Uncertainty Quantification for Extreme Quantile Estimation with stochastic computer models

Qiyun Pan, Young Myoung Ko, *Member, IEEE,* and Eunshin Byon, *Member, IEEE*

***Abstract*—Extreme quantiles are important measures in reliability analysis. At the system design stage, quantiles are often estimated via stochastic simulations. This study aims to quantify quantile estimation uncertainties by constructing confidence intervals using importance sampling when quantiles are estimated via stochastic computer models. We validate the asymptotic normality for the importance sampling quantile estimator and construct a theoretically valid confidence interval in a closed form. A drawback of the theoretical confidence interval is that it needs to consistently estimate a variance parameter. To resolve the limitation of the theoretical confidence interval, we present batching-based approaches which are also built upon the asymptotic normality of the quantile estimator. We compare the estimation performance of studied methods and other alternative methods using numerical examples and wind turbine case study.**

***Index Terms*—Batching, confidence interval, importance sampling, reliability, sectioning, variance reduction**

ACRONYMS AND ABBREVIATIONS

UQ	uncertainty quantification
CDF	cumulative density function
PDF	probability density function
IEC	International Electrotechnical Commission
NREL	National Renewable Energy Laboratory
MCS	Monte Carlo sampling
CMC	crude Monte Carlo
SIS	stochastic importance sampling
CI	confidence interval
CLT	central limit theorem
POE	probability of exceedance
DLC	design load case

NOTATION

Y	response variable
F_Y	CDF of Y
f_Y	PDF of Y
α	probability that Y exceeds the resistance level
y_α	upper α-quantile of Y
X	input vector
$f_{\mathbf{X}}$	pdf of X
$\Omega_{\mathbf{X}}$	support of $f_{\mathbf{X}}$
$q_{\mathbf{X}}$	importance sampling density
n	sample size
$\hat{P}_n$	failure probability estimator
q_X^*	optimal importance sampling density with stochastic black box computer models
C_q	normalizing constant
s	conditional failure probability given $X = x$
$\hat{s}$	metamodel of s
$\hat{y}_{\alpha,n}$	α-quantile estimator
p_y	failure probability
$\hat{P}_n(y)$	failure probability estimator
σ_y^2	asymptotic variance of $\sqrt{n}\hat{P}_n(y)$
R_n, R'_n	remainders in Taylor expansion
ξ	random vector embedded inside a computer model
$f_{X\xi}$	joint pdf of X and ξ
$f_{\xi\|X}$	conditional pdf of ξ, given $X = x$
$q_{X\xi}$	joint importance sampling density of X and ξ
L	likelihood ratio
h_n	bandwidth parameter
b	number of batches
r	batch size
$\hat{P}_{r,k}$	Failure probability estimator at the k^{th} batch
$\hat{y}_{\alpha,r,k}$	quantile estimator at the k^{th} batch
$\hat{y}_{\alpha,b,bat}$	sample average of batch quantile estimates
$S^2_{b,bat}$	sample variance in batching
$S^2_{b,sec}$	sample variance in sectioning

I. Introduction

This study considers uncertainty quantification (UQ) in estimating quantiles via stochastic computer models. Quantiles are important measures in the reliability analysis for physical and engineering systems, or risk analysis for social, environmental, and financial systems [1]. Consider a continuous random variable Y with its cumulative density function (CDF) $F_Y(y) = P(Y \leq y)$. The upper α-quantile (called α-quantile in this paper) is defined as the constant y_α such that $F_Y(y_\alpha) = 1 - \alpha$. In the reliability analysis, y_α is also known as a resistance level [2].

Q. Pan and E. Byon are with the Department of Industrial & Operations Engineering, University of Michigan, Ann Arbor, MI, 48109 USA e-mail: qiyun@umich.edu and ebyon@umich.edu (corresponding author).

Young Myoung Ko is with the Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, Gyeongbuk, 37673, Korea e-mail: youngko@postech.ac.kr.

This work was supported by the U.S. National Science Foundation under Grants IIS-1741166 and CMMI-1662553.

In particular, we consider a system that operates under stochastic conditions. When the reliability of such systems is assessed at the design stage, system manufacturers may conduct a field measurement campaign during a short period of time [3]. However, when real operating data is scarce, field data is usually not sufficient due to rare occurrences of extreme events. To supplement data scarcity, simulation models are often employed. For example, in the wind industry, estimating extreme load responses becomes increasingly crucial for determining design parameters of large-scale wind turbines. Accordingly, the International Electrotechnical Commission (IEC)'s design standard [4] requires wind turbine designers to estimate the extreme load response associated with a pre-specified small failure probability. In response, aeroelastic simulators have been developed in the wind energy community. For instance, the U.S Department of Energy's National Renewable Energy Laboratory (NREL) developed a set of simulators to generate load response data for a turbine

operating under stochastic operating environment [5], [6].

There are two major approaches for extreme quantile estimation with simulation models. The first approach is to develop an emulator that can approximate the simulation model and estimate the quantile with the resulting emulator. For example, the Gaussian process has been widely used in the computer experiment literature [7], [8]. For estimating extreme load responses in a wind turbine, statistical models based on extreme value distribution have been studied in [3], [9]. However, the main purpose of such emulators is to estimate general characteristics of the response surface over the input area. Such approaches often show poor estimation for analyzing tail probability [2], [10].

Another approach is to use Monte Carlo sampling (MCS) that generates data via simulation [11]. The most brute-force MCS method is a so-called crude Monte Carlo (CMC) that uses the original input distribution to run the simulation. However it has been well-known that CMC typically requires extensive computational resources for extreme quantile estimation and its estimation variance is high [2], [12]. To address these issues, various variance reduction techniques have been studied in the literature, among which importance sampling has been proven to be a powerful tool [13]. Most studies in importance sampling consider simulation models, which are called *deterministic computer models* in this study, which generate a deterministic output given the same input. Recently, in an attempt to resemble actual stochastic systems more realistically, some modern simulators employ *stochastic computer models* where random outputs are generated even at the same input. The NREL simulator is one example of such stochastic computer models.

Choe et al. [14] developed an importance sampling method that can reduce the variance for failure probability estimation using stochastic computer models, referred to as stochastic importance sampling (SIS). The SIS method was applied to the NREL simulator for estimating extreme load responses in a wind turbine in [2], showing great advantages of SIS over CMC, in terms of both computational efficiency and variance reduction. Both studies in [2], [14] focused on point estimations of the failure probability and quantile.

In the reliability analysis, UQ is as important as the point estimation for evaluating the estimation accuracy [15]. In general, UQ can be done by establishing a confidence interval (CI) of an estimator, which is typically constructed based on the central limit theorem (CLT). For example, Choe et al. [16] derived the CLT and asymptotic CI of the failure probability when SIS is applied to stochastic computer models.

Typically, CLT can be driven for the estimator in the form of the sample average. Deriving CLT for the quantile estimator is, however, nontrivial because it does not take the form of sample average. In the literature, Sun and Hong [17] studied the asymptotic normality of the quantile estimator when importance sampling is used. But the resulting CI includes an unknown variance parameter, which prevents one from implementing it in practice. Chu and Nakayama [12] further advanced the theoretical results and present certain conditions where the CLT for a quantile estimator hold. Asumussen and Glynn [18] introduced the batching-based approach to construct a quantile CI. Based on the asymptotic normality of the quantile estimator established in [12], Nakayama [19] also constructed several batching-based approaches, including batching, sectioning and sectioning-batching, using variance reduction techniques under certain conditions. The CIs developed in these studies, however, have been generally applied to deterministic computer models.

The main contribution of this study is to derive the asymptotic normality of the quantile estimator when quantile is estimated using stochastic computer models (Figure 1). The resulting validity allows us to develop an asymptotic quantile CI in a closed-form. It also provides a theoretical foundation to construct CIs with batching-based approaches. Built upon the asymptotic result, we show that the batching, sectioning and sectioning-batching approaches [19] can be applicable for obtaining quantile CIs with stochastic computer models. To the best of our knowledge, this is the first study that constructs quantile CIs with stochastic computer models.

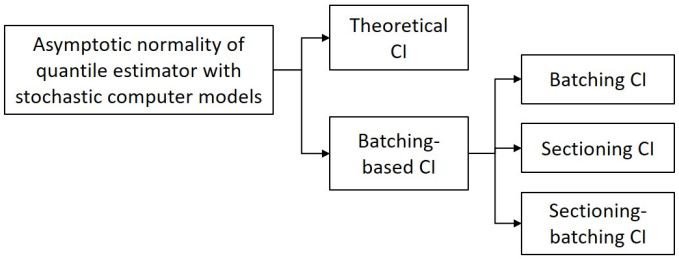

Fig. 1: Proposed quantile CI construction, based on the asymptotic normality property of the quantile estimator with stochastic computer models

We implement the CIs from the theoretical approach and batching-based approaches through a numerical example and case study. The CI performance of these approaches are compared with other alternatives, including bootstrapping and Jackknife. The results suggest that the batching-based approach provides stable results. Specifically, the sectioning-batching generates narrow confidence intervals with high coverage rates in most cases.

