Received: 27 June 2019 | Revised: 26 February 2020 | Accepted: 8 July 2020

DOI: 10.1002/nav.21938

RESEARCH ARTICLE

WILEY

Adaptive importance sampling for extreme quantile estimation with stochastic black box computer models

Qiyun Pan[1] | **Eunshin Byon**[1] | **Young Myoung Ko**[2] | **Henry Lam**[3]

[1]Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan

[2]Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, South Korea

[3]Department of Industrial Engineering and Operations Research, Columbia University, New York, New York

Correspondence

Eunshin Byon, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan, 48109.
Email: ebyon@umich.edu

Funding information

National Science Foundation, Division of Information and Intelligent Systems, Grant/Award Numbers: IIS-1741166, IIS-1849280. National Science Foundation, Division of Civil, Mechanical and Manufacturing Innovation, Grant/Award Number: CMMI-1834710. National Research Foundation of Korea (NRF), Basic Science Research Program, Grant/Award Number: NRF-2016R1D1A1B04933453.

Abstract

Quantile is an important quantity in reliability analysis, as it is related to the resistance level for defining failure events. This study develops a computationally efficient sampling method for estimating extreme quantiles using stochastic black box computer models. Importance sampling has been widely employed as a powerful variance reduction technique to reduce estimation uncertainty and improve computational efficiency in many reliability studies. However, when applied to quantile estimation, importance sampling faces challenges, because a good choice of the importance sampling density relies on information about the unknown quantile. We propose an adaptive method that refines the importance sampling density parameter toward the unknown target quantile value along the iterations. The proposed adaptive scheme allows us to use the simulation outcomes obtained in previous iterations for steering the simulation process to focus on important input areas. We prove some convergence properties of the proposed method and show that our approach can achieve variance reduction over crude Monte Carlo sampling. We demonstrate its estimation efficiency through numerical examples and wind turbine case study.

KEYWORDS

Monte Carlo sampling, reliability, variance reduction

1 | INTRODUCTION

This study concerns the quantile estimation of an output of interest in a system using stochastic computer models, which can help determine an important design parameter of a system. In particular, this study is motivated by estimating extreme load responses in a wind turbine (Ragan & Manuel, 2008). To avoid catastrophic failures of the wind turbine structure, the International Electrotechnical Commission (IEC)'s design standard requires estimating extreme load responses imposed on turbine subsystems such as blades (IEC, 2005). At the design stage, wind turbine manufacturers can install a prototype turbine to collect data, but doing so is very expensive and time-consuming (Lee, Byon, Ntaimo, & Ding, 2013). Recent advancements in numerical computer modeling provide opportunities to quantify load responses and their variability. For example, an aeroelastic simulator has been developed by the U.S Department of Energy's National Renewable Energy Laboratory (NREL) to help design reliable turbines (B. J. Jonkman, 2009; J. M. Jonkman & Buhl, 2005).

Simulating the load response with the NREL simulator uses a nested procedure where a random input (e.g., wind speed), $\mathbf{X} \in \mathbb{R}^p$, is first generated from its prespecified probability density function (pdf), $p(\mathbf{x})$, and then fed into the simulator to generate the load response (e.g., blade bending moment), Y (Choe, Lam, & Byon, 2018). The NREL simulator uses a stochastic (or noisy) computer model which generates random outputs even at the same input. This is because it embeds a high-dimensional random vector, $\boldsymbol{\xi}$, inside the simulator to generate stochastic turbulence around rotor blades (B. J. Jonkman, 2009; J. M. Jonkman & Buhl, 2005). The embedded ξ may, or may not, depend on $\mathbf{X}$. In either case, ξ is hidden inside the black box computer model and thus, one cannot sample ξ from its distribution, but can sample $\mathbf{X}$ only from $p(\mathbf{x})$. Related types of simulation models also arise in several

TABLE 1 Computer experiments for black box computer models, serving different purposes

	Emulator (metamodeling)	Reliability analysis
Deterministic black box computer model	Emulators such as GP (Ba & Joseph, 2012; Bastos & O'Hagan, 2009; Oakley, 2004; Ranjan et al., 2008; Yang et al., 2007)	Importance sampling and other variance reduction techniques (Cannamela et al., 2008; Chu & Nakayama, 2012; Glynn, 1996; Hesterberg, 1995; Kurtz & Song, 2013; Neddermeyer, 2009; Zhang, 1996)
Stochastic black box computer model	GP with nugget effect, stochatic krigging (Ankenman et al., 2010; Binois et al., 2019; Chen et al., 2012; Wang & Hu, 2015)	Stochastic importance sampling (Choe, Byon, & Chen, 2015)

other applications (Ankenman, Nelson, & Staum, 2010; Shi & Chen, 2018; Sun, Apley, & Staum, 2011).

When a system response depends on the probabilistic input condition, $\mathbf{X}$, the failure probability, $\mathbb{P}(Y > y)$, is generally expressed as

$$\mathbb{P}(Y > y) = \int_{\mathbf{X}} \mathbb{P}(Y > y|\mathbf{X} = \mathbf{x})p(\mathbf{x})dx. \quad (1)$$

Here, $p(\mathbf{x})$ is assumed to be known. At the design stage, $p(\mathbf{x})$ is often specified in the design standard (IEC, 2005). This failure probability is also called the probability of exceedance (POE).

Given a prespecified failure probability, α, the $(1-\alpha)$-quantile is defined as

$$y_\alpha = \inf\{y : \mathbb{P}(Y > y) \le \alpha\}, \quad (2)$$

where "inf" represents the infimum. In the reliability analysis, y_α implies a resistance level for guaranteeing a failure probability, α. For designing a highly reliable system, it is crucial to accurately estimate the resistance level that can satisfy a target failure probability. For estimating y_α, one needs to accurately estimate the tail distribution. This type of problems is inherently challenging, because the simulator output is stochastic, the density of Y is unknown, the input-output relationship is complex and cannot be prescribed analytically due to the black box nature, and running the simulator takes time.

In the computer experiment literature, emulator-based approaches are commonly used (Ba & Joseph, 2012; Bastos & O'Hagan, 2009; Oakley, 2004; Ranjan, Bingham, & Michailidis, 2008; Yang, Ankenman, & Nelson, 2007). Recently, Gaussian process (GP) modeling, or stochastic Kriging, becomes the most common among many different choices of metamodeling approaches with stochastic computer models. Wang and Hu (2015) show that the prediction performance of stochastic Kriging, measured by the mean squared error (MSE), monotonically improves as the number of sampling points increases in a sequential computer experimental setting. Stochastic kriging is also employed in Chen, Nelson, and Kim (2012) for estimating the conditional value-at-risk. Binois, Huang, Gramacy, and Ludkovski (2019) further develop a new algorithm that sequentially decides sampling points for obtaining a globally accurate GP metamodel where the accuracy is defined with the integrated MSE. Other nonparametric approaches have been also studied. Hong, Juneja, and Liu (2017) use the kernel smoothing to estimate the conditional expectation of the portfolio loss given the risk factor. The focus of these studies is, however, to improve the metamodel accuracy for estimating the computer model's response surface in general. When the problem is to characterize extreme tail properties of Y, such approach can lose estimation accuracy, as discussed in Cannamela, Garnier, and Iooss (2008).

It is conceivable that a method for reliability will have to involve some type of variance reduction techniques that can guide the simulation process to generate outputs of interest (large Y values in our case). Among various variance reduction methods, importance sampling (IS) has been proven to be a powerful tool in many applications (Bulteau & El Khadiri, 2002; Cannamela et al., 2008; Chu & Nakayama, 2012; Hesterberg, 1995). Rather than sampling the input from the original density, $p(\mathbf{x})$, IS uses a biased density, $q(\mathbf{x})$, to sample $\mathbf{X}$, aiming to allocate greater sampling efforts over important input regions.

Most studies that develop the IS methods consider simulators that generate a deterministic output at the same input. The line of work on IS with deterministic computer models can be viewed as the reliability counterpart of emulator modeling (or metamodeling) for deterministic computer experiments (Table 1). Then, the line of work on IS with stochastic computer models is the reliability counterpart of metamodeling for stochastic computer experiments.

Recently, Choe et al. (2015) develop a new IS method, called stochastic importance sampling (SIS), for estimating reliability with stochastic black box computer models. The results in Choe et al. (2015) suggest that SIS is effective for estimating the failure probability of 1% or higher. In real life analyses, this probability will have to be smaller, for example, 10^{-4}. The approach in Choe et al. (2015) develops a nonadaptive (i.e., one-time) IS density. To estimate the extreme quantile associated with a very small probability, it is understandable that the SIS method could be reinforced with additional adaptive mechanisms.

This study develops a sequential method that informatively updates the IS density for efficiently estimating the extreme quantile with stochastic black box computer models. Specifically, as we iterate our quantile estimate, we use updated information to adjust the IS density parameter. To the best of our knowledge, this study is the first to develop an adaptive IS scheme for quantile estimation in the setting of

stochastic black box computer models. We study some convergence properties of our approach and demonstrate its benefits through numerical examples with a wide range of settings and a wind turbine case study. Implementation results suggest that our proposed method elicits substantial computational improvements over alternative approaches.

