This result points to the fact that AK-MCS leads to unnecessary training of the Kriging model; however, *ESC* can address this issue and dramatically reduce the number of calls. For the strict error threshold of $\epsilon_{thr} = 0.01$, the number of calls for *ESC* + *EFF* is 42 and for *ESC* + *U* is 44, which are still smaller than those for *AK-MCS* + *EFF* ($N_{call} = 61$) and *AK-MCS* + *U* ($N_{call} = 60$). Thus, the stopping criterion *ESC* successfully alleviates the unnecessary training of the Kriging model in adaptive reliability analysis methods.

Table 6. Reliability analysis results for *MCS, ESC* + *EFF*, *AK-MCS* + *EFF*, *ESC* + *U* and *AK-MCS* + *U* for different ϵ_{thr} (COV_{thr}= 0.022, $N_S = 10^5$, and $N_{\Delta_S} = 10^5$)

ϵ_{thr}	Methodology	N_{call}	$\hat{P}_f(COV_{\hat{P}_f})$	$\hat{\epsilon}_{max}$	ϵ
-	Monte Carlo	2×10^5	2.770×10^{-3} (<2.2%)	-	-
	AK-MCS + *EFF*	12 + 49	2.770×10^{-3}	No estimation	0
	AK-MCS + *U*	12 + 48	2.770×10^{-3}	No estimation	0
0.05	*ESC* + *EFF*	12 + 11	2.772×10^{-3}	0.0414	0.0007
	ESC + *U*	12 + 12	2.770×10^{-3}	0.0433	0
0.03	*ESC* + *EFF*	12 + 18	2.770×10^{-3}	0.0278	0
	ESC + *U*	12 + 17	2.770×10^{-3}	0.0195	0
0.01	*ESC* + *EFF*	12 + 30	2.770×10^{-3}	0.0091	0
	ESC + *U*	12 + 32	2.770×10^{-3}	0.0092	0

4.5 Truss structure with ten dimensions

A 23-bar truss structure with 10 random variables is considered here [36], [46]. The configuration of the truss is shown in Fig. 10. The implicit nonlinear performance function is defined as:

$$g(x) = 0.14 - |dis(\boldsymbol{x})|, \tag{52}$$

where $dis(\boldsymbol{x})$ is the vertical displacement of the truss at point E. The truss is subject to six vertical loadings, P_1 to P_6, which follow Gumbel distributions. A_1 and A_2 are the cross-section areas and E_1 and E_2 are the Young's modulus of the horizontal and diagonal bars, respectively. The 10 mutually independent random variables are described in Table 7.

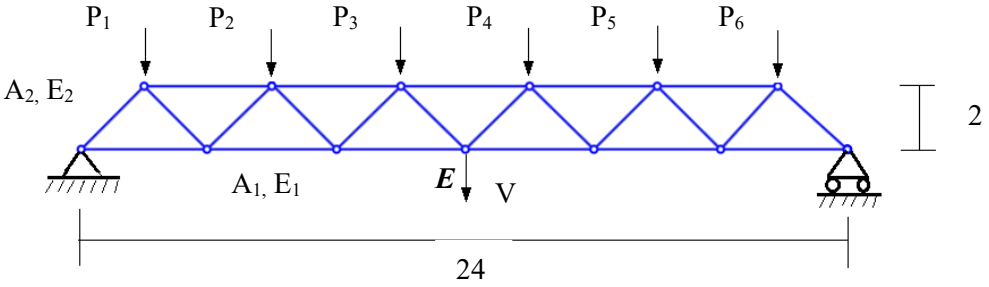

Fig. 10 Example 5, the truss with 10 random variables.

Table 7. Random variables in example 5.

Random variable	Distribution	Mean	Standard deviation
$P_1 - P_6$	Gumbel	6.5×10^4	6.5×10^3
A_1	Lognormal	2×10^{-3}	2×10^{-4}
A_2	Lognormal	1×10^{-3}	1×10^{-4}
E_1	Lognormal	2.1×10^{11}	2.1×10^{11}
E_2	Lognormal	2.1×10^{11}	2.1×10^{11}