The remaining of the paper is structured as follows. Section II provides background and reviews the importance sampling for stochastic computer models. Section III theoretically derives the quantile CI and presents the CIs from three batching-based approaches. Section IV compares the CIs from different methods with a numerical example. Section V presents a wind turbine case study using the NREL simulators. Section VI concludes the paper.

II. Review of Importance Sampling for Stochastic Computer Models

This paper considers reliability against extreme shocks (or extreme loads). Let $\mathbf{X} \in \mathbb{R}^p$ denote a random input vector with its pdf $f_{\mathbf{X}}(x)$ that represents stochastic operating condition. For the system operating under stochastic operating condition,

the failure probability, or probability of exceedance (POE), is typically defined as

$$P(Y > y) = \int_{\Omega_{\mathbf{X}}} P(Y > y \mid X = x) f_{\mathbf{X}}(x) dx, \quad (1)$$

where $\Omega_{\mathbf{X}}$ represents the support of the input density $f_{\mathbf{X}}$, i.e., $\Omega_{\mathbf{X}} = \{x \mid f_{\mathbf{X}}(x) > 0\}$, and y is a resistance level (or threshold level). This failure probability is the same as the counter cumulative distribution function of Y, that is, $1 - F_Y(y)$.

The α-quantile, denoted by y_α, is the value y that satisfies $P(Y > y_\alpha) = \alpha$. Mathematically, it can be defined as

$$y_\alpha = \inf\{y \in \mathbb{R} : P(Y > y) \leq \alpha\}. \quad (2)$$

To estimate y_α, one can use CMC that draws the input $\mathbf{X}$ from $f_{\mathbf{X}}$ to run the simulator and obtain the output Y. However, when α is small, one needs a large number of simulation runs to observe the exceedance event $\{Y > y_\alpha\}$ sufficiently many times, so as to get an accurate estimate of y_α. As a result, CMC is usually computationally inefficient, leading to large estimation variance when the computational resource is limited.

On the contrary, importance sampling draws the input from a biased density, so that more sampling effort can be allocated to the input area that generates the event of interest. Consider a biased importance sampling density $q_{\mathbf{X}}$. Let X_i, $i = 1, \cdots, n$, be a sample generated from $q_{\mathbf{X}}$ and Y_i denote the simulation output at each X_i. The failure probability estimator, or the POE estimator, becomes

$$\hat{P}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(Y_i > y | X_i = x_i\right) \frac{f_X(X_i)}{q_X(X_i; y)}, \quad (3)$$

where $f_X(X_i)/q_X(X_i; y)$ is used to recover unbiasness of $\hat{P}_n(y)$ because X_i is drawn from q_X.

Importance sampling has been widely applied to deterministic computer models where the output is uniquely determined at a fixed input. Recently, Choe et al. [14] developed its stochastic counterpart and derived the optimal importance sampling density q_X^* that minimizes the variance of $\hat{P}_n(y)$. The optimal SIS density q_X^* that minimizes the variance of $\hat{P}_n(y)$ evaluated at $y = y_0$ is given by

$$q_X^*(x; y_0) = \frac{1}{C_{q*}} f_X(x) \sqrt{s(x; y_0)}, \quad (4)$$

where $C_{q*} = \int_{\Omega_{\mathbf{X}}} f_X(x) \sqrt{s(x; y_0)} dx$ (> 0) is the normalizing constant satisfying $C_{q*} > 0$ and $s(x; y_0)$ is the conditional POE at x,

$$s(x; y_0) = P(Y > y_0 \mid X = x). \quad (5)$$

In reality, $s(x; y_0)$ is unknown when the simulator is treated as a black-box. Choe et al. [14] suggest using a metamodel $\hat{s}(x; y_0)$ that approximates $s(x; y_0)$. With $\hat{s}(x; y_0)$, the importance sampler becomes

$$q_X(x; y_0) = \frac{1}{C_q} f_X(x) \sqrt{\hat{s}(x; y_0)}, \quad (6)$$

where $C_q = \int_{\Omega_{\mathbf{X}}} f_X(x) \sqrt{\hat{s}(x; y_0)} dx$ (> 0) is the normalizing constant for $q(x; y_0)$. Here, to make the POE estimator unbiased, the support of $\hat{s}(x; y_0)$ should include the support of $s(x; y_0) = P(Y > y_0 \mid X = x)$. This condition can be readily satisfied when $\hat{s}(x; y_0)$ is strictly positive in $\Omega_{\mathbf{X}}$ [16].

Then one can sample X_i, $i = 1, 2, \cdots, n$, from q_X and run simulation at each X_i to obtain Y_i. Once the data is collected, the α-quantile estimator [2] can be obtained by

$$\hat{y}_{\alpha,n} = \inf\{y \geq y_0 : \hat{P}_n(y) \leq \alpha\}. \quad (7)$$

It has been shown that the POE estimator $\hat{P}_n(y)$ in (3) obeys the CLT as follows [16], [20].

$$\frac{\sqrt{n}}{\sigma_y}(\hat{P}_n(y) - p_y) \xrightarrow{d} N(0, 1), \quad (8)$$

where p_y denotes $P(Y > y)$ in (1) and σ_y^2 is the asymptotic variance of $\sqrt{n}\hat{P}_n(y)$ [16], [20].

The CLT for $\hat{P}_n(y)$ is built upon the average of independent samples, as shown in (3). However, the quantile estimator $\hat{y}_{\alpha,n}$ in (7) does not have such a natural sample average form. Section III derives the CLT for $\hat{y}_{\alpha,n}$ and various types of the quantile CI.

As a remark, Choe et al. [14] considered a more general POE estimator that allowed multiple runs at each sampled X_i. Specifically, this general framework runs simulation n_i times at each X_i. The POE estimator in (3) and the importance sampler in (6) is a special case by setting $n_i = 1$. When multiple runs are allowed, the resulting output samples are correlated, which makes the CI derivation extremely challenging. In this study we consider the case where we run simulation once at each input and estimate the extreme quantile using (7) with the POE defined in (3).

III. Methodology

This section establishes the CLT for the SIS quantile estimator to construct the CIs of extreme quantiles. We first derive a theoretically valid asymptotic CI in an explicit form and then present batching-based approaches.

A. Theoretical CI

Building a theoretical CI of some unknown quantity typically requires that its associated estimator takes the form of sample average and obeys the CLT. As discussed earlier, one of the difficulties in constructing a CI of a quantile is rooted from the fact that the quantile estimator does not take a form of sample average. However, by applying Taylor's expansion on the POE, we can relate the CLT for the quantile estimator to the CLT for the POE estimator.

Specifically, it holds

$$\begin{aligned} P(Y > \hat{y}_{\alpha,n}) =& P(Y > y_\alpha) + f_Y(y_\alpha)(\hat{y}_{\alpha,n} - y_\alpha) + R'_n. \\ =& \alpha + f_Y(y_\alpha)(\hat{y}_{\alpha,n} - y_\alpha) + R'_n, \end{aligned} \quad (9)$$

for $\hat{y}_{\alpha,n} \geq y_0$, where R'_n denotes a remainder. For large sample size n, by the Glivenko-Cantelli Theorem [21] we obtain

$$\hat{y}_{\alpha,n} = y_\alpha - \frac{\hat{P}_n(y_\alpha) - \alpha}{f_Y(y_\alpha)} + R_n, \quad (10)$$

Here R_n is a remainder. Note that we use a different remainder due to the replacement of $P(Y > \hat{y}_{\alpha,n})$ with $\hat{P}_n(y_\alpha)$ [12].

Recall that $\hat{P}_n(y_\alpha)$ asymptotically obeys the CLT under the SIS method, as shown in (8). Therefore, linking the probability estimation with the quantile estimation, Equation (10) implies that the quantile estimator $\hat{y}_{\alpha,n}$ can also follow the CLT if the remainder R_n vanishes for n sufficiently large. In particular, considering the CLT in (8), we need the following condition for R_n

$$\sqrt{n}R_n \xrightarrow{d} 0, \tag{11}$$

where $\xrightarrow{d}$ denotes the convergence in distribution. Equations (10) and (11) together are called a weak Bahadur representation in the literature [12], [19].

It has been shown that the Bahadur representation holds when the importance sampling density satisfies the condition $E_q[\mathbb{1}(Y > y_\alpha - \delta)L_Y^{2+\epsilon}] < \infty$ for some $\delta > 0$ and $\epsilon > 0$ where the likelihood ratio L_Y is defined in the domain of Y (see Theorems 3.1, 3.2 and 4.1 in [12]). In deterministic computer models that generate the same output Y given the same input X, it is straightforward to extend the result to show that the similar condition $E_q[\mathbb{1}(Y > y_\alpha - \delta)L_X^{2+\epsilon}] < \infty$ with L_X denoting a likelihood ratio of the input vector X guarantees the Bahadur representation.