This paper is structured as follows. Section 2 reviews relevant studies and discusses challenging issues. Section 3 develops a new adaptive approach and provides its properties. Sections 4 and 5 present numerical examples and conduct a case study for the wind turbine extreme load estimation, respectively. Section 6 concludes the paper.

2 | BACKGROUND AND LITERATURE REVIEW

2.1 | Importance sampling with deterministic black box computer models

Crude Monte Carlo (CMC) sampling, which samples simulation inputs from $p(\mathbf{x})$, is the simplest way. However it is ineffective, because it generates samples most frequently in the main part of the density of Y. Unlike CMC, IS modifies its sampling focus on a different region of the density, for example, upper tail density.

Most studies that develop IS consider deterministic computer models that generate a fixed output given the input where the conditional failure probability, $\mathbb{P}(Y > y \mid \mathbf{X} = \mathbf{x})$ in (1), becomes an indicator function, that is, $\mathbb{I}(Y > y|\mathbf{X} = \mathbf{x})$ (Cannamela et al., 2008; Glynn, 1996; Hesterberg, 1995). When the target quantile is y_α, the optimal IS density that asymptotically minimizes the estimation variance is

$$q_{DIS}(\mathbf{x}) = \frac{1}{C_{DIS}} p(\mathbf{x}) \mathbb{I}(Y > y_\alpha|\mathbf{X} = \mathbf{x}), \tag{3}$$

where C_{DIS} is a normalizing constant (Morio, 2012).

Although $q_{DIS}(\mathbf{x})$ in (3) is theoretically optimal, it is not directly implementable in practice, because $\mathbb{I}(Y > y_\alpha|\mathbf{X} = \mathbf{x})$ and y_α are unknown. Therefore, estimating quantiles using IS requires approximating the unknown optimal IS density. In the literature with deterministic computer models, the metamodel approximation has been used in obtaining a good IS density. Using the Taylor expansion, Glasserman, Heidelberger, and Perwez (1999); Glasserman, Heidelberger, and Shahabuddin (2000) employ the delta and delta-gamma approximations to the financial loss in the portfolio value. Cannamela et al. (2008) state that a metamodel can be available from a previous study or from a physical model in industrial practice.

2.2 | Nested simulation and adaptive importance sampling

This section reviews two prominent research areas relevant to this study, namely, nested simulation and adaptive IS. First, the nested simulation schemes have been actively studied in the portfolio risk measurement literature. Glasserman et al. (1999, 2000) propose a quantile (value-at-risk) estimation method using the combination of IS and stratified sampling. They design the IS density with the exponential tilting by changing the density parameter in an exponential distribution family.

Gordy and Juneja (2010) dealt with the risk measurement problem that inevitably requires nested simulation due to the uncertainty between risk evaluation point and the horizon. They used two risk measures: value-at-risk and the probability of large loss. Having a limited budget of simulation, they provided a method to allocate the number of runs between outer and inner simulations minimizing the MSE. Similarly, Broadie, Du, and Moallemi (2011) consider the probability of large loss as a risk measure and propose a sequential approach for allocating more simulation budget to the inner simulation of the outer scenarios located close to the boundary of the tail probability, that is, close to y_α for the estimator of $P(Y > y_\alpha)$, using the optimization problem that maximizes the probability of a sign change. Gordy and Juneja (2010) and Broadie et al. (2011), however, do not consider the IS scheme. Recently Hong et al. (2017) use the kernel smoothing to estimate the conditional expectation of the portfolio loss given the risk factor, but they do not use the kernel estimator in Monte Carlo simulation.

Regarding the adaptive IS, Au and Beck (1999) propose a kernel-based sampling scheme for reliability estimation with a deterministic computer model. They devise a two-step algorithm where the first step uses Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) to generate input points lying in the failure region and the second step constructs kernel-based IS with the generated samples. They call their approach adaptive IS, because the next input sample is adaptively generated from the current sample in the Metropolis algorithm. Therefore, their adaptivity is different from the general notion of iterative updating of the importance density toward the unknown optimal density.

Recent studies provide more adaptive features that iteratively update the IS density using past samples, similar to the adaptivity implied in this study. Balesdent, Morio, and Marzat (2013) combine the Kriging metamodeling technique into the IS scheme. Specifically, they estimate the response surface with Kriging model and choose next sample points that can minimize the estimation uncertainty measured by the standard deviation in the Kriging response surface. Cornuet, Marin, Mirea, and Robert (2012), building upon the deterministic multiple mixture IS technique (Owen & Zhou, 2000), recompute importance weights of all simulated inputs generated from multiple densities. This approach is different from the standard approach that defines the importance weight as the likelihood ratio of the original input density to a single importance sampling density. Extensions on multiple IS have been made in Elvira, Martino, Luengo, and Bugallo (2017, 2019) where theoretical properties, including consistency and

variance reduction over standard weight scheme, are derived. These studies focus on estimating probability estimators.

For the quantile estimation, Morio (2012) uses the quantile estimate to update the IS density iteratively in a nonparametric framework, however, without any theoretical justification. Another adaptive approach is the stochastic approximation (SA) approach, which is a stochastic analog to the gradient descent method in deterministic nonlinear programming (Kushner & Yin, 2003). SA sequentially updates the quantile estimate, based on the difference between the failure probability estimate and the target probability. In the literature (Bardou, Frikha, & Pages, 2009; Egloff & Leippold, 2010; Kohler, KrzyZak, & Walk, 2014), SA is applied to find the root for a variance minimization problem to approximate the optimal IS density with deterministic computer models.

The adaptive IS scheme has been also studied in the Bayesian inference when a posterior density is known up to a normalizing constant. Comprehensive review of adaptive IS for the Bayesian inference as well as variance reduction is available in Bugallo et al. (2017).

2.3 | Importance sampling with stochastic black box computer models

With deterministic black box computer models, the original joint density of all random variables used in the simulation is known and takes a closed-form expression. This prerequisite is not satisfied for the stochastic black box computer model where the internal process is unknown and the input-output relationship is not deterministic. As discussed in Section 1, the stochastic computer model generates stochastic outputs even at the same input, because the random vector, $\boldsymbol{\xi}$, is hidden inside the model. In the nested simulation with stochastic black box computer models, the input $\mathbf{X}$ is first sampled and then the black box simulator, which embeds random vector ξ, generates the random output Y given $\mathbf{X}$ (Choe et al., 2018). The embeded ξ may, or may not, depend on $\mathbf{X}$. In either case, ξ is hidden inside the black box computer model and thus, one cannot sample ξ from its distribution. Consider the NREL simulator. It embeds over 8 million random variables and the joint density of $\mathbf{X}$ and $\boldsymbol{\xi}$ is not known to simulator users (instead, only the density of $\mathbf{X}$ is known).

Below we review the SIS method which minimizes the POE estimation variance using stochastic computer models (Choe et al., 2015). Let $\mathbf{X}_i$ ($i = 1, 2, \ldots, m$) denote the ith input sample drawn from the IS density, $q(\mathbf{x};\theta)$ for some parameter θ, and m be the input sample size. Due to the randomness in the output, SIS runs the simulator multiple times (say n_i times) at each $\mathbf{X}_i$ to obtain n_i outputs of Y_{ij} ($j = 1, 2, \ldots, n_i$). Then the POE estimator for the probability that Y exceeds the resistance level, y, becomes

$$\widehat{P}_{SIS}(y) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{I}(Y_{ij} > y)\right)\frac{p(\mathbf{X}_i)}{q(\mathbf{X}_i;\theta)}. \quad (4)$$

The estimator, $\widehat{P}_{SIS}(y)$, is unbiased when the support of $q(\mathbf{x};\theta)$, denoted as $supp\{q(\mathbf{x};\theta)\}$, includes $supp\{\mathbb{P}(Y > y \mid \mathbf{X} = \mathbf{x})p(\mathbf{x}))\}$. In other words, the following condition is required for $\widehat{P}_{SIS}(y)$ to be unbiased: *If* $q(\mathbf{x};\theta) = 0$, *then* $\mathbb{P}(Y > y \mid \mathbf{X} = \mathbf{x})p(\mathbf{x}) = 0$ *for any* $\mathbf{x}$. The unbiasedness condition can be also satisfied by the uniform continuity condition $q(\mathbf{x};\theta) = 0$ whenever $p(\mathbf{x}) = 0$.

Given the total number of simulation runs n_T, Choe et al. (2015) show that the optimal IS density that minimizes the variance of $\widehat{P}_{SIS}(y)$ is

$$q(\mathbf{x};\theta) = \frac{1}{C_q}p(\mathbf{x})\sqrt{s(\mathbf{x};\theta)}\sqrt{\frac{1}{n_T} + \left(1 - \frac{1}{n_T}\right)s(\mathbf{x};\theta)}, \quad (5)$$

where C_q is the normalizing constant. In $q(\mathbf{x};\theta)$, θ can be viewed as a density parameter where the optimal value for minimizing the variance of $\widehat{P}_{SIS}(y)$ is given by $\theta = y$, and $s(\mathbf{x};\theta)$ represents the conditional POE,

$$s(\mathbf{x};\theta) = \mathbb{P}(Y > \theta | \mathbf{X} = \mathbf{x}). \quad (6)$$

Suppose that m inputs, $\mathbf{x}_i$ ($i = 1, \ldots, m$), are sampled from $q(\mathbf{x};\theta)$. Choe et al. (2015) further show that the optimal run size at each $\mathbf{x}_i$ is

$$n_i = n_T\frac{\sqrt{\frac{n_T(1-s(\mathbf{x}_i;\theta))}{1+(n_T-1)s(\mathbf{x}_i;\theta)}}}{\sum_{i=1}^{m}\sqrt{\frac{n_T(1-s(\mathbf{x}_i;\theta))}{1+(n_T-1)s(\mathbf{x}_i;\theta)}}}. \quad (7)$$

When n_i is not an integer, it can be rounded to the nearest integer subject to $n_i \geq 1$. With rounding, we lose the theoretical optimality, but the loss would not be significant. Note that we use $\mathbf{x}_i$ to denote the realized value of the random variable $\mathbf{X}_i$.