For this nonlinear example, the threshold for $\mathrm{COV_{thr}}$ is set as 0.05, and the initial number of candidate design points is $\mathrm{N_S} = 10^4$ with $\mathrm{N_{\Delta_S}} = 10^4$. The simulation results are presented in Table 8. An interesting observation in this example is the inability of *AK-MCS+EFF* to converge to an accurate estimate of failure probability ($\epsilon = 0.0543$) because of the premature termination of the training process. The premise behind *EFF*-based stopping criterion is that the threshold of 10^{-3} in $\max(EFF) \leq 10^{-3}$ is small enough to ensure that the adaptive reliability analysis yields accurate $\widehat{\mathrm{P}}_\mathrm{f}$. As argued in this paper earlier, *EFF* and in general existing stopping criteria do not directly correspond to the extent of error in failure probability estimates. Thus as observed in this example, the threshold is not sufficiently small. Reducing the threshold for the maximum of *EFF* to values smaller than 10^{-3}, however, may considerably increase the unnecessary calls to the performance function for other problems. In this example, *AK-MCS + U* provides an accurate estimate of $\widehat{\mathrm{P}}_\mathrm{f}$ with total 67 simulations. The proposed *ESC + EFF* requires 49, 54 and 58 evaluations to the performance function for error thresholds of 0.05, 0.03, and 0.01, respectively. This indicates that adaptive reliability analysis using the proposed stopping criterion can properly converge to true failure probability. Similar trend can be observed for the case of *ESC + U* approach.

Table 8. Reliability analysis results for *MCS, ESC + EFF*, *AK-MCS + EFF*, *ESC + U* and *AK-MCS + U* for different ϵ_{thr} (COV_{thr}= 0.05, $N_S = 10^4$, and $N_{\Delta_S} = 10^4$)

ϵ_{thr}	Methodology	N_{call}	$\hat{P}_f(COV_{\hat{P}_f})$	$\hat{\epsilon}_{max}$	ϵ
-	Monte Carlo	5×10^4	9.200×10^{-3} (<5%)	-	-
	AK-MCS + EFF	12 + 28	9.700×10^{-3}	No estimation	0.0543
	AK-MCS + U	12 + 55	9.200×10^{-3}	No estimation	0
0.05	*ESC + EFF*	12 + 37	9.400×10^{-3}	0.0309	0.0217
	ESC + U	12 + 29	9.100×10^{-3}	0.0299	0.0109
0.03	*ESC + EFF*	12 + 42	9.300×10^{-3}	0.0152	0.0109
	ESC + U	12 + 37	9.300×10^{-3}	0.0148	0.0109
0.01	*ESC + EFF*	12 + 46	9.200×10^{-3}	0.0022	0
	ESC + U	12 + 55	9.200×10^{-3}	0	0

5. Conclusion

In this article, an Error-based stopping criterion (*ESC*) for Kriging-based reliability analysis algorithms is proposed. It is shown that the total number of candidate design samples wrongly categorized by Kriging as safe and failed follows a Poisson binomial distribution. Based on the statistical properties of this distribution, the confidence intervals of the number of points with wrong sign estimation in both failure and safe domains are derived for a given confidence level. These bounds are subsequently used to determine the maximum error for estimated failure probabilities in the adaptive Kriging reliability analysis process. Finally, a new stopping criterion based on the derived maximum error is proposed. It is shown that the proposed stopping criterion, *ESC*, can avoid unnecessary calls to performance functions and guarantee convergence to target accuracies in Kriging-based reliability analysis methods for the considered confidence level. This new feature helps researchers to establish a balance between computational demand and required accuracy. Five examples are considered to investigate the performance of *ESC* and two other widely used stopping criteria. Numerical results showcase the high efficiency and accuracy of adaptive reliability analysis using *ESC* compared to existing methods.

Acknowledgements

This research has been partly funded by the U.S. National Science Foundation (NSF) through awards CMMI-1462183, 1563372, 1635569, and 1762918. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo method*. John Wiley & Sons, 2016.

[2] G. Fishman, *Monte Carlo: concepts, algorithms, and applications*. Springer Science & Business Media, 2013.

[3] B. Echard, N. Gayton, M. Lemaire, and N. Relun, "A combined importance sampling and kriging reliability method for small failure probabilities with time-demanding numerical models," *Reliab. Eng. Syst. Saf.*, vol. 111, pp. 232–240, 2013.

[4] X. Huang, J. Chen, and H. Zhu, "Assessing small failure probabilities by AK–SS: an active learning method combining Kriging and subset simulation," *Struct. Saf.*, vol. 59, pp. 86–95, 2016.

[5] O. Ditlevsen and H. O. Madsen, *Structural reliability methods*, vol. 178. Wiley New York, 1996.

[6] M. Lemaire, *Structural reliability*. John Wiley & Sons, 2013.