The challenge in our case is that the simulation with stochastic computer models takes the two-level simulation (or nested simulation) procedure where the input X is sampled in the first level and then the output Y is randomly generated from the black box computer model, given X [16]. The reason of observing random outputs is that a random vector ξ is embedded inside the computer model [14]. Unlike the input X which has a known pdf $f_X(x)$, the density $f_{\xi|X}(\xi|x)$ of ξ, given $X = x$, is unknown, due to the simulator's black box nature (Figure 2). In other words, the density $f_{\xi|X}(\xi|x)$, which generates ξ, is hidden inside the computer model.

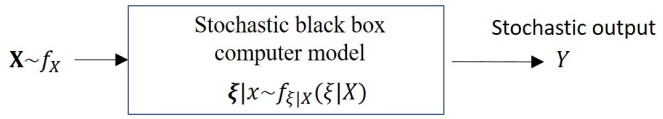

Fig. 2: Two-level simulation with stochastic black box computer model

Although we cannot sample ξ from $f_{\xi|X}(\xi|x)$, we note that the output Y becomes fixed, given x and ξ. Therefore, a stochastic computer model can be viewed as a special case of deterministic computer models with controllable input X and non-controllable hidden input ξ. With this insight, we will show that the condition for satisfying the Bahadur representation, which is $E_q[\mathbb{1}(Y > y_\alpha - \delta)L^{2+\epsilon}] < \infty$ for some $\delta > 0$ and $\epsilon > 0$, holds with the SIS density under some mild condition, where L is the likelihood ratio of the entire input vector X and ξ in this case.

With the two types of inputs, $E_q[\mathbb{1}(Y > y_\alpha - \delta)L^{2+\epsilon}]$ becomes

$$\begin{aligned} &E_q[I(Y > y_\alpha - \delta)L^{2+\epsilon}] \\ &= \int_{\Omega_\mathbf{x}} \int_{\Omega_\xi} I(Y > y_\alpha - \delta|x,\xi) \left(\frac{f_{X\xi}(x,\xi)}{q_{X\xi}(x,\xi)}\right)^{2+\epsilon} \\ &\qquad q_{X\xi}(x,\xi)d\xi dx, \end{aligned} \tag{12}$$

where $f_{X\xi}(x,\xi)$ is the original joint density of X and ξ, while $q_{X\xi}(x,\xi)$ denotes the joint importance sampling density.

Under the importance sampling scheme, the likelihood in (12) becomes

$$\frac{f_{X\xi}(x,\xi)}{q_{X\xi}(x,\xi)} = \frac{f_{\xi|X}(\xi|x)f_X(x)}{f_{\xi|X}(\xi|x)q_X(x;y_0)} = \frac{f_X(x)}{q_X(x;y_0)}. \tag{13}$$

Here, because $f_{\xi|X}$ is unknown, we cannot bias the conditional density. Instead we can bias the density of X only. So, at each x sampled from q_X, the simulator randomly generates ξ with $f_{\xi|X}$ which is hidden inside the black box computer model. This is why the joint importance sampling density $q_{X\xi}(x,\xi)$ in the denominator becomes $f_{\xi|X}(\xi|x)q_X(x;y_0)$.

Using the likelihood ratio in (13) and $q_{X\xi}(x,\xi) = f_{\xi|X}(\xi|x)q_X(x;y_0)$, we get

$$\begin{aligned} &E[I(Y > y_\alpha - \delta)L^{2+\epsilon}] \\ &= \int_{\Omega_\mathbf{x}} \int_{\Omega_\xi} I(Y > y_\alpha - \delta|x,\xi) \left(\frac{f_X(x)}{q_X(x;y_0)}\right)^{2+\epsilon} f_{\xi|X}(\xi|x)d\xi \\ &\qquad q_X(x;y_0)dx \end{aligned} \tag{14}$$

$$\begin{aligned} &= \int_{\Omega_\mathbf{x}} \left(\int_{\Omega_\xi} I(Y > y_\alpha - \delta|x,\xi) f_{\xi|X}(\xi|x)d\xi \right) \\ &\qquad \frac{f_X(x)^{2+\epsilon}}{q_X(x;y_0)^{1+\epsilon}}dx \end{aligned} \tag{15}$$

$$= \int_{\Omega_\mathbf{x}} P(Y > y_\alpha - \delta \mid x) \frac{f_X(x)^{2+\epsilon}}{q_X(x;y_0)^{1+\epsilon}}dx \tag{16}$$

$$= \int_{\Omega_\mathbf{x}} P(Y > y_\alpha - \delta \mid x) f_X(x) \left(\frac{f_X(x)}{q_X(x;y_0)}\right)^{1+\epsilon} dx, \tag{17}$$

for some $\delta > 0$ and $\epsilon > 0$, where Ω_ξ denotes the support of $f_{\xi|X}$.

Next we show that $E[I(Y > y_\alpha - \delta)L^{2+\epsilon}]$ is finite when the SIS density is used. We plug the SIS density $q_X(x;y_0)$ defined in (6) into $f_X(x)/q_X(x;y_0)$ to obtain

$$\frac{f_X(x)}{q_X(x;y_0)} = \frac{C_q}{\sqrt{\hat{s}(x;y_0)}}. \tag{18}$$

From (18), the condition becomes

$$E[I(Y > y_\alpha - \delta)L^{2+\epsilon}] \tag{19}$$

$$= \int_{\Omega_\mathbf{x}} P(Y > y_\alpha - \delta \mid x) f_X(x) \frac{C_q^{1+\epsilon}}{\hat{s}(x;y_0)^{\frac{1+\epsilon}{2}}}dx \tag{20}$$

$$\leq C_q^{1+\epsilon} \int_{\Omega_\mathbf{x}} f_X(x)\hat{s}(x;y_0)^{-\frac{1+\epsilon}{2}}dx, \tag{21}$$

where the last inequality holds because of $P(Y > y_\alpha - \delta \mid x) \leq 1$. Considering that the normalizing constant C_q is finite, $E[I(Y > y_\alpha - \delta)L^{2+\epsilon}]$ is bounded if

$$\int_{\Omega_\mathbf{x}} f_X(x)\hat{s}(x;y_0)^{-\frac{1+\epsilon}{2}}dx < \infty, \tag{22}$$

for some $\epsilon > 0$.

The condition in (22) can be easily satisfied in practice. We present a couple of cases where the condition can be met. The first case is when $\Omega_{\mathbf{X}}$ is bounded and closed (i.e., compact set) and $\hat{s}(x; y_0)$ is strictly positive and continuous. This condition is satisfied in the wind turbine application case study (to be detailed in Section V). When $\Omega_{\mathbf{X}}$ is bounded and closed and $\hat{s}(x; y_0)$ is strictly positive and continuous, the minimum value of $\hat{s}(x; y_0)$, denoted by $\hat{s}_{min}$, can be defined as

$$\hat{s}_{min} := \min_{x \in \Omega_{\mathbf{X}}} \hat{s}(x; y_0). \tag{23}$$

We then get

$$\int_{\Omega_{\mathbf{X}}} f_X(x)\hat{s}(x; y_0)^{-\frac{1+\epsilon}{2}} dx \leq \hat{s}_{min}^{-\frac{1+\epsilon}{2}} \int_{\Omega_{\mathbf{X}}} f_X(x)dx \tag{24}$$

$$= \hat{s}_{min}^{-\frac{1+\epsilon}{2}} \tag{25}$$

$$< \infty, \tag{26}$$

for some $\epsilon > 0$.

Second, without imposing any assumptions on $\Omega_{\mathbf{X}}$, the condition in (22) can be met as long as $\hat{s}(x; y_0)$ is bounded above some positive constant. To make $\hat{s}(x; y_0)$ strictly positive everywhere, a small number $s_0 (> 0)$ can be added to the metamodel. Let $\hat{s}'(x; y_0)$ denote an original matamodel that approximates $s(x; y_0)$. Then, $\hat{s}(x; y_0)$ can be defined as

$$\hat{s}(x; y_0) = \hat{s}'(x; y_0) + s_0. \tag{27}$$

Here it should be noted that, although $\hat{s}(x; y_0)$ can possibly exceeds 1 in (27), the importance sampler in (6) can be still well-defined thanks to the normalizing constant C_q. Then, because of $s(x; y_0) \geq s_0 \forall x \in \Omega_{\mathbf{X}}$, the condition in (22) is satisfied by following a procedure similar to (24)-(26).