In this approach the choice of θ is critical, because it affects the estimation efficiency. When we estimate $\mathbb{P}(Y > y)$ with a prespecified y, the optimal θ in $q(\mathbf{x};\theta)$ is y, because it provides the unbiased POE estimation and minimizes the estimation variance. This paper considers quantile estimation problem. Given a pre-specified failure probability, quantile is defined in (1). When the cumulative density of Y is continuous and strictly monotonic, the quantile can be rewritten as $y_\alpha = F^{-1}(1-\alpha)$, where F denotes a cdf of Y. Therefore, we can view the quantile estimation problem as the inverse of the POE estimation problem. However, y_α in our case is unknown a priori. In the next section we present an adaptive approach that steers the SIS density toward the optimal density, when quantiles are estimated via stochastic black box computer models.

3 | METHODOLOGY

3.1 | Adaptive importance sampling

The ideal IS density for quantile estimation is the one used to estimate the POE, $\mathbb{P}(Y > y_\alpha)$. Here, the "ideal" implies the optimality in terms of variance minimization. It has been

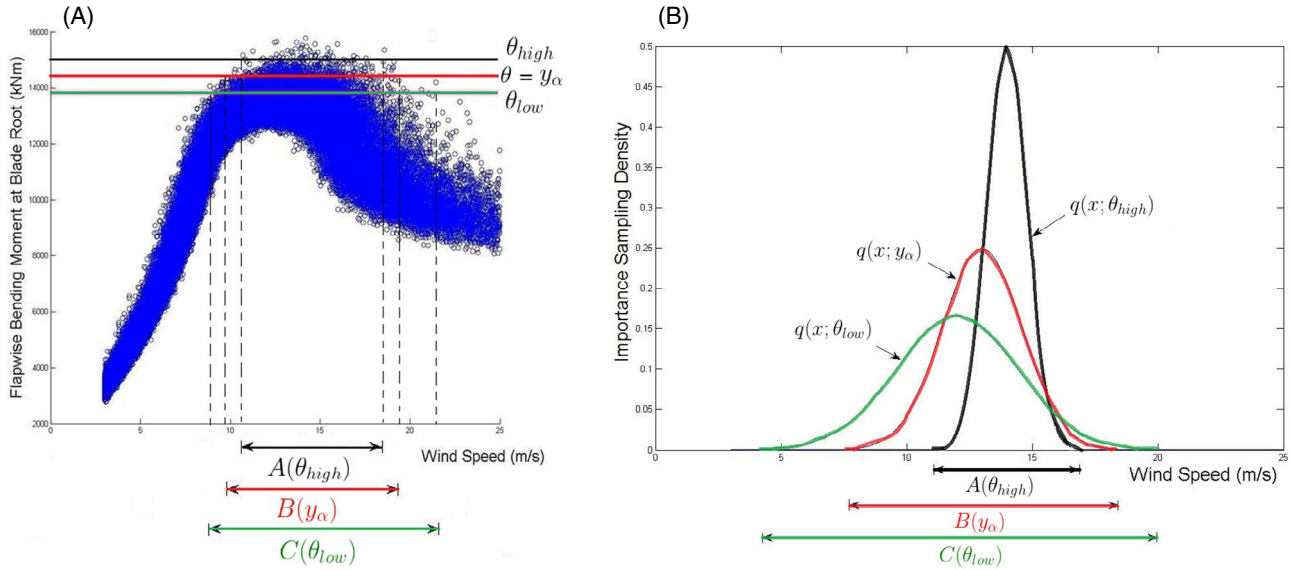

FIGURE 1 Example of wind turbine load response. (A) Flapwise bending moment. (B) SIS densities with different parameters [Colour figure can be viewed at wileyonlinelibrary.com]

shown that using Taylor expansion, we have

$$\hat{y}_\alpha = y_\alpha - \frac{\hat{P}_{SIS}(y_\alpha) - \alpha}{f_Y(y_\alpha)} + R_n, \tag{8}$$

where $\hat{y}_\alpha$ denotes the quantile estimate and R_n is a remainder which vanishes as the sample size grows under certain conditions (Chu & Nakayama, 2012; Pan, Byon, & Ko, 2020). Therefore, to minimize the variance of $\hat{y}_\alpha$, we need to minimize the variance of $\hat{P}_{SIS}(y_\alpha)$, and the density that minimizes the POE estimation variance also minimizes the quantile estimation variance. For the stochastic black box models, it is $q(\mathbf{x};\theta)$ in (5) and n_i in (7) with $\theta = y_\alpha$. Based on these key properties, our approach is to refine θ sequentially toward y_α throughout the iterative process.

We first examine the impact of θ on the estimation performance. Note that $q(\mathbf{x};\theta)$ in SIS allocates sampling efforts on the area.

$$supp\{q(\mathbf{x};\theta)\} = supp\{s(x;\theta)p(\mathbf{x})\} \tag{9}$$

$$= supp\{\mathbb{P}(Y > \theta|\mathbf{X} = \mathbf{x})p(\mathbf{x})\}, \tag{10}$$

where $supp\{\mathbb{P}(Y > \theta \mid \mathbf{X} = \mathbf{x})p(\mathbf{x})\}$ implies the input sampling area that the exceedance event, $\{Y > \theta\}$, can possibly happen. Therefore, the density parameter, θ, controls the input sampling area, which further affects the output samples that can be obtained from the simulator. When $q(\mathbf{x};\theta)$ uses a large θ (e.g., θ_{high} in Figure 1), the sampling efforts unduly focus on the narrow input region in practice, so the resulting quantile estimate can be substantially different from the true quantile. On the other hand, a too small θ (e.g., θ_{low} in Figure 1) distracts sampling efforts over unnecessarily large input areas (see $C(\theta_{low})$ in Figure 1b), losing simulation efficiency.

Consider an iterative simulation process. Let θ_k denote the IS density parameter used at the k^{th} iteration where K is the total number of iterations. During the simulation process, θ_k is determined based on the generated data, so it becomes random and even a carefully selected θ_k can possibly deviate from y_α. To handle the randomness of θ_k, we employ a new sampling density, $\widetilde{q}(x;\theta_k)$, that supports on the whole input space, $\Omega_{\mathbf{X}}$. Specifically, similar to the defensive sampling approach (Hesterberg, 1995), we modify $q_k(x;\theta_k)$ as.

$$\widetilde{q}(\mathbf{x};\theta_k) = \frac{1}{C_{\widetilde{q}}} p(\mathbf{x})\sqrt{\widetilde{s}(\mathbf{x};\theta_k)} \cdot \sqrt{\frac{1}{n_T} + \left(1 - \frac{1}{n_T}\right)\widetilde{s}(\mathbf{x};\theta_k)}, \tag{11}$$

where $C_{\widetilde{q}}$ is the normalizing constant and,

$$\widetilde{s}(\mathbf{x};\theta_k) = \left(1 - \frac{\delta}{k^\beta}\right) s(\mathbf{x};\theta_k) + \frac{\delta}{k^\beta}(1 - s(\mathbf{x};\theta_k)) \tag{12}$$

$$= \left(1 - \frac{2\delta}{k^\beta}\right) s(\mathbf{x};\theta_k) + \frac{\delta}{k^\beta}, \tag{13}$$

with some positive constants $\delta(<0.5)$ and β. Here, $\widetilde{s}(\mathbf{x};\theta_k)$ ranges between 0 and 1 (i.e., $0 < \widetilde{s}(\mathbf{x};\theta_k) < 1$). For $\delta < 0.5$, $\widetilde{s}(\mathbf{x};\theta_k)$ increases as $s(\mathbf{x};\theta_k)$ increases. With small δ, the first term in (13) enables the sampling efforts to be focused on the important input area with high failure probability, whereas the second term allows some portion of sampling efforts to be allocated over the entire input domain. The construction of $\widetilde{s}(\mathbf{x};\theta_k)$ in (13) guarantees that the variance of the POE estimator is bounded, which is proved in Lemma 1 and used in showing the consistency properties later.

At each iteration, we sample m inputs, $\mathbf{x}_{i,k}$ $(i = 1, 2, \ldots, m)$, from $\widetilde{q}(\mathbf{x};\theta_k)$. At each $\mathbf{x}_{i,k}$, we also modify the allocation size, n_i in (7), to

$$\widetilde{n}_{i,k} = n_T \frac{\sqrt{\frac{1-\widetilde{s}(\mathbf{x}_{i,k};\theta_k)}{1+(n_T-1)\widetilde{s}(\mathbf{x}_{i,k};\theta_k)}}}{\sum_{i=1}^{m}\sqrt{\frac{1-\widetilde{s}(\mathbf{x}_{i,k};\theta_k)}{1+(n_T-1)\widetilde{s}(\mathbf{x}_{i,k};\theta_k)}}}. \tag{14}$$

When $\widetilde{n}_{i,k}$ is not an integer, we round it to the nearest integer. If the nearest integer is zero, we set $\widetilde{n}_{i,k} = 1$.