[7] V. J. Romero, L. P. Swiler, and A. A. Giunta, "Construction of response surfaces based on progressive-lattice-sampling experimental designs with application to uncertainty propagation," *Struct. Saf.*, vol. 26, no. 2, pp. 201–219, 2004.

[8] A. A. Giunta, J. M. McFarland, L. P. Swiler, and M. S. Eldred, "The promise and peril of uncertainty quantification using response surface approximations," *Struct. Infrastruct. Eng.*, vol. 2, no. 3–4, pp. 175–189, 2006.

[9] W. Zhao, F. Fan, and W. Wang, "Non-linear partial least squares response surface method for structural reliability analysis," *Reliab. Eng. Syst. Saf.*, vol. 161, pp. 69–77, May 2017.

[10] G. Blatman and B. Sudret, "An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis," *Probabilistic Eng. Mech.*, vol. 25, no. 2, pp. 183–197, 2010.

[11] H. Dai, H. Zhang, W. Wang, and G. Xue, "Structural reliability assessment by local approximation of limit state functions using adaptive Markov chain simulation and support vector regression," *Comput.-Aided Civ. Infrastruct. Eng.*, vol. 27, no. 9, pp. 676–686, 2012.

[12] J.-M. Bourinet, "Rare-event probability estimation with adaptive support vector regression surrogates," *Reliab. Eng. Syst. Saf.*, vol. 150, pp. 210–221, 2016.

[13] B. Echard, N. Gayton, and M. Lemaire, "AK-MCS: an active learning reliability method combining Kriging and Monte Carlo simulation," *Struct. Saf.*, vol. 33, no. 2, pp. 145–154, 2011.

[14] W. Fauriat and N. Gayton, "AK-SYS: An adaptation of the AK-MCS method for system reliability," *Reliab. Eng. Syst. Saf.*, vol. 123, pp. 137–144, 2014.

[15] Z. Wang and A. Shafieezadeh, "REAK: Reliability analysis through Error rate-based Adaptive Kriging," *Reliab. Eng. Syst. Saf.*, vol. 182, pp. 33–45, Feb. 2019.

[16] I. Kaymaz, "Application of kriging method to structural reliability problems," *Struct. Saf.*, vol. 27, no. 2, pp. 133–151, 2005.

[17] B. Gaspar, A. P. Teixeira, and C. G. Soares, "Assessment of the efficiency of Kriging surrogate models for structural reliability analysis," *Probabilistic Eng. Mech.*, vol. 37, pp. 24–34, Jul. 2014.

[18] B. J. Bichon, M. S. Eldred, L. P. Swiler, S. Mahadevan, and J. M. McFarland, "Efficient global reliability analysis for nonlinear implicit performance functions," *AIAA J.*, vol. 46, no. 10, pp. 2459–2468, 2008.

[19] M. Balesdent, J. Morio, and J. Marzat, "Kriging-based adaptive Importance Sampling algorithms for rare event estimation," *Struct. Saf.*, vol. 44, pp. 1–10, Sep. 2013.

[20] V. Dubourg, B. Sudret, and F. Deheeger, "Metamodel-based importance sampling for structural reliability analysis," *Probabilistic Eng. Mech.*, vol. 33, pp. 47–57, Jul. 2013.

[21] V. Dubourg, B. Sudret, and J.-M. Bourinet, "Reliability-based design optimization using kriging surrogates and subset simulation," *Struct. Multidiscip. Optim.*, vol. 44, no. 5, pp. 673–690, Nov. 2011.

[22] N. Pedroni and E. Zio, "An Adaptive Metamodel-Based Subset Importance Sampling approach for the assessment of the functional failure probability of a thermal-hydraulic passive system," *Appl. Math. Model.*, vol. 48, pp. 269–288, Aug. 2017.