In either case, the Bahadur representation holds, i.e., $\sqrt{n}R_n \xrightarrow{d} 0$ holds. Accordingly, the CLT of the probability estimation in (8) is translated to the CLT of the quantile estimation, making the asymptotic normality of the SIS quantile estimator valid. Consequently, from (10) with the $\sqrt{n}R_n \xrightarrow{d} 0$, we obtain

$$\frac{f_Y(y_\alpha) \cdot \sqrt{n}(\hat{y}_{\alpha,n} - y_\alpha)}{\sigma_y} \xrightarrow{d} N(0, 1). \tag{28}$$

By setting $\kappa_\alpha = \phi_{\alpha,n} \cdot \sigma_{y_\alpha}$ with $\phi_\alpha = 1/f_Y(y_\alpha)$, we get

$$\frac{\sqrt{n}}{\kappa_\alpha}(\hat{y}_{\alpha,n} - y_\alpha) \xrightarrow{d} N(0, 1). \tag{29}$$

Note that the pdf $f_Y(y_\alpha)$ of Y at y_α and the asymptotic variance $\sigma^2_{y_\alpha}$ in (29) are unknown. Therefore, we need to find consistent estimators of ϕ_α and $\sigma^2_{y_\alpha}$ to obtain the CI of a quantile. First, to handle ϕ_α, several consistent estimators have been suggested in the literature. One of the most widely employed estimators is the finite difference estimator, defined as

$$\hat{\phi}_{\alpha,n}(h_n) = \frac{\hat{y}_{\alpha+h_n,n} - \hat{y}_{\alpha-h_n,n}}{2h_n}, \tag{30}$$

where $h_n > 0$ is called the bandwidth parameter [19], [22]. It has been shown that $\hat{\phi}_{\alpha,n}(h_n)$ with h_n satisfying $1/h_n = O(\sqrt{n})$ is a consistent estimator of $\phi_{\alpha,n}(h_n)$ [12]. In defining h_n, Chu and Nakayama [12] use $h_n = cn^{-\nu}$, where c is a positive constant. Several variants of the finite difference estimator are also discussed in [12]. Instead of using the finite difference estimator, a kernel estimator can be used [23], [24].

Next, according to the studies in [12], [16], [20], $\sigma^2_{y_\alpha}$ can be consistently estimated by

$$\hat{\sigma}^2_{y_\alpha} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbb{I}\left(Y_i > \hat{y}_{\alpha,n} | X_i = x_i\right) L_i - \hat{P}_n(\hat{y}_{\alpha,n}))^2. \tag{31}$$

Then the product of $\phi_{\alpha,n}$ and σ_{y_α} in (29) can be consistently estimated by $\hat{\kappa} = \hat{\phi}_{\alpha,n}(h_n) \cdot \hat{\sigma}^2_{y_\alpha}$ by Slutsky's theorem, which make the CLT for $\hat{y}_{\alpha,n}$ in (29) hold. Consequently, assuming $f_Y(y_\alpha) > 0$, the CLT for $\hat{y}_{\alpha,n}$ holds under the condition in (22). That is,

$$\frac{\sqrt{n}}{\hat{\kappa}_\alpha}(\hat{y}_{\alpha,n} - y_\alpha) \xrightarrow{d} N(0, 1), \tag{32}$$

where $\hat{\kappa}_\alpha = \hat{\phi}_{\alpha,n}(h_n) \cdot \hat{\sigma}_{y_\alpha}$ with $\hat{\phi}_{\alpha,n}(h_n)$ defined in (30) for h_n satisfying $1/h_n = O(\sqrt{n})$ and $\hat{\sigma}^2_{y_\alpha}$ defined in (31). Lastly, the asymptotically valid $100(1-\beta)\%$ CI of y_α is given by

$$[\hat{y}_{\alpha,n} \pm z_{\beta/2}\hat{\kappa}_{\alpha,n}/\sqrt{n}]. \tag{33}$$

We summarize the implementation procedure in Algorithm 1.

Algorithm 1 Procedure for constructing a theoretical CI

1: Set input parameters: y_0, c, ν.
2: Sample x_i from $q(x; y_0)$ in (6) and run simulation at each x_i to generate y_i, $i = 1, \cdots, n$.
3: Sort y_i from the smallest to the largest. Let $y_{(i)}$ denote the i^{th} smallest output among n outputs.
4: Compute the POE estimate $\hat{P}_n(y_{(i)})$ in (3) at each $y_{(i)}$.
5: Obtain the α-quantile estimate $\hat{y}_{\alpha,n}$ in (7).
6: Obtain $\hat{\kappa}_\alpha = \hat{\phi}_{\alpha,n}(h_n) \cdot \hat{\sigma}_{y_\alpha}$ with $\hat{\phi}_{\alpha,n}(h_n)$ and $\hat{\sigma}^2_{y_\alpha}$ defined in (30) and (31), respectively.
7: Compute the α-quantile CI in (33).

In the first step of the algorithm, y_0 needs to be pre-specified for defining the importance sampler q_X in (6). To get the unbiased estimation, y_0 should be smaller than, or equal to, the target unknown extreme quantile y_α [2]. To set such y_0, one can use domain knowledge. Alternatively, the metamodel $\hat{s}(x; y)$ can be used to get a rough estimate for y_0. The value of y_0 affects the estimation efficiency in SIS. Properly defining its value is out of the scope of this paper, rather it is a subject of our future research.

For ν in the first step, to satisfy the condition for h_n, which is $1/h_n = O(\sqrt{n})$, we use $\nu = 0.5$ in our implementation. Deciding the appropriate value for c will be discussed in Sections IV and V. In the second step, for drawing samples from $q(x; y_0)$, the acceptance-rejection sampling method can be used [18].

Although the proposed asymptotic CI is theoretically valid and takes a closed form, the result is highly sensitive to the choice of h_n. A good choice of h_n depends on several factors, including the distribution of Y and α. Similar issues

arise with other finite difference estimators and the kernel estimator [19], [25], [26]. Our numerical experience indicates that even with carefully tuned h_n, the resulting asymptotic CI tends to be overly conservative with a large width. Due to the difficulty in finding an appropriate bandwidth parameter in the theoretical CI, we consider alternative approaches in the following section.

B. Batching-Based Approaches

Batching-based approaches have wide applicability thanks to their simple procedure [19]. However, their CIs are valid only when the normality assumption is satisfied. In our case, the asymptotic normality of the quantile estimator discussed in the previous section provides a theoretical basis for constructing batching-based CIs (Figure 1). This section presents the three batching-based CI procedures, namely, batching, sectioning and sectioning-batching [19].

First, batching randomly divides n output samples into b non-overlapping batches. The asymptotic justification of the batching-based CI is from the fact that SIS quantile estimator $\hat{y}_{\alpha,n}$ obeys CLT, as shown in (32). Let $r = n/b$ be the sample size of each batch. Here we consider equal-size batches for notation simplicity. When r is not integer, we can make small adjustment in our notations.

From each batch we obtain the SIS α-quantile estimator using r samples. Specifically, let $\hat{P}_{r,k}(y_\alpha)$ represent the SIS POE estimator with r samples from the k^{th} batch, i.e.,

$$\hat{P}_{r,k}(y_\alpha) = \frac{1}{r}\sum_{i=1}^{r} \mathbb{I}\left(Y_{i,k} > y_\alpha | X_{i,k}\right) \frac{f_X(X_{i,k})}{q_X(X_{i,k})}, \tag{34}$$

where $X_{i,k}$ and $Y_{i,k}$, respectively, represent the i^{th} input and output in the k^{th} batch. Similar to (7), we can obtain the SIS quantile estimator, denoted by $\hat{y}_{\alpha,r,k}$, from the k^{th} batch as

$$\hat{y}_{\alpha,r,k} = \inf\{y \in \mathbb{R} : \hat{P}_{r,k}(y) \le \alpha\}. \tag{35}$$

Based on the asymptotic normality of $\hat{y}_{\alpha,n}$, each batch quantile estimator $\hat{y}_{\alpha,r,k}$, $k = 1, 2, \cdots, b$, is also asymptotically normally distributed when the batch sample size r is sufficiently large (Figure 3). Therefore, the average of b batch quantile estimators is also asymptotically normal. Such property allows us to use the following sample mean and variance of quantile estimates from b batches as the point and variance estimates, respectively, in constructing a CI.

$$\hat{y}_{\alpha,b,bat} = \frac{1}{b}\sum_{k=1}^{b} \hat{y}_{\alpha,r,k}, \tag{36}$$

$$S^2_{b,bat} = \frac{1}{b-1}\sum_{k=1}^{b} (\hat{y}_{\alpha,r,k} - \hat{y}_{\alpha,b,bat})^2. \tag{37}$$

Because the bias in $\hat{y}_{\alpha,b,bat}$ diminishes as r gets large, $(\hat{y}_{\alpha,bat} - y_\alpha)/(S_{b,bat}/\sqrt{b})$ asymptotically follows the t distribution with $b-1$ degrees of freedom. That is,

$$\frac{\hat{y}_{\alpha,bat} - y_\alpha}{S_{b,bat}/\sqrt{b}} \sim t_{b-1} \tag{38}$$

for large batch size r. Accordingly we obtain the $100(1-\beta)\%$ batching CI of y_α as

$$CI_{b,bat} = \left(\hat{y}_{\alpha,b,bat} \pm t_{b-1,\beta/2}\frac{S_{b,bat}}{\sqrt{b}}\right). \tag{39}$$

The accuracy of batching CI highly depends on the batch size. Recall that the batching CI relies on the asymptotic normality of each batch's quantile estimator $\hat{y}_{\alpha,r,k}$ whose bias vanishes when the batch size $r = n/b$ is large. As such, a large batch size is required for ensuring the CLT. When r is small, the estimation bias could be significant, possibly causing poor CI coverage. To circumvent the limitation of batching when the batch size is small, sectioning modifies batching by replacing the batching point estimator $\hat{y}_{\alpha,b,bat}$ with the overall quantile estimator $\hat{y}_{\alpha,n}$. Specifically, we replace $\hat{y}_{\alpha,b,bat}$ with $\hat{y}_{\alpha,n}$ in both (36) and (37) to obtain the sectioning $100(1-\beta)\%$ CI as

$$CI_{b,sec} = \left(\hat{y}_{\alpha,n} \pm t_{b-1,\beta/2}\frac{\hat{S}_{b,sec}}{\sqrt{b}}\right) \tag{40}$$

with

$$S^2_{b,sec} = \frac{1}{b-1}\sum_{k=1}^{b} (\hat{y}_{\alpha,r,k} - \hat{y}_{\alpha,n})^2. \tag{41}$$

Note that the sectioning CI uses $\hat{y}_{\alpha,n}$ in (40) and (41), whereas the batching CI uses $\hat{y}_{\alpha,b,bat}$ in both central position and sample variance.