It should be noted that the variance minimizing properties in (5)–(7) are not completely carried to Equations (11)–(14) due to $\delta \neq 0$. However, as iterations proceed, the second term in (13) diminishes. Thus, if θ_k converges to the target quantile y_α, the variance minimizing properties become more clear at later iterations. On the other hand, earlier iterations explore wider input areas at the cost of increased variance, but such wider exploration is needed for accommodating insufficient information in choosing right θ_k.

In practice, the conditional failure probability, $s(\mathbf{x};\theta_k)$, is not available. A reasonable approximation is to use its metamodel as a substitute for $s(\mathbf{x};\theta_k)$. We present our method and its properties with the exact $s(\mathbf{x};\theta_k)$ and then extend the analysis when $s(\mathbf{x};\theta_k)$ is approximated by its metamodel.

Now we discuss how to choose θ_k at each iteration. Considering that the most desirable density parameter is y_α, we propose to use the quantile estimate to guide the simulation process (Morio, 2012). Specifically, to get the quantile estimate, we use the following combined POE estimator,

$$\widehat{P}_{1:K}(y) = \frac{1}{K}\sum_{k=1}^{K}\widehat{P}_k(y) \tag{15}$$

with

$$\widehat{P}_k(y) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{\widetilde{n}_{i,k}}\sum_{j=1}^{\widetilde{n}_{i,k}}\mathbb{I}(Y_{ij,k} > y)\right)\frac{p(\mathbf{X}_{i,k})}{\widetilde{q}(\mathbf{X}_{i,k};\theta_k)}, \tag{16}$$

where $Y_{ij,k}$ is the jth ($j = 1, 2, \ldots, \widetilde{n}_{i,k}$) output at each $\mathbf{x}_{i,k}$, $\widehat{P}_k(y)$ is the individual POE estimator at the k^{th} iteration, and $\widehat{P}_{1:K}(y)$ aggregates the K POE estimators to fully utilize the information obtained from all iterations.

Because $\widetilde{s}(\mathbf{x};\theta_k)$ is strictly positive over $\Omega_\mathbf{x}$, $\widetilde{q}(\mathbf{x};\theta_k) = 0$ implies $\mathbb{P}(Y > y \mid \mathbf{X} = \mathbf{x})p(\mathbf{x}) = 0$ for any $\mathbf{x} \in \Omega_\mathbf{X}$. Therefore, the POE estimator, $\widehat{P}_{1:K}(y)$, is unbiased, $\forall y \in \Omega_Y$, where Ω_Y denotes the output space. Moreover, the variance of the POE estimator is bounded, thanks to the construction of $\widetilde{s}(\mathbf{x};\theta_k)$ in (13), as shown in Lemma 1.

Lemma 1 *(a) Variance of $\widehat{P}_k(y)$ in (16) is bounded, $\forall y \in \Omega_Y$. (b) Variance of $\widehat{P}_{1:K}(y)$ in (15) is also bounded, $\forall y \in \Omega_Y$.*

Using the combined POE estimator, the intermediate quantile estimate after the k^{th} iteration is defined as

$$\widehat{y}_k^\alpha = min\{y : 0 < \widehat{P}_{1:k}(y) \leq \alpha\}. \tag{17}$$

or

$$\widehat{y}_{k,\alpha} = max\{y : \widehat{P}_{1:k}(y) \geq \alpha\}, \tag{18}$$

where $\widehat{y}_k^\alpha$ and $\widehat{y}_{k,\alpha}$ can be obtained using order statistics among the outputs obtained up to the current iteration (Choe, Pan, & Byon, 2016). Any of these two estimates can be used as the next density parameter, θ_{k+1}. In our implementation, we use $\widehat{y}_{k,\alpha}$, that is, $\theta_{k+1} = \widehat{y}_{k,\alpha}$. Specifically we sort the outputs ($Y_{ij,h}$, $i = 1, \ldots, m$, $j = 1, \ldots, \widetilde{n}_{i,k}$, $h = 1, \ldots, k$) obtained up to the kth iteration. Let $Y_{(s)}$ denote the sth smallest values among all $Y_{ij,h}$'s. We sequentially compute $\widehat{P}_{1:k}(Y_{(s)})$ from the largest value. Then the order statistic $Y_{(s)}$ that satisfies $\widehat{P}_{1:k}(Y_{(s)}) \geq \alpha$ and $\widehat{P}_{1:k}(Y_{(s+1)}) \leq \alpha$ is identified as θ_{k+1}. In our implementation we use the "sort" function in Matlab to obtain order statistics. With the kn_T samples obtained up to the kth iteration, the complexity is $O(kn_T \cdot log(kn_T))$ on average (Mathworks, 2004).

As a remark, instead of the POE estimator in (15) and (16), we can also use the self-normalized estimator (Owen, 2013). Both estimators are consistent estimators (Owen, 2013), so they can be used with the proposed scheme. We brifely compare the two estimators. First, the estimator in (15) and (16) provides the unbiased probability estimation with any sample size, and this form of the estimator has been widely used in the IS literature (Bucklew, 2004). The self-normalized estimator is *asymptotically* unbiased, that is, it converges to the true probability when the sample size gets large (Owen, 2013). Second, the self-normalized estimator is beneficial when an unnormalized version of p or $\widetilde{q}$ is only available. Lastly, it is more complicated to obtain the self-normalized POE variance estimate and its bound, in particular, in an adaptive setting. Therefore, we employ the original estimator in (15) and (16) and present asymptotic properties of the proposed approach in the next section. In our future study, we plan to compare the theoretical properties and estimation performance between the two estimators.

3.2 | Asymptotic properties

This section establishes some asymptotic properties of the proposed adaptive approach. In particular, we prove consistency and variance reduction properties of our approach. The relevant proofs and derivations are available in the Appendix. A key issue in showing the consistency properties is that θ_k is random. Suppose that the importance sampler in (11), $\widetilde{q}$, is employed with $\widetilde{n}_i$ ($i = 1, \ldots, m$) in (14) and that θ_k is refined with the quantile estimate.

Assumption 1 *The cdf of Y is continuous and strictly increasing.*

First, based on the results in Lemma 1, Theorem 1 specifies two conditions on β to make the POE estimator converge to the true POE, $\mathbb{P}(Y > y_\alpha)$, in probability and almost surely. The results suggest that a too large β may make the POE estimator fail to be consistent. This is because a large β shrinks the support of IS density rapidly. So β should be chosen with care.

Theorem 1 *Suppose that Assumption 1 holds. Then $\widehat{P}_{1:K}(y) \xrightarrow{P} \mathbb{P}(Y > y)$, $\forall y \in \Omega_Y$, as $K \to \infty$, for $0 < \beta < 1$. Moreover, $\widehat{P}_{1:K}(y) \xrightarrow{a.s.} \mathbb{P}(Y > y)$, $\forall y \in \Omega_Y$, as $K \to \infty$, for $0 < \beta < 0.5$.*

Next we show the consistency properties of the quantile estimators. First, Corollary 1 shows that $\widehat{y}_K^\alpha$ in (17) is a consistent estimator of the target quantile, y_α.

Corollary 1 *Suppose that Assumption 1 holds. Then* $\widehat{y}_K^{\alpha} \xrightarrow{P} y_\alpha$, *as* $K \to \infty$, *for* $0 < \beta < 1$.

Recall that we use $\widehat{y}_{k,\alpha}$ as the next density parameter value. Corollary 2 also shows the convergence of θ_K to y_α as K becomes large.

Corollary 2 *Suppose that Assumption 1 holds. Then* $\theta_K \xrightarrow{P} y_\alpha$, *as* $K \to \infty$, *for* $0 < \beta < 1$.

These consistency properties are important, because they indicate that $\widetilde{q}(\mathbf{x}; \theta_k)$ in (11) approaches the ideal density, $q(\mathbf{x}; y_\alpha)$ in (5), as K gets larger. This result can be translated into variance reduction of our approach over CMC. Consider the following POE estimator of CMC with the same computational budget, Kn_T.

$$\widehat{P}_{CMC}(y_\alpha) = \frac{1}{Kn_T} \sum_{i=1}^{Kn_T} \mathbb{I}(Y_i > y_\alpha), \tag{19}$$

where each input $\mathbf{x}_i$ is sampled from $p(\mathbf{x})$ and Y_i is simulated at $\mathbf{x}_i$. The variance of $\widehat{P}_{CMC}(y_\alpha)$ is given by

$$Var[\widehat{P}_{CMC}(y_\alpha)] = \frac{\mathbb{P}(Y > y_\alpha)[1 - \mathbb{P}(Y > y_\alpha)]}{Kn_T} = \frac{\alpha(1-\alpha)}{Kn_T}. \tag{20}$$

or equivalently,

$$Var[\sqrt{Kn_T}\widehat{P}_{CMC}(y_\alpha)] = \alpha(1-\alpha). \tag{21}$$

We can also consider another CMC sampling scheme that allows multiple runs at each sampled input, referred to as CMC2. Given the total computational budget, Kn_T, we generate m inputs, $\mathbf{X}_i$, $i = 1, \ldots, m$, from $p(\mathbf{x})$. At each $\mathbf{X}_i$, CMC2 obtains n_i outputs of Y_{ij}'s, such that $\sum_{i=i}^{m} n_i = Kn_T$. Therefore, the CMC2's POE estimator becomes.

$$\widehat{P}_{CMC2}(y_\alpha) = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} \mathbb{I}(Y_{ij} > y_\alpha) \right). \tag{22}$$

While this CMC2 scheme shares some similarity with SIS in that multiple replications are allowed at each sampled input, it uses the input density $p(\mathbf{x})$ only, ignoring the geometric structure of response surface. When we use the equal sample sizes, that is, $n_i = (Kn_T)/m$, it turns out that the variance of $\widehat{P}_{CMC2}(y_\alpha)$ is larger than that of $\widehat{P}_{CMC}(y_\alpha)$. Detailed discussion and derivation are available in Appendix A7. As such, allowing multiple replicates is not beneficial in the CMC procedure and thus, we do not consider the CMC2 scheme in the subsequent discussion.