[23] Z. Wen, H. Pei, H. Liu, and Z. Yue, "A Sequential Kriging reliability analysis method with characteristics of adaptive sampling regions and parallelizability," *Reliab. Eng. Syst. Saf.*, vol. 153, pp. 170–179, 2016.
[24] X. Yang, Y. Liu, C. Mi, and C. Tang, "System reliability analysis through active learning Kriging model with truncated candidate region," *Reliab. Eng. Syst. Saf.*, vol. 169, pp. 235–241, 2018.
[25] Z. Hu and X. Du, "Mixed Efficient Global Optimization for Time-Dependent Reliability Analysis," *J. Mech. Des.*, vol. 137, no. 5, pp. 051401-051401–9, May 2015.
[26] Z. Lv, Z. Lu, and P. Wang, "A new learning function for Kriging and its applications to solve reliability problems in engineering," *Comput. Math. Appl.*, vol. 70, no. 5, pp. 1182–1197, Sep. 2015.
[27] Z. Sun, J. Wang, R. Li, and C. Tong, "LIF: A new Kriging based learning function and its application to structural reliability analysis," *Reliab. Eng. Syst. Saf.*, vol. 157, pp. 152–165, 2017.
[28] N.-C. Xiao, M. J. Zuo, and C. Zhou, "A new adaptive sequential sampling method to construct surrogate models for efficient reliability analysis," *Reliab. Eng. Syst. Saf.*, vol. 169, pp. 330–338, Jan. 2018.
[29] Z. hu and S. Mahadevan, "A Single-Loop Kriging (SILK) Surrogate Modeling for Time-Dependent Reliability Analysis," *J. Mech. Des.*, vol. 138, Apr. 2016.
[30] N. V. Queipo, R. T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, and P. Kevin Tucker, "Surrogate-based analysis and optimization," *Prog. Aerosp. Sci.*, vol. 41, no. 1, pp. 1–28, Jan. 2005.
[31] A. Chaudhuri and R. T. Haftka, "Effectiveness Indicators for Stopping Criteria based on Minimum Required Improvement," in *56th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, 0 vols., American Institute of Aeronautics and Astronautics, 2015.
[32] B. Gaspar, A. P. Teixeira, and C. G. Soares, "Adaptive surrogate model with active refinement combining Kriging and a trust region method," *Reliab. Eng. Syst. Saf.*, 2017.
[33] Z. Hu and S. Mahadevan, "Global sensitivity analysis-enhanced surrogate (GSAS) modeling for reliability analysis," *Struct. Multidiscip. Optim.*, vol. 53, no. 3, pp. 501–521, Mar. 2016.
[34] "UQLab Kriging (Gaussian process modelling) manual," *UQLab, the Framework for Uncertainty Quantification*. [Online]. Available: http://www.uqlab.com/userguidekriging. [Accessed: 13-May-2017].
[35] "UQLab sensitivity analysis user manual," *UQLab, the Framework for Uncertainty Quantification*. [Online]. Available: http://www.uqlab.com/userguide-reliability. [Accessed: 13-May-2017].
[36] J. Wang, Z. Sun, Q. Yang, and R. Li, "Two accuracy measures of the Kriging model for structural reliability analysis," *Reliab. Eng. Syst. Saf.*, vol. 167, pp. 494–505, Nov. 2017.
[37] S. N. Lophaven, H. B. Nielsen, and J. Søndergaard, "DACE-A Matlab Kriging toolbox, version 2.0," 2002.
[38] S. N. Lophaven, H. B. Nielsen, and J. Søndergaard, "Aspects of the matlab toolbox DACE," Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2002.
[39] L. LeCam, "On the Distribution of Sums of Independent Random Variables," in *Bernoulli 1713 Bayes 1763 Laplace 1813*, Springer, Berlin, Heidelberg, 1965, pp. 179–202.
[40] O. T. Johnson, *Information Theory and the Central Limit Theorem*. World Scientific, 2004.
[41] L. L. Cam, "An approximation theorem for the Poisson binomial distribution.," *Pac. J. Math.*, vol. 10, no. 4, pp. 1181–1197, 1960.
[42] S. K. Au and J. L. Beck, "Subset simulation and its application to seismic risk based on dynamic analysis," *J. Eng. Mech.*, vol. 129, no. 8, pp. 901–917, 2003.
[43] L. Schueremans and D. Van Gemert, "Benefit of splines and neural networks in simulation based structural reliability analysis," *Struct. Saf.*, vol. 27, no. 3, pp. 246–261, 2005.
[44] M. R. Rajashekhar and B. R. Ellingwood, "A new look at the response surface approach for reliability analysis," *Struct. Saf.*, vol. 12, no. 3, pp. 205–220, 1993.
[45] N. Gayton, J. M. Bourinet, and M. Lemaire, "CQ2RS: a new statistical approach to the response surface method for reliability analysis," *Struct. Saf.*, vol. 25, no. 1, pp. 99–121, 2003.
[46] N. Roussouly, F. Petitjean, and M. Salaun, "A new adaptive response surface method for reliability analysis," *Probabilistic Eng. Mech.*, vol. 32, pp. 103–115, Apr. 2013.