Because $\hat{y}_{\alpha,n}$ is a quantile estimator with a larger sample size, the sectioning approach can reduce the estimation bias, compared to batching. However, it has a drawback that its CI could be much wider than the batching CI when the individual batching estimator $\hat{y}_{\alpha,r,k}$, $r = 1, 2, \cdots, b$, is largely different from the overall estimator $\hat{y}_{\alpha,n}$.

To address the limitations of batching (large bias) and sectioning (large variance), sectioning-batching combines both sectioning and batching by taking the advantages of both approaches. Specifically, it uses $\hat{y}_{\alpha,n}$ as the CI center point

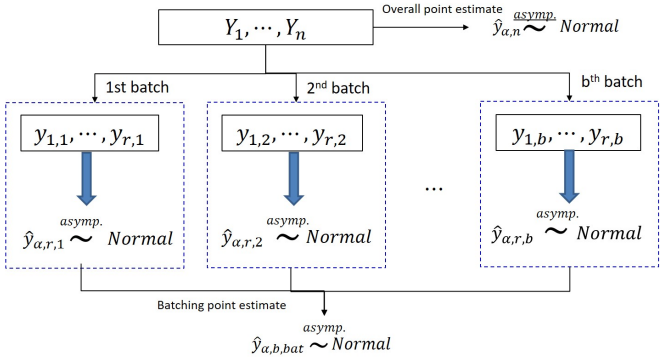

Fig. 3: Overview of batching (each batching quantile estimator $\hat{y}_{\alpha,r,k}$, $k = 1, 2, \cdots, b$, is asymptotically normally distributed when the batch size r is large, due to the CLT result discussed in Section III.A)

as in sectioning, while employing the batching sample variance $\hat{y}_{\alpha,b,bat}$ as in batching. The resulting sectioning-batching $100(1-\beta)\%$ CI is

$$CI_{b,sec-bat} = \left(\hat{y}_{\alpha,n} \pm t_{b-1,\beta/2} \frac{S_{b,bat}}{\sqrt{b}} \right). \tag{42}$$

Under the condition in (22), $E_q[\mathbb{1}(Y > y_\alpha - \delta)L^{2+\epsilon}(Y)] < \infty$ holds and thus, the coverage rate of quantile CIs in (39), (40) and (42) converge to the theoretical target coverage rate [19], that is,

$$P(y_\alpha \in CI) \to 1-\beta \tag{43}$$

as $r \to \infty$ with b fixed for the CI being CI_{bat}, CI_{sec} or $CI_{sec-bat}$ in (39), (40), and (42), respectively.

Algorithm 2 summarizes the implementation procedure for constructing the batching CI. The algorithms for the sectioning CI and sectioning-batching CI can be stated in a similar manner, but they are omitted to save space.

Algorithm 2 Procedure for constructing a batching CI

1: Set input parameter y_0.
2: Sample x_i, $i = 1, \cdots, n$, from $q(x; y_0)$ in (6) and run simulation at each x_i to generate y_i $(i = 1, \cdots, n)$.
3: Randomly divide the outputs into b batches as $(y_{1,1}, \cdots, y_{r,1})$, $(y_{1,2}, \cdots, y_{r,2})$, $\cdots$, $(y_{1,b}, \cdots, y_{r,b})$.
4: Obtain α-quantile estimate $\hat{y}_{\alpha,r,k}$ in (35) in each k^{th} batch, $k = 1, 2, \cdots, b$.
5: Obtain the batching CI in (39).

IV. Numerical Example

To evaluate the performance of proposed approaches, we slightly modify the numerical example presented in [14] and use the following data generating structure

$$Y|\mathbf{X} \sim N(\mu(\mathbf{X}), \sigma^2(\mathbf{X})) \tag{44}$$

with $\mu(\mathbf{X}) = 0.95\mathbf{X}^2(1 + 0.5\cos(10\mathbf{X}) + 0.5\cos(20\mathbf{X}))$ and $\sigma(\mathbf{X}) = 1 + 0.7|\mathbf{X}| + 0.4\cos(\mathbf{X}) + 0.3\cos(14\mathbf{X})$ and $\mathbf{X}$ following a truncated standard normal distribution in [-100,100]. In this example, it is assumed that we know the conditional POE $s(x; y_0)$ with $\mu(\mathbf{X})$ and $\sigma(\mathbf{X})$ with $y_0 = 3$ for defining the importance sampler q_X. In Section V, we estimate the conditional POE using a metamodel.

We first evaluate the CI estimation performance with $n = 1,000$. Later we conduct sensitivity analysis with different sample sizes. In constructing the theoretical CI, we need to set c and ν for defining the bandwidth $h_n = cn^{-\nu}$ in (30). To satisfy $1/h_n = O(\sqrt{n})$, we use $\nu = 0.5$, as in [19]. For c, $c = 0.1$ is used in [19]. However, it generates overly large CIs, so we use different c values (to be discussed later). In batching-based approaches, we first use $b = 10$ and investigate how the performance changes with different batch sizes.

A. Alternative approaches

We compare the theoretical and batching-based approaches with two alternative methods, bootstrapping and Jackknife [18]. First, bootstrapping resamples data from the original set $\{y_i\}_{i=1}^n$ with replacement and constructs a CI by finding estimates for Δ_1 and Δ_2, such that $P(\Delta_1 < \hat{y}_{\alpha,n} - y_\alpha < \Delta_2) = 1-\beta$ holds. Then $[\hat{y}_{\alpha,n} - \Delta_2, \hat{y}_{\alpha,n} - \Delta_1]$ becomes the $(1-\beta)100\%$ confidence interval for $\hat{y}_{\alpha,n}$.

Suppose T sets of bootstrapping samples are generated. Let $\hat{y}^{bs}_{\alpha,s,t}$ denote the quantile estimator for the t^{th} bootstrapped set of samples of size s, $t = 1, 2, \cdots, T$. Typically s is set to be equal to n. Let $\hat{\Delta}_1$ and $\hat{\Delta}_2$ be the $\beta/2$ lower and upper quantiles of $\{\hat{y}^{bs}_{\alpha,s,t} - \hat{y}_{\alpha,n}\}_{t=1,\cdots,T}$, respectively. Then the bootstrapping confidence interval is given by

$$CI_{bsp} := [\hat{y}_{\alpha,n} - \hat{\Delta}_2, \hat{y}_{\alpha,n} - \hat{\Delta}_1]. \tag{45}$$

In our implementation we use $T = 100$.

The asymptotic property of the CI of quantile from the bootstrapping approach has not been well studied. Liu and Yang [27] established the asymptotic distribution for bootstrapping extreme quantile variance estimation in importance sampling when the likelihood ratio function has a specific exponential form. However, theoretical results under the general form of importance sampling framework have not been fully developed in the literature.

Unlike the bootstrapping that resamples the output samples, Jackknife leaves one sample out and obtains the point quantile estimator as

$$\bar{J} = \frac{1}{n}\sum_{i=1}^{n} J_i, \tag{46}$$

with

$$J_i = n\hat{y}_{\alpha,n} - (n-1)\hat{y}_{\alpha,(i)}, \tag{47}$$

where $\hat{y}_{\alpha,(i)}$ is called a "leave-the-i^{th}-sample-out" estimate, i.e., the quantile estimate with samples $Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n$. Then the sample variance estimate in Jackknife is

$$S^2_{jac} = \frac{1}{n-1}\sum_{i=1}^{n} \left(J_i - \bar{J}\right)^2. \tag{48}$$

Finally, the Jackknife $100(1-\beta)\%$ CI has the form of

$$CI_{jac} := \left(\bar{J} \pm z_{\beta/2}\sqrt{\frac{S^2_{jac}}{n}} \right). \tag{49}$$

It has been known that Jackknife can reduce the estimation bias and thus, its asymptotic bias is smaller than the bias of the general quantile estimator [18]. However, the main issue is that its sample variance estimator is not consistent, because the Jackknife samples J_i's are highly correlated. In particular, J_i's can be similar in different i's, so the sample variance tends to be unduly underestimated, resulting in a narrow CI width with a low coverage rate. We will discuss the performance of these methods in details, as compared to the theoretical CI and batching-based CIs, in the next section.