Theorem 2 states that our approach can achieve variance reduction over CMC. It indicates that our method is always beneficial over CMC, unless the conditional POE at y_α, $s(\mathbf{x}; y_\alpha)$, is constant with respect to $\mathbf{x}$. When $s(\mathbf{x}; y_\alpha)$ is constant, $\widetilde{q}(\mathbf{x}; \theta_K)$ converges to $p(\mathbf{x})$. In this special case there is no need to bias the input density, so equality holds in (23).

Theorem 2 *Suppose that Assumption 1 holds. Then*

$$\lim_{K\to\infty} Var[\sqrt{Kn_T}\widehat{P}_{1:K}(y_\alpha)] \le \alpha(1-\alpha), \tag{23}$$

for $0 < \beta < 1$, *where the equality in (23) holds if and only if* $s(\mathbf{x}; y_\alpha)$ *is constant over the entire input domain,* $\Omega_{\mathbf{x}}$. *In other words, the asymptotic variance of the POE estimator in the proposed approach is always strictly smaller than CMC's except the special case where* $s(\mathbf{x}; y_\alpha)$ *is constant over* $\Omega_{\mathbf{x}}$.

The aforementioned convergence properties are established for K tending to infinity. In practice, it could be impractical to have a large K, when simulation is computationally expensive. However, the asymptotic results developed in this study highlights the benefit of using the adaptive procedure we propose. Numerical studies in Sections 4–6 show that the quantile estimates from our approach become close to the target quantile within a relatively small number of iterations, for example, 25 iterations, in many cases.

3.3 | Approximation of $s(\mathbf{x}; \theta_k)$ and implementation summary

The proposed approach requires information on $s(\mathbf{x}; \theta_k)$ in (6) in order to define $\widetilde{q}(\mathbf{x}; \theta_k)$ and $\widetilde{n}_{i,k}$ in (11) and (14), respectively. In practice, $s(\mathbf{x}; \theta_k)$ is unknown for stochastic black box computer models, so it needs to be approximated. Depending on applications, different statistical models, for example, GP, can be employed. For the wind turbine simulation, Choe et al. (2015) suggest using the generalized additive model for location, scale and shape (GAMLSS) (Rigby & Stasinopoulos, 2005) (more details will be discussed in Section 6).

Let $s^a(\mathbf{x}; \theta_k)$ denote a metamodel that approximates $s(\mathbf{x}; \theta_k)$ satisfying $0 \le s^a(\mathbf{x}; \theta_k) \le 1, \forall \mathbf{x} \in \Omega_{\mathbf{X}}$. Suppose that we replace $s(\mathbf{x}; \theta_k)$ with $s^a(\mathbf{x}; \theta_k)$ in the importance sampler defined in (11)–(14) and the POE estimator in (15). With $s^a(\mathbf{x}; \theta_k)$, the results in Theorem 1, Corollaries 1 and 2, and Theorem 2 still hold. To prove this, we just need to replace $s(\mathbf{x}; \theta_k)$ with $s^a(\mathbf{x}; \theta_k)$ in our derivations provided in the Appendix.

However, achieving the variance reduction over CMC with $s^a(\mathbf{x}; \theta_k)$, similar to the result in Theorem 2, requires accurate approximation of $s(\mathbf{x}; \theta_k)$. Below we show that variance reduction can hold under certain conditions. Let $\|\cdot\|$ denote the norm on the continuous function space w.r.t. the input vector, that is, $\|s(\mathbf{x}; y)\| := \max_{\mathbf{x}\in\Omega_{\mathbf{X}}} |s(\mathbf{x}; y)|$.

Theorem 3 *Let* $\overline{F}(y) = \mathbb{P}(Y > y)$. *With Assumption 1, suppose that* $p_{max} := \max_{\mathbf{x}\in\Omega_{\mathbf{X}}} p(\mathbf{x}) < \infty$ *and* $\|s^a(\mathbf{x}; \theta_k) - s(\mathbf{x}; \theta_k)\| = o(k^{-\beta})$. *We further assume that* $s(\mathbf{x}; y)$ *and* $\overline{F}^{-1}(p)$ *are locally Lipschitz continuous at* $y = y_\alpha$ *and* $p = \alpha$, *respectively. Then, after* $s(\mathbf{x}; \theta_k)$ *is replaced with* $s^a(\mathbf{x}; \theta_k)$ *in (11)–(14), it holds*

$$\lim_{K\to\infty} Var[\sqrt{Kn_T}\widehat{P}_{1:K}(y_\alpha)] \le \alpha(1-\alpha), \quad (24)$$

for $0 < \beta < 0.5$.

In Theorem 3, $\|s^a(\mathbf{x};\theta_k) - s(\mathbf{x};\theta_k)\| = o(k^{-\beta})$ implies that the maximum difference between the estimated and true conditional failure probability decreases at a rate faster than $k^{-\beta}$ as iterations proceed. In other words, this condition requires a high-quality metamodel in the tail portion of the conditional output density. Admittedly, this condition is strong and it is difficult to show whether this condition is satisfied for stochastic black box computer models. The simulation process can be possibly steered in a wrong direction with poor approximation of $s(\mathbf{x};\theta_k)$.

In the literature the metamodel approximation has been used in obtaining a good IS density (Balesdent et al., 2013; Cannamela et al., 2008; Glasserman et al., 1999, 2000). The focus of this study is to develop a procedure for estimating extreme quantiles, assuming a good metamodel is available. The proposed approach, regardless of the metamodel quality, provides a unbiased POE estimation, which leads to an unbiased quantile estimation with the sample size sufficiently large. Our numerical results with different metamodel qualities in Section 4 suggest that the proposed adaptive approach is robust to the approximation quality. However, admittedly the metamodel quality affects the efficiency of the procedure. To the best of our knowledge, how the metamodel approximation error affects the efficiency in Monte Carlo simulation has not been studied yet in the literature. Understanding how the approximation error is transferred to the SIS density is a subject of our future research.

We call the proposed approach adaptive SIS (shortly, A-SIS). In particular, when $\widehat{y}_K^\alpha$ is used for estimating y_α, we refer the method to as A-SIS$_1$, while estimating y_α with $\widehat{y}_{K,\alpha}$ is referred to as A-SIS$_2$. Both A-SIS$_1$ and A-SIS$_2$ are collectively called A-SIS in the subsequent discussion. We assume the metamodel, $s^a(\mathbf{x};y)$, for approximating $s(\mathbf{x};y)$, is available. When it is not available, we can build it using pilot samples. Below we summarize the implementation procedure of A-SIS.

Algorithm 1. ASIS quantile estimation procedure

Initialization: Set parameters β, δ, m, n_T, K and the initial parameter θ_1. Set $k = 1$.

1: Sample $\mathbf{x}_{i,k}$ from $\widetilde{q}(\mathbf{x};\theta_k)$ in (11) and determine the allocation size $\widetilde{n}_{i,k}$ in (14) for each $\mathbf{x}_{i,k}$ $(i = 1, \ldots, m)$.

2: Run simulation $n_{i,k}$ times at each $\mathbf{x}_{i,k}$ to generate $Y_{ij,k}$ $(i = 1, \cdots, m, j = 1, \ldots, n_{i,k})$.

3: Obtain θ_{k+1} in (18). If $k < K$, set $k = k+1$ and go to Step 1. Otherwise, go to Step 4.

4: Obtain the (1-α)-quantile estimate using $\widehat{y}_K^\alpha$ in A-SIS$_1$, or θ_{K+1} in A-SIS$_2$.

Remark 1 In Step 1 of Algorithm 1, we can use the acceptance-rejection algorithm for drawing samples from $\widetilde{q}(\mathbf{x};\theta_k)$ (Asmussen & Glynn, 2007). We note that acceptance-rejection may have a low acceptance rate, so it may not lead to overall computation efficiency improvement in situations where the efficiency is based on the number of input generation to draw samples from $\widetilde{q}(\mathbf{x};\theta_k)$. In our case, however, the computational bottleneck is the evaluation of the computer model given the input, not the generation of the inputs. Thus, we can afford to sample a large number of inputs to generate the IS density. For example, consider an experiment with $m = 30$ and $n_T = 100$ at each iteration. In our wind turbine case study, it takes about 0.01 seconds to draw inputs from the IS density, whereas running the simulator takes about 100 minutes at each iteration. Therefore, the computational overhead to draw samples from the proposed IS density can be considered as negligible. Other sampling methods, for example, Markov chain Monte Carlo (MCMC), can also be used for sampling the inputs.