B. Implementation results

Table I summarizes the results for the 95% CI from 1,000 independent experiments, including the average estimation error, average half CI width and coverage rate. The average estimation error is the averaged difference between the center value in the CI and the true quantile. Because the true quantile is unknown, we get 10^6 CMC samples and obtain the true quantiles for three different values of α, $\alpha = 0.1, 0.05, 0.01$, as $y_{0.1} = 3.77, y_{0.05} = 5.11$ and $y_{0.01} = 8.82$, respectively. The coverage rate is the proportion of the 1,000 CIs including the true quantile. Ideally, the coverage rate should be close to the normal coverage rate 95%.

TABLE I: Average point estimation error (Error), average half width (Half-width) and coverage rate (Coverage) of 95% quantile CI from 1,000 experiments

Method	Criteria	α		
		0.1	0.05	0.01
Theoretical approach	Error	0.026	-0.095	-0.058
	Half-width	2.691	1.831	0.810
	Coverage	99.9%	97.5%	96.6%
Batching	Error	0.091	0.083	0.240
	Half-width	0.177	0.204	0.508
	Coverage	92.4%	98.6%	97.7%
Sectioning	Error	0.026	-0.095	-0.058
	Half-width	0.555	0.743	1.675
	Coverage	100.0%	100.0%	98.9%
Sectioning-batching	Error	0.026	-0.095	-0.058
	Half-width	0.177	0.204	0.508
	Coverage	99.8%	100.0%	97.8%
Bootstrapping	Error	0.006	-0.240	-0.055
	Half-width	0.066	0.150	0.018
	Coverage	100.0%	0.1%	0.4%
Jackknife	Error	0.023	-0.096	-0.059
	Half-width	0.006	0.002	0.003
	Coverage	9.0%	0.3%	0.0%

The theoretical CI performance appears sensitive to the choice of $h_n = cn^{-\nu}$ in (30), in particular, the value for c (the results do not change significantly with different values of ν). When we use $c = 0.1$ as in [19], the resulting CIs are 13-27 times larger than those in the batching CI, depending on α. Therefore, we investigate the CI performance with different values of h_n in a wide range of $10^{-5} - 10^{-1}$ and obtain narrower CI widths when c is $10^{-4}, 10^{-4}$ and 5×10^{-3} for $\alpha = 0.1,\ 0.05$ and 0.01, respectively. Note that the chosen c values do not exhibit any systematic trend. The third to fifth rows of Table I report the theoretical CI results with the chosen c values. We also investigate other approaches for defining the bandwidth, including the weighted sum of the forward and central finite difference estimators [12], but do not get better results in constructing the theoretical CIs.

We summarize the comparison results among different approaches as follows.

- Theoretical CI: The coverage rate of the asymptotic CI is close to the nominal rate. However, even with the tuned bandwidth parameters, the asymptotic CI widths are wider, compared with those from other approaches, which are less informative.
- Batching: Overall, batching provides reasonable coverage rates, but its estimation error is generally larger than that from sectioning. In this example, batching uses $r = 100$ data points in each batch. As a result, the batch point estimator $\hat{y}_{\alpha,b,bat}$ tends to yield a larger bias, compared to the overall quantile estimator that uses $n = 1,000$. Such larger bias causes slightly lower coverage rates, compared to sectioning and sectioning-batching CIs.
- Sectioning: Sectioning uses the overall quantile estimator $\hat{y}_{\alpha,n}$ in both center point and sample variance. It provides more accurate point estimates, because it uses a large number of samples, compared to batching. However, it yields larger sample variances, resulting in wider CIs. In this example, the sectioning CI is about three times wider than the batching CI. It is because each batch estimator $\hat{y}_{\alpha,b,bat}$ is largely different from the overall estimator $\hat{y}_{\alpha,n}$.
- Sectioning-batching: Sectioning-batching generally outperforms the other two batching-based methods by employing the point quantile estimator from sectioning and sample variance estimator from batching. Its CI widths are the same as those from batching, but its coverage rates are higher. This result indicates that the sectioning-batching overcomes the limitations of batching and sectioning.
- Bootstrapping: Bootstrapping does not provide consistent results. It generates a narrow CI width with a high coverage rate for $\alpha = 0.1$. However, for other α values, its performance rapidly deteriorates. For $\alpha = 0.05$, it generates a large estimation error, whereas its narrow CI width makes its coverage rate greatly decrease for $\alpha = 0.01$.
- Jackknife: Jackknife's point estimation error is small, but its CI is overly narrow, causing poor coverage rates. This is likely due to the fact that the Jackknife samples are highly correlated.

It is also interesting to compare the patterns of CI width with different α values. As we estimate more extreme quantiles with smaller α, the estimation uncertainty increases and thus, the CI width is expected to increase. The batching-based approaches show such increasing pattern. However, the theoretical CI, bootstrapping and Jackknife do not show any specific trend. The width of the theoretical CI depends on the bandwidth parameter. Because we use different bandwidth parameters to attain reasonable CI widths, no trend is observed. In bootstrapping, the difference from the bootstrapping quantile estimate and the overall estimate does not necessary increase as α gets smaller. Rather, it depends on data in each bootstrap, which are resampled from original outputs. Similarly, the sampling variance of J_i's in Jackknife does not necessarily increase as α gets smaller.

In summary, among different approaches for constructing the CIs of quantiles, the batching and sectioning-batching methods outperform other methods in terms of the CI width and coverage rate. Between these two methods, the sectioning-batching provides smaller point estimation errors and slightly higher coverage rates, while its CI width remains the same as the batching's CI width.

C. Sensitivity Analysis

This section performs sensitivity analysis with different settings. Specifically, we investigate the CI performance with different sample sizes in all methods and with different batch sizes in batching-based methods.

First, Table II shows the 95% CI results with $n = 500, 1,000$ and $5,000$ for $\alpha = 0.01$, obtained from 1,000 independent experiments. In constructing the theoretical CI, we tune c values and use 7×10^{-3}, 5×10^{-3}, 3×10^{-2} for $n = 500$, $1,000$ and $5,000$, respectively, for identifying the appropriate bandwidth h_n. Nevertheless, the theoretical CI widths are wider in all different n's, compared with other approaches. Moreover, the CI width does not necessarily get narrower with a larger n, because different bandwidth parameters are used for different n's to attain narrow CIs in our implementation.

TABLE II: Sensitivity analysis with different sample sizes for $\alpha = 0.05$

Method	Criteria	Sample size (n)		
		500	1,000	5,000
Theoretical approach	Error	-0.144	-0.095	0.009
	Half-width	3.996	1.831	10.636
	Coverage	95.9%	97.5%	100.0%
Batching	Error	0.189	0.083	0.045
	Half-width	0.490	0.204	0.173
	Coverage	99.6%	98.6%	100.0%
Sectioning	Error	-0.144	-0.095	0.009
	Half-width	1.667	0.743	0.526
	Coverage	100.0%	100.0%	100.0%
Sectioning-batching	Error	-0.144	-0.095	0.009
	Half-width	0.490	0.204	0.173
	Coverage	94.8%	100.0%	100.0%
Bootstrapping	Error	-0.303	-0.240	-0.012
	Half-width	0.180	0.150	0.000
	Coverage	0.0%	0.1%	0.0%
Jackknife	Error	-0.148	-0.096	0.006
	Half-width	0.007	0.002	0.005
	Coverage	1.1%	0.3%	15.8%

On the contrary, the CI width gets narrower in batching-based approaches as n increases, implying that uncertainty can be reduced with a larger sample size. The CI coverage rates are generally close to the nominal value, 95%, with different n's, although they are slightly larger than 95% in most cases. Among them, the sectioning-batching method provides the narrow CI width while maintaining its coverage rate close to the nominal rate.

When bootstrapping is used, its estimation error is large when n is small (e.g., $n = 500$ and $1,000$). Moreover, its CI widths are narrower than those from sectioning-batching in all cases. As a result, it produces very low coverage rates. Similarly, Jackknife CI has the lowest coverage rate in all n values, because the sample variance estimation is overly small, leading to a very narrow CI.

Next, Table III shows the 95% CI estimation results for $\alpha = 0.05$ using batching-based approaches with different batch sizes. We note that the coverage rate of batching slightly gets deteriorated when b increases. This is because each batch contains a smaller number of data points with a larger b, leading to higher bias in the batch quantile estimator. With $b = 20$, the batching estimation error gets three times larger than that with $b = 10$. Unlike batching, the CI coverage rate in sectioning-batching does not deteriorate with a larger b. This is because, by using the overall point quantile estimate as the center point in the CI, the bias in sectioning-batching is not affected by the number of batches.

TABLE III: Sensitivity analysis with different batch sizes for $\alpha = 0.05$

Method	Criteria	Number of batches (b)	
		10	20
Batching	Error	0.083	0.240
	Half-width	0.204	0.355
	Coverage	98.6%	94.7%
Sectioning	Error	-0.095	-0.095
	Half-width	0.743	1.707
	Coverage	100.0%	100.0%
Sectioning-batching	Error	-0.095	-0.095
	Half-width	0.204	0.355
	Coverage	100.0%	99.0%

In sectioning, the CI width increases greatly as b increases, whereas the changes are less significant in batching and sectioning-batching. Recall that sectioning uses the overall point estimate when calculating the sample variance. With a large number of batches, the quantile estimate from each small-size batch can substantially deviate from the overall quantile point estimate, leading to an increased sample variance in sectioning. On the contrary, in batching and sectioning-batching, the sample variance of quantile estimators from the batches is more stable with different b's. As a result, the CI widths in batching and sectioning-batching do not change significantly.

In summary, among all studied methods, sectioning-batching generates most satisfactory results. Its performance is robust to the sample size and batch size.

V. Case Study

We apply the studied methods to construct the CIs of extreme load responses in a wind turbine. In this case study, we use the NREL's aeroelastic simulators, TurbSim [5] and FAST [6]. Specifically, we use 10-minute average wind speed as a simulation input. Among several design load cases (DLCs) in IEC 61400-1 [4], DLC 1.1 specifies the input wind condition between the cut-in and cut-out wind speeds under which a turbine normally operates. According to the IEC design standard [4], we employ a truncated Rayleigh distribution on $[3, 25]$ (m/s) with the scale parameter of $\sqrt{2/\pi} \cdot 10$.

After feeding the sampled wind speed into TurbSim [5], TurbSim generates time series of wind profile and passes this profile into FAST [6] to generate stochastic load responses. Each simulation run takes about 1 minute. In this study, we consider flapwise bending moment as simulation output response, which is considered to be an important load type in the wind turbine reliability analysis [28], [29].

Because the simulators are treated as black box computer models, the conditional POE $s(\mathbf{x}; y_0)$ is unknown. We estimate $s(\mathbf{x}; y_0)$ by fitting a non-homogeneous generalized extreme

value (GEV) distribution with a small pilot sample consisting of 600 samples where its location and scale parameter functions are modeled with cubic smoothing spline functions, so $\hat{s}(x; y_0)$ is strictly positive and continuous. The detailed procedure of obtaining $\hat{s}(\mathbf{x}; y_0)$ is available in [2], [14].

In this case study, $\Omega_{\mathbf{X}}$ is bounded and closed in $[3, 25]$ and the metamodel with the nonhomogeneous GEV distribution is strictly continuous and positive, satisfying the condition for holding the Bahadur representation. Other types of metamodeling techniques with different parametric or non-parametric functions can be alternatively used to obtain $\hat{s}(x; y_0)$ [30]. Because $\hat{s}(x; y_0)$ represents the estimated conditional failure probability, it can be easily formulated as a strictly positive and continuous function.

In our implementation we use $n = 30,000$ sample to build the CI of extreme load at $\alpha = 1/1,000, 1/3,000$ and $1/5,000$ level. In defining the SIS density in (6), we use $y_0 = 14,600$. In obtaining the theoretical CIs, we tune the values of c in the bandwidth to get narrow CI widths. In batching-based approaches we use $b = 10$ as in the numerical example, whereas we use $T = 100$ in bootstrapping.

A. Implementation results

Table IV shows the results for 95% CIs obtained from 25 independent experiments. In the numerical example, we estimate the true quantile from 10^6 CMC samples and compute the coverage from 1,000 experiments. In this case study, we cannot perform such extensive experiments due to the limited computational resource available to us. In Table IV, we report the average point estimates and half widths obtained from 25 independent experiments.

TABLE IV: Implementation results for 95% CI of flapwise bending moment (unit: kNm)

Method	Criteria	α		
		$1/1,000$	$1/3,000$	$1/5,000$
Theoretical approach	Point Estimate	15,000	15,233	15,370
	Half-width	8,103	4,922	3,687
Batching	Point Estimate	14,959	15,286	15,384
	Half-width	63	109	139
Sectioning	Point Estimate	15,000	15,233	15,370
	Half-width	214	352	424
Sectioning-batching	Point Estimate	15,000	15,233	15,370
	Half-width	63	109	139
Bootstrapping	Point Estimate	15,071	15,268	15,403
	Half-width	54	62	119
Jackknife	Point Estimate	15,000	15,225	15,340
	Half-width	0	16	59

Overall we obtain narrow CI widths in all methods except the theoretical CI. It is because we use a large sample size, $n = 30,000$, in this case study. However, even with this large sample size, the theoretical CI widths are overly large for all α values. To choose appropriate bandwidths, from the wide range of c, we choose $c = 7 \times 10^{-2}$, 1×10^{-3}, and 2×10^{-3} for $\alpha = 1/1,000$, $1/3,000$ and $1/5,000$, respectively. Even with these tuned bandwidths, the theoretical CI widths are still much larger than those from other approaches.

Even though we do not know the exact quantile value y_α in this case study, sectioning and sectioning-batching may provide more accurate point estimates, because they use the overall quantile estimate as the CI center value. Moreover, their point estimates are close to Jackknife's. Considering Jackknife provides a small estimation bias, the bias of sectioning and sectioning-batching is likely smaller than the bias of batching. However, the sectioning CI widths are larger than those of batching, whereas sectioning-batching reduces the CI width by using the batching sampling variance. Bootstrapping has narrower CI widths in this case, however, according to the analysis in the numerical example, it does not guarantee a good CI coverage in general. Jackknife has very narrow CI widths, which unlikely produce reasonable and stable coverage rates.

In summary, the results in this case study echo what we observe in the numerical example in Section IV. Sectioning-batching appears to provide the most stable results among all studied approaches. In addition, bootstrapping and Jackknife are computationally expensive with a large sample size.

B. Comparison with CMC

We further compare the CI from SIS with that from CMC. In CMC we sample input from the original distribution, i.e., truncated Rayleigh distribution in this case study. We use the same computational resource of $n = 30,000$ in CMC as in SIS. In both SIS and CMC, we use sectioning-batching to obtain CIs. Figure 4 demonstrates the 95% sectioning-batching CI using SIS (darker area) and CMC (lighter area). The CI from SIS sectioning-batching lies inside the CMC's with much narrower width across different α values in the y-axis. In general, the CMC's CI width for the flapwise bending moment is about 2 times wider than that from SIS, indicating that SIS can reduce extreme load estimation uncertainties.

Moreover, with the same computational resource, SIS can estimate the CI of the extreme load response at smaller α levels than CMC. For example, with $n = 30,000$ and $b = 10$, the sectioning-batching with CMC can obtain the CI of α equal to, or larger than, $1/3000$, while SIS can obtain the CI of the extreme quantile associated with $\alpha = 2 \times 10^{-4}$, as shown in Figure 4.

It should be noted that the sample size n should be large enough to obtain the quantile estimate (i.e, $\hat{y}_{\alpha,n}$ in the theoretical approach, or $\hat{y}_{\alpha,r,k}$ in the batching-based approaches). If the quantile estimate is not obtained due to the small sample size, additional computational runs are needed. For example, to obtain a quantile estimate and its CI at α smaller than 2×10^{-4} in our case study, we need to increase the sample size n.

However, the required sample size for getting quantile estimates in SIS is less than that in CMC. For example, to get the quantile estimate at $\alpha = 2 \times 10^{-4}$ with sectioning-batching, CMC needs at least $50,000 (= b/(2 \times 10^{-4}))$ runs for $b = 10$. Even with $n = 50,000$, the resulting CI from CMC would be much wider than that from the proposed approach. The computational advantage of SIS is due to the fact that the input vectors are sampled from the region where $q_{\mathbf{X}}$ is high. When $f_{\mathbf{X}}$ and $q_{\mathbf{X}}$ are substantially different, the likelihood ratio $f_{\mathbf{X}}/q_{\mathbf{X}}$ becomes smaller than 1 in most sampled inputs, which allows us to obtain smaller POE [2].

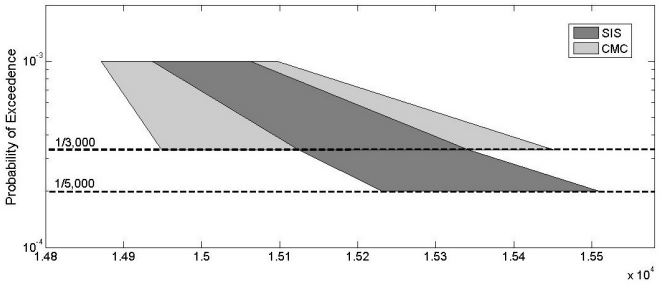

Fig. 4: 95% CI of extreme quantile for wind turbine flapwise bending moment (unit: kNm) using sectioning-batching in SIS and CMC (x-axis: y_α, y-axis: α)

VI. Summary

This study examines multiple approaches for constructing the quantile CI when importance sampling is applied to stochastic computer models. We verify the asymptotic normality of the SIS quantile estimator under some mild condition and derive an explicit formula for the theoretical CI in a closed form. The theoretical validity of the quantile estimator allows us to build the CIs from the three batching-based approaches, namely, batching, sectioning, sectioning-batching.

The CI estimation performance of the studied approaches is examined through the numerical example and wind turbine case study. The results consistently show that sectioning-batching outperforms other approaches. Compared with the theoretical method, batching-based approaches avoid the necessity of parameter tuning. In particular, the sectioning-batching method takes advantage of both sectioning and batching and thus, gives better CI performance than the other methods. Our implementation results also demonstrate that SIS can greatly reduce estimation uncertainty over CMC and that it can construct the CI of more extreme quantiles associated with smaller failure probability, compared to CMC.