Remark 2 Although our approach requires approximating $s(\mathbf{x};\theta_k)$, it is different from the emulator-based approach that replaces the computer model with a metamodel (or surrogate model). In our approach, the metamodel is used to approximate the true conditional failure probability, thus to guide the adaptive IS procedure.

4 | EXAMPLE 1

To investigate the performance of the proposed method, we employ the numerical example with the following data generating structure.

$$\mathbf{X} \sim N(0, \sigma_{\mathbf{X}}^2) \quad (25)$$

$$Y|\mathbf{X} \sim N(\mathbf{X}, \sigma_{Y|\mathbf{X}}^2) \quad (26)$$

with $\sigma_{\mathbf{X}} = 5$ and $\sigma_{Y|\mathbf{X}} = 1$. Therefore, the conditional POE in this example becomes.

$$s(\mathbf{x};\theta_k) = \mathbb{P}(Y > \theta_k|\mathbf{X} = \mathbf{x}) = 1 - \Phi\left(\frac{\theta_k - \mathbf{X}}{\sigma_{Y|\mathbf{X}}}\right), \quad (27)$$

where Φ denotes the standard normal cdf. Plugging (27) into Equation (13), we can get $\widetilde{s}(\mathbf{x};\theta_k)$, which in turn provides $\widetilde{q}_k(\mathbf{x};\theta_k)$ in (11) and $\widetilde{n}_{i,k}$ in (14). We first consider the perfect metamodel and use $\beta = 0.1$, $\delta = 0.1$, and $\theta_1 = 1$ as a baseline setting. We also set $m = 30$, $n_T = 100$ and $K = 25$. Then we conduct sensitivity analysis with other settings, including imperfect metamodels. In all cases, we focus on estimating the extreme quantile for 10^{-4}. In this data generating structure,

$Y \sim N(0, \sigma_{\mathbf{X}}^2 + \sigma_{Y|\mathbf{X}}^2)$, so the true quantile can be calculated explicitly. With $\alpha = 10^{-4}$, the true quantile is $y_\alpha = 19.0$.

As a remark, when the density of Y does not take a closed-form, we obtain the true quantile estimate using CMC in evaluating the estimation performance. For example, in the wind turbine case study in Section 6, we use the CMC estimate with 10^6 replications. To check if 10^6 replications are sufficient, we conduct 25 CMC experiments (each with 10^6 replications) with the above example. The standard deviation and MSE of the CMC estimates obtained from 25 experiments are 0.013341 and 0.000179, respectively (note that we use the true quantile, $y_\alpha = 19.0$, when we compute MSE). The average difference between individual CMC estimates and true value is 0.002370. These results justify the use of CMC quantile estimate with 10^6 replications in 1-dimensional case study in Section 6.

4.1 | Alternative methods

We compare the estimation performance of A-SIS with alternative approaches. We first consider the nonadaptive SIS (NA-SIS) method where we use $\widetilde{q}_k(\mathbf{x}; \theta_k)$ in (11) with $\theta_1 = 1$ as an IS density and do not update the IS density. By comparing A-SIS with NA-SIS, we can evaluate the advantage of parameter updating.

Also, considering SA has been used as an adaptive IS approach for deterministic computer models (Bardou et al., 2009; Kohler et al., 2014), we implement SA in the stochastic setting. The Robbins-Monro algorithm (Robbins & Monro, 1951) provides a prototypical SA method, and Polyak (1990) and Ruppert (1988) further improve the Robbins-Monro algorithm by introducing an averaging idea. Specifically, we use the same importance sampler, $\widetilde{q}_k(\mathbf{x}; \theta_k)$, in (11) and update θ_k using the averaging idea (Polyak, 1990; Ruppert, 1988). That is, the IS density parameter is updated as follows.

$$\theta_{k+1}^{SA} = \frac{1}{k+1} \sum_{s=1}^{k+1} V_s, \tag{28}$$

with.

$$V_{k+1} = \theta_k^{SA} + \frac{a}{k^\gamma} \cdot (\widehat{P}_k(\theta_k) - \alpha), \tag{29}$$

where $\widehat{P}_k(\theta_k)$ is the POE estimator defined in (16). After the last iteration, θ_K^{SA} becomes the SA's quantile estimator. In implementing SA, we use $a = 100$ and $\gamma = 0.5$.

Note that the implemented SA with (28) and (29) is similar to A-SIS, in the sense that they use the same importance sampler, $\widetilde{q}_k(\mathbf{x}; \theta_k)$, and update the density parameter throughout iterations. The main difference is the updating rule: A-SIS updates the IS density parameter based on the quantile estimate using all the past samples, whereas SA updates it based on the difference between the target and estimated POEs.

4.2 | Implementation results

Table 2 summarizes the implementation results from 100 experiments under the baseline setting. The average difference (Avg. diff.) in the third column denotes the averaged difference between the true quantile and quantile estimates from 100 experiments. The results indicate that the estimated quantiles from A-SIS$_1$ and A-SIS$_2$ are close to y_α with small difference. The NA-SIS's average difference is more than three times larger than A-SIS, mainly because it does not update the IS density.

TABLE 2 Quantile estimation results from 100 experiments under the baseline setting

Methods	Sample std.	Avg. diff	MSE
A-SIS$_1$	1.9	0.5	3.8
A-SIS$_2$	1.2	−0.9	2.3
NA-SIS	1.2	−3.3	12.5
SA	1.6	3.1	12.0

TABLE 3 Quantile estimation results with different θ_1 (In SA, $a = 50$, 200, and 1000 are used in SA for $\theta_1 = 1$, 8, and 15, respectively)

θ_1	Methods	Sample std.	Avg. diff	MSE
1	A-SIS$_1$	1.9	0.5	3.8
	A-SIS$_2$	1.2	−0.9	2.3
	NA-SIS	1.2	−3.3	12.5
	SA	1.6	3.1	12.0
8	A-SIS$_1$	1.8	0.5	3.5
	A-SIS$_2$	1.4	−0.8	2.5
	NA-SIS	1.2	−1.9	5.1
	SA	3.6	1.0	13.9
15	A-SIS$_1$	1.5	0.4	2.5
	A-SIS$_2$	1.0	−0.9	1.1
	NA-SIS	1.1	−1.3	3.0
	SA	2.2	−1.5	7.2

It should be noted that the result of SA is highly sensitive to the choice of a. In Section 4.3, detailed sensitivity analysis results are discussed. In this example we explore a wide range of a and choose an appropriate value that generates small estimation errors, which is $a = 50$. Even after carefully tuning a, SA yields a large difference, because its sequence does not converge within 25 iterations.

4.3 | Sensitivity analysis

We conduct sensitivity analysis under widely different parameter settings. First we compare our approach with NA-SIS and SA with different initial parameters. Table 3 summarizes results with three different θ_1 values. The estimation performance of NA-SIS differs, depending on θ_1. When θ_1 is closer to the target quantile $y_\alpha = 19.0$, its estimation results generally become better. With $\theta_1 = 15$, the initial IS density is already close to the optimal one, so NA-SIS produces small estimation errors. These results indicate that the NA-SIS's estimation capability highly depend on the initial parameter choice. In particular, the average difference from NA-SIS gets larger, as the initial parameter is more different from the target

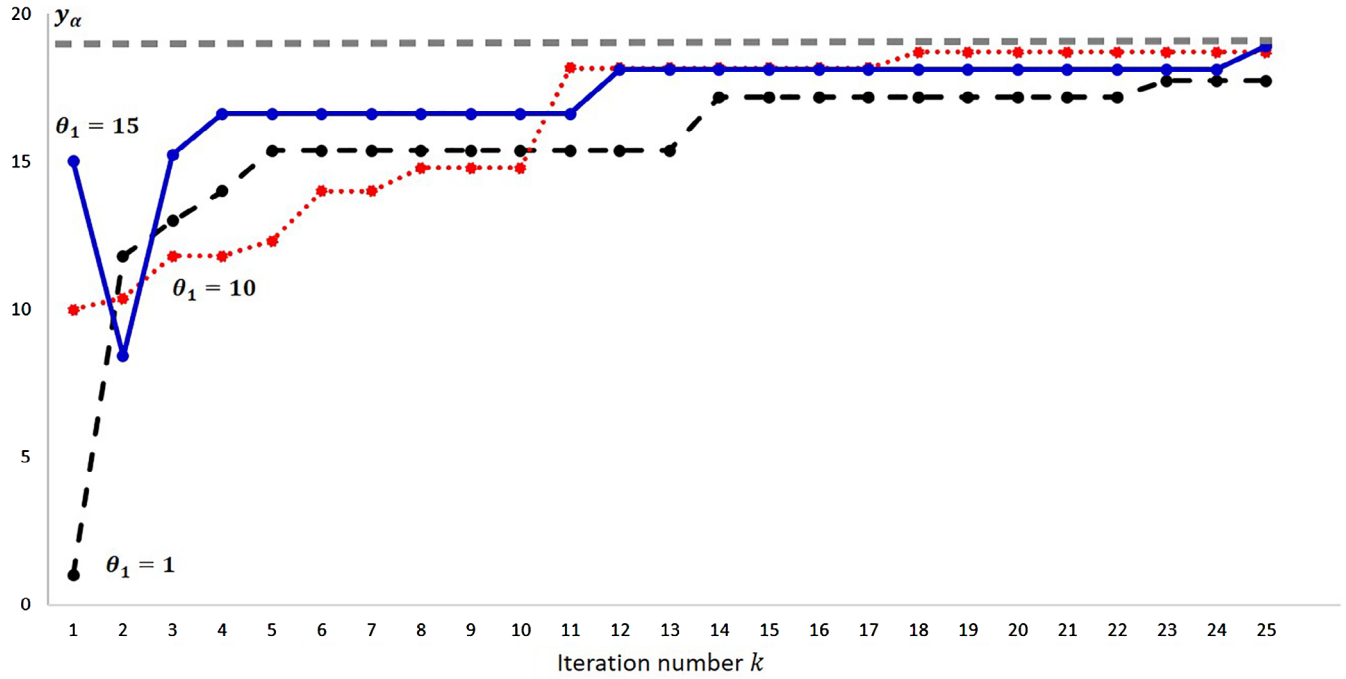

FIGURE 2 Parameter sequence in A-SIS with different θ_1 values [Colour figure can be viewed at wileyonlinelibrary.com]