In the future, we will investigate alternative procedures for estimating the variance constant in the theoretical CI. Compared to the batching-based approaches, the theoretical approach can construct the CI of the extreme quantile at smaller α levels, because it uses a larger number of samples. As such, a well-tuned theoretical CI could be more beneficial. Another possible approach is to apply the Knight's identity to our problem [31]. The Knight's identity has been used for deriving the convergence property of the regression parameters in a quantile regression where the conditional quantile of y given $\mathbf{X}$ is of interest [32]. Noting that the failure probability estimator takes the form of sample average, we will explore the possibility of using the Knight identity by exploiting the relationship between the probability estimation and quantile estimation.

Moreover, we plan to extend the SIS method to combine with other variance reduction techniques such as stratified sampling and control variate. Even though the sectioning-batching provides satisfactory results, its coverage rate is higher than the nominal rate in most cases, which indicates that its CI width could be further reduced. By combining with other variance reduction techniques, we hope to further reduce the estimation uncertainty. Finally, the current form of the SIS density may face challenges for problems with high dimensional inputs, because it is hard, if not impossible, to find a good metamodel and to sample directly from $q_{\mathbf{X}}$. We will study alternative methods, such as cross-entropy method [33], [34] and non-parametric approach [30].

References

[1] D. Grabaskas, M. K. Nakayama, R. Denning, and T. Aldemir, "Advantages of variance reduction techniques in establishing confidence intervals for quantiles," *Reliab. Eng. & Syst. Safety*, vol. 149, pp. 187 – 203, 2016.

[2] Y. Choe, Q. Pan, and E. Byon, "Computationally efficient uncertainty minimization in wind turbine extreme load assessments," *ASME J. Sol. Energy. Eng.*, vol. 138, no. 4, pp. 041 012–041 020, 2016.

[3] G. Lee, E. Byon, L. Ntaimo, and Y. Ding, "Bayesian spline method for assessing extreme loads on wind turbines," *Ann. Appl. Stat.*, vol. 7, no. 4, pp. 2034–2061, 2013.

[4] "International electrotechnical commission, wind turbines - part 1: Design requirements, IEC/TC88,61400-1 ed.3," 2005.

[5] B. J. Jonkman, "Turbsim user's guide: Version 1.50," National Renewable Energy Laboratory, Tech. Rep., 2009.

[6] J. M. Jonkman and M. L. Buhl, "Fast user's guide, No. NREL/EL-500-38230," National Renewable Energy Laboratory, Tech. Rep., 2005.

[7] L. S. Bastos and A. O'Hagan, "Diagnostics for Gaussian process emulators," *Technometrics*, vol. 51, no. 4, pp. 425–438, 2009.

[8] S. Ba and V. R. Joseph, "Composite Gaussian process models for emulating expensive functions," *Ann. Appl. Stat.*, vol. 6, no. 4, pp. 1838–1860, 2012.

[9] J. Fogle, P. Agarwal, and L. Manuel, "Towards an improved understanding of statistical extrapolation for wind turbine extreme loads," *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, vol. 11, no. 6, pp. 613–635, 2008.

[10] C. Cannamela, J. Garnier, and B. Iooss, "Controlled stratification for quantile estimation," *Ann. Appl. Stat.*, vol. 2, no. 4, pp. 1554–1580, 2008.

[11] I. Romani de Oliveira, J. Musiak, and J. Musiak, "A method for estimating the probability of extremely rare accidents in complex systems," *IEEE Trans. Reliab.*, vol. 68, no. 2, pp. 583–598, 2019.

[12] F. Chu and M. K. Nakayama, "Confidence intervals for quantiles when applying variance-reduction techniques," *ACM Trans. Model. Comput. Simul.*, vol. 22, no. 10, pp. 10.1–10.25, 2012.

[13] P. L'Ecuyer, G. Rubino, S. Saggadi, and B. Tuffin, "Approximate zero-variance importance sampling for static network reliability estimation," *IEEE Trans. Reliab.*, vol. 60, no. 3, pp. 590–604, 2011.

[14] Y. Choe, E. Byon, and N. Chen, "Importance sampling for reliability evaluation with stochastic simulation models," *Technometrics*, vol. 57, no. 3, pp. 351–361, 2015.

[15] S. Sankararaman, M. J. Daigle, and K. Goebel, "Uncertainty quantification in remaining useful life prediction using first-order reliability methods," *IEEE Trans. Reliab.*, vol. 63, no. 2, pp. 603–619, 2014.

[16] Y. Choe, H. Lam, and E. Byon, "Uncertainty quantification of stochastic simulation for black-box computer experiments," *Meth. Comp. Appl. Prob.*, vol. 20, no. 4, pp. 1155–1172, 2018.

[17] L. Sun and L. J. Hong, "Asymptotic representations for importance-sampling estimators of value-at-risk and conditional value-at-risk," *Oper. Res. Lett.*, vol. 38, no. 4, pp. 246–251, 2010.

[18] S. Asmussen and P. W. Glynn, *Stochastic simulation: algorithms and analysis*. Springer Science and Business Media, 2007.

[19] M. K. Nakayama, "Confidence intervals for quantiles using sectioning when applying variance-reduction techniques," *ACM Trans. Model. Comput. Simul.*, vol. 24, no. 19, pp. 19:1–19:21, 2014.

[20] Y. Choe, "Computationally efficient reliability evaluation with stochastic simulation models," Ph.D. dissertation, University of Michigan, 2015.

[21] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*. Wiley-Interscience, 1981.

[22] P. Glasserman, *Monte Carlo Methods in Financial Engineering*. Springer, New York, 2004.

[23] M. K. Nakayama, "Asymptotic properties of Kernel density estimators when applying importance sampling," *Proceedings of the 2011 Winter Simulation Conference*, pp. 556–568, 2011.

[24] M. P. Wand and M. C. Jones, *Kernel smoothing*. Chapman and Hall/CRC, 1994.

[25] P. W. Glynn, "Importance sampling for Monte Carlo estimation of quantiles," *Proceedings of the Second International Workshop on Mathematical Methods in Stochastic Simulation and Experimental Design*, pp. 180–185, 1996.
[26] P. Glasserman, P. Heidelberger, and P. Shahabuddin, "Variance reduction techniques for estimating value-at-risk," *Manag. Sci.*, vol. 46, no. 10, pp. 1349–1364, 2000.
[27] J. Liu and X. Yang, "The convergence rate and asymptotic distribution of the bootstrap quantile variance estimator for importance sampling," *Adv. Appl. Prob.*, vol. 44, no. 3, pp. 815–841, 2012.
[28] E. Byon, Y. Choe, and N. Yampikulsakul, "Adaptive learning in time-variant processes with application to wind power systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 2, pp. 997–1007, Apr. 2016.
[29] M. Barone, J. Paquette, B. Resor, and L. Manuel, "Decades of wind turbine load simulation," in *50th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition*, 2012.
[30] J. Morio, "Extreme quantile estimation with nonparametric adaptive importance sampling," *Simul. Model. Pract. and Theory*, vol. 27, pp. 76–89, 2012.
[31] K. Knight, "Limiting distributions for l_1 regression estimators under general conditions," *Ann. Statist.*, vol. 26, no. 2, pp. 755–770, 1998.
[32] R. Koenker, *Quantile Regression*, ser. Econometric Society Monographs. Cambridge University Press, 2005.
[33] N. Kurtz and J. Song, "Cross-entropy-based adaptive importance sampling using Gaussian mixture," *Struct. Safety*, vol. 42, pp. 35–44, 2013.
[34] Q. D. Cao and Y. Choe, "Cross-entropy based importance sampling for stochastic simulation models," *Reliab. Eng. & Syst. Safety*, vol. 191, p. 106526, 2019.

Qiyun Pan received her Ph.D. in Industrial & Operations Engineering Department from the University of Michigan, Ann Arbor, MI, USA in 2019. Her research interests include simulations with a focus on adaptive variance reduction methods and uncertainty quantification, as well as predictive modeling with machine learning methods.

Young Myoung Ko received his B.S., and M.S. degrees in Industrial Engineering from Seoul National University, Korea, and his Ph.D. degree in Industrial Engineering from Texas A&M University, United States. As an associate professor in the Department of Industrial and Management Engineering at Pohang University of Science and Technology (POSTECH), Korea, he focuses on the reliability and maintenance optimization of large-scale stochastic systems such as service systems, telecommunication networks, ICT infrastructure, and renewable energy systems.

Eunshin Byon received the B.S., and the M.S. degrees in industrial and systems engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejon, Korea, and the Ph.D. degree in industrial and systems engineering from Texas A&M University, College Station, TX, USA. She is an Associate Professor with the Department of Industrial and Operations Engineering at the University of Michigan, Ann Arbor, MI, USA. Her research interests include data analytics, industrial informatics, quality and reliability engineering, and uncertainty quantification. Prof. Byon is a member of IIE, IEEE, and INFORMS.