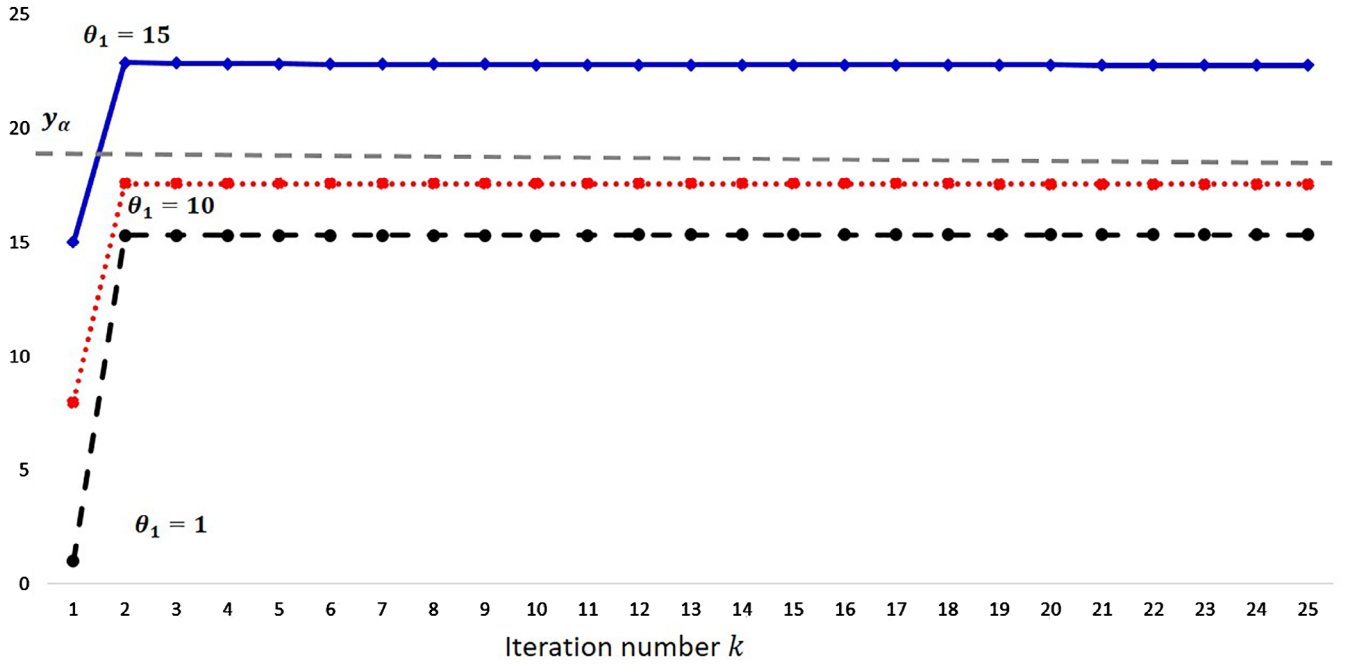

FIGURE 3 Parameter sequence in SA with different θ_1 values [Colour figure can be viewed at wileyonlinelibrary.com]

value. While we observe a similar pattern in SA, the results also depend on the step size, a. Table 3 reports the SA results with carefully tuned step sizes.

Unlike SA and NA-SIS, our approach is robust to the choice of initial parameter and consistently generates lower errors with all three different θ_1's. Figure 2 further depicts the trajectories of θ_k along the iterations in A-SIS. Even with very small θ_1 (e.g., $\theta_1 = 1$), θ_k increases reasonably fast after a small number of iterations and become quite close to y_α within 25 iterations. On the contrary, Figure 3 shows that SA does not appropriately update the parameter after the first few iterations and cannot reach the target value within 25 iterations.

We further investigate the impacts of metamodel quality. For reflecting a metamodel approximation error, we use a metamodel that incorrectly specifies the conditional distribution. Specifically, the standardized conditional density of Y given X is assumed to follow the t-distribution in the metamodel. Table 4, which summarizes the result with different degrees of freedom in the studentized t-distribution, demonstrates that the proposed approach generates robust performance. The performance of the proposed approach is comparable to that with the perfect metamodel in Table 2.

Next, we study how the values of β and δ affect the estimation capability. Tables 5 and 6 summarize the results of

TABLE 4 Quantile estimation results with different degrees of freedom in the metamodel with the studentized t-distribution

	A-SIS$_1$			A-SIS$_2$			SA		
	Sample	Avg.		Sample	Avg.		Sample	Avg.	
df	Std.	Diff.	MSE	Std.	Diff.	MSE	Std.	Diff.	MSE
5	1.4	0.2	1.8	1.1	−1.2	2.7	1.8	3.2	13.1
15	1.9	0.3	3.5	1.2	−1.2	2.8	1.5	3.0	11.3
25	1.7	0.3	2.9	1.0	−1.0	2.1	1.9	3.3	14.0

TABLE 5 Quantile estimation results with different β values

	A-SIS$_1$			A-SIS$_2$			SA		
	Sample	Avg.		Sample	Avg.		Sample	Avg.	
β	Std.	Diff.	MSE	Std.	Diff.	MSE	Std.	Diff.	MSE
0.01	1.5	0.0	2.1	1.1	−1.4	3.2	1.6	3.2	12.6
0.1	0.9	0.5	3.8	1.2	−0.9	2.3	1.6	3.1	12.0
0.2	1.4	−0.1	1.9	1.1	−1.2	2.5	1.6	3.3	13.1

TABLE 6 Quantile estimation results with different δ values

	A-SIS			A-SIS			SA		
	Sample	Avg.		Sample	Avg.		Sample	Avg.	
δ	Std.	Diff.	MSE	Std.	Diff.	MSE	Std.	Diff.	MSE
0.01	0.6	−0.2	0.5	0.6	−0.5	0.7	0.9	3.3	11.9
0.1	0.9	0.5	3.8	1.2	−0.9	2.3	1.6	3.1	12.0
0.2	1.9	−0.4	3.6	1.1	−1.9	4.9	2.3	3.2	15.6

TABLE 7 Quantile estimation results with different a and γ in SA

		SA		
a	γ	Sample std.	Avg. diff	MSE
25	0.1	0.7	−6.7	44.8
	0.5	0.7	−7.0	49.3
	0.9	0.8	−7.2	51.8
50	0.1	1.6	3.1	12.1
	0.5	1.6	3.1	12.0
	0.9	1.7	3.3	13.7
75	0.1	2.6	14.0	202.4
	0.5	2.6	13.6	192.5
	0.9	2.4	13.6	190.0

our approach with different β and δ values, respectively. The implementation results with a wide range of settings suggest that our procedure generates stable estimations, demonstrating its robust performance. In all cases, A-SIS provides better estimation results, compared with SA.

It is worthwhile to mention that one critical disadvantage of SA is that its performance is sensitive to the choice of step parameters, a and γ. Table 7 demonstrates that SA's estimation performance varies substantially, depending on the step parameters, in particular, the value of a.

In summary, the implementation results with a wide range of settings suggest that the proposed method is robust to the parameter setting. It also consistently provides better results, compared to alternative approaches. Between A-SIS$_1$ and A-SIS$_2$, A-SIS$_1$ generates quantile estimates closer to the target quantile in general. It is mainly because A-SIS$_1$ uses the higher order statistics, $\widehat{y}_k^{\alpha}$, than A-SIS$_2$ with $\widehat{y}_{k,\alpha}$. While A-SIS$_1$ appears to perform slightly better when K is small, the estimates from A-SIS$_1$ and A-SIS$_2$ would become closer to each other with larger K.

4.4 | Computational budget allocation

This section examines the impact of computational budget allocation on the estimation performance. In our study, given the total computational resource of $K \cdot n_T$, the budget allocation rules involve the number of sample points (m), the number of replications for each sampled point (n_i), the computational budget at each iteration (n_T) and the number of iterations (K).

First, in the original SIS method presented in Choe et al. (2015), given the number of sample points (m) and the computational budget (n_T), variance-minimizing n_i at each sampled x_i is decided with Equation (7). Moreover, Choe et al. (2015) empirically demonstrate that the estimation performance, in terms of the variance, is not sensitive to the choice of m, given n_T in their experiments in a wide range of setting.

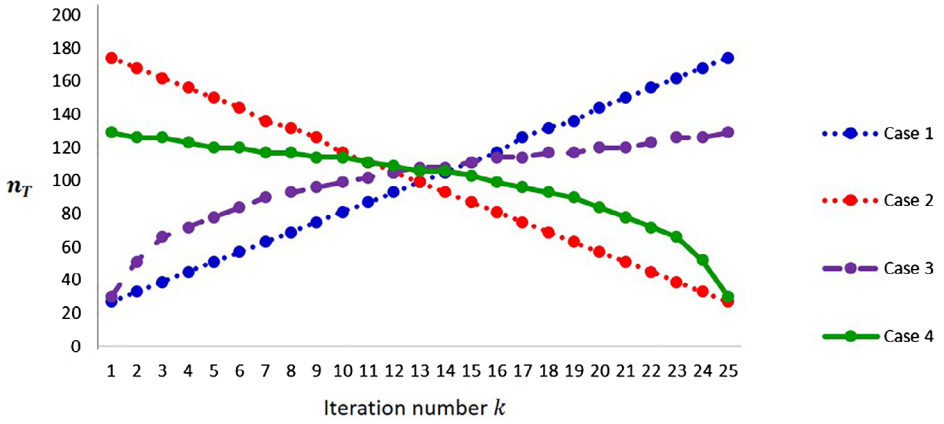

FIGURE 4 Multiple cases with different n_T sequences in A-SIS [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 8 Quantile estimation results when n_T is linearly increasing (Case 1), linearly decreasing (Case 2), logarithmically increasing (Case 3) and logarithmically decreasing (Case 4) throughout iterations

	A-SIS			A-SIS		
	Sample Std.	Avg. Diff.	MSE	Sample Std.	Avg. Diff.	MSE
Case 1	1.2	−0.7	1.9	0.9	−1.6	3.3
Case 2	1.7	0.5	3.1	1.1	−1.0	2.0
Case 3	1.0	−0.9	1.7	0.8	−1.6	3.2
Case 4	1.9	0.5	3.7	1.2	−0.9	2.1

In the proposed sequential procedure, we need to further decide n_T, given the total resource $K \cdot n_T$. In our study, we assign an equal budget to all K iterations with a fixed n_T. A small n_T (or large K) increases the variance of individual POE estimator in (16) at each iteration. On the other hand, with large n_T (or small K), θ_k may not converge to the target quantile, given the fixed budget $K \cdot n_T$. To handle this trade-off, one possible way is to use different sample sizes at each iteration, considering potentially different variances of individual POE estimators in (16) over k. We empirically evaluate the estimation performance with different forms of n_T throughout iterations. We consider multiple cases where n_T is linearly increasing (Case 1), linearly decreasing (Case 2), logarithmically increasing (Case 3) and logarithmically decreasing (Case 4), as shown in Figure 4. The total budget is set to be 2500 in all cases. Table 8 summarizes the results, indicating that there are no clear patterns in the estimation performance.

Although varying the budget allocation throughout iterations do not show clear benefits in this example, such treatment could further enhance the IS procedure in our adaptive framework. We hope to extend our framework for further improving the budget allocation rules and analyzing theoretical properties with adaptive sample sizes in our future study.

5 | EXAMPLE 2

We evaluate the proposed approach for a multi-dimensional input case. Let p denote the dimension of the input vector.

TABLE 9 Quantile estimation results with multidimensional input vector (In SA, $a = 25$, 50, and 50 are used in SA for $p = 2$, 3, and 5, respectively)

p	y_α	Methods	Sample Std.	Avg. diff	MSE
2	26.0	A-SIS	3.3	−1.3	12.4
		A-SIS	2.3	−3.9	20.3
		NA-SIS	1.5	−4.8	25.4
		SA	1.2	−8.0	65.8
3	28.8	A-SIS	3.1	−1.2	10.8
		A-SIS	2.6	−3.1	15.9
		NA-SIS	1.8	−4.7	25.5
		SA	2.8	8.3	76.7
5	33.2	A-SIS	2.8	−2.0	11.8
		A-SIS	2.0	−4.3	21.9
		NA-SIS	2.8	−5.0	32.5
		SA	2.7	6.2	45.9

We consider the following data generating structure.

$$\mathbf{X} \sim MVN(0, \sigma^2_{\mathbf{X}} \cdot I_{p\times p}) \tag{30}$$

$$Y|\mathbf{X} \sim N(\mu(\mathbf{X}), \sigma^2_{Y|\mathbf{X}}), \tag{31}$$

with $\sigma^2_{\mathbf{X}} = 5$, $\mu(\mathbf{X}) = \|\mathbf{X}\|_2$ and $\sigma^2_{Y|\mathbf{X}} = \|\mathbf{X}\|_2$, where $\|\cdot\|_2$ denotes a 2-norm. We investigate the quantile estimation for $\alpha = 10^{-4}$ with the same baseline parameter setting in Example 1.

Table 9 summarizes the results from 25 experiments, assuming the perfect metamodel. The SA performance greatly varies, depending on a. We test the SA performance with different values of a and choose the value that provides small performance error. While our adaptive procedure's standard deviation is comparable to those in NA-SIS and NA, it estimates the true quantile much closely, resulting in smaller average difference and MSE.

6 | WIND TURBINE CASE STUDY

This section estimates the extreme load response in a wind turbine using the set of NREL simulators, TurbSim (version

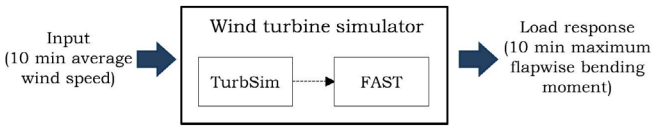

FIGURE 5 Simulation process using NREL's wind turbine simulator [Colour figure can be viewed at wileyonlinelibrary.com]

1.50) (B. J. Jonkman, 2009) and FAST (version 7.01.00a-bjj) (J. M. Jonkman & Buhl, 2005). Following the design specification in the international standard, IEC 61400-1 (International Electrotechnical Commission, 2005), we consider the 10-minute average wind speed as the simulation input, which is assumed to follow a truncated Rayleigh distribution on the interval [3, 25] (m/s) with the scale parameter of $\sqrt{2/\pi} \cdot 10$. Given the input wind speed, Turbsim generates stochastic turbulence around blade rotor plane. In doing so, 8 million random variables (ξ) are used, but TurbSim itself automatically draws ξ from its density embedded inside TurbSim. Because the density of ξ is hidden inside TurbSim, a simulator user does not know its density and is not allowed to sample ξ. Then, taking the turbulence generated from Turbsim, FAST generates structural responses such as flapwise bending moment (see Figure 5).

In particular, we consider the 10-minute maximum flapwise bending moment, which is one of important load types in the wind turbine reliability analysis (Byon, Choe, & Yampikulsakul, 2016; Moriarty, 2008; Yampikulsakul, Byon, Huang, Sheng, & You, 2014). In Choe et al. (2015, 2018), the flapwise bending moments were calculated using the results from the FAST outputs, following the procedure in Moriarty (2008). This study uses a newer version of TurbSim and obtain the flapwsie bending moments directly generated from FAST (Manuel, Nguyen, & Barone, 2013). The CPU time for each run takes about 1 minute.

In NREL simulators we approximate the conditional failure probability $s(\mathbf{x}; \theta)$, as suggested in Choe et al. (2015). Specifically, we fit a nonhomogeneous GEV distribution under the GAMLSS framework with small-scale pilot samples obtained from 600 runs. We model the location and scale parameters of the GEV distribution using the cubic smoothing spline functions of input and estimate the model parameters that maximize the log-likelihood penalized by the roughness of parameters (Rigby & Stasinopoulos, 2005). We evaluate the goodness-of-fit using the Kolmogorov-Smirnov test. Detailed procedure for approximating $s(\mathbf{x}; \theta)$ is available in Choe et al. (2015).

We conduct 25 experiments with $\theta_1 = 12\ 000$ (kNm). Table 10 summarizes the estimation results for $\alpha = 10^{-4}$. The theoretical quantile, estimated from 1 050 000 CMC samples, is $y_\alpha \approx 15\ 589$. The average difference results suggest the estimated quantiles from A-SIS$_1$ and A-SIS$_2$ are closer to y_α than those from NA-SIS and SA. The proposed approach also generates smaller sample standard deviations and MSEs.

In comparison with CMC, we conduct 25 experiments each with 10^4 runs and obtain the sample standard deviation, average difference and MSE of 386.2, −257.5 and 214,922.6, respectively. Note that A-SIS uses 600 pilot samples and 2500 runs in each experiment. Therefore, even accounting for the overhead of constructing the metamodel with 600 samples, A-SIS achieves much better estimation performance than CMC with a smaller than one third of CMC computational runs.

TABLE 10 Quantile estimation results for flapwise bending moment (unit: kNm)

Methods	Sample Std.	Avg. diff	MSE
A-SIS$_1$	131.5	86.6	24 100.2
A-SIS$_2$	136.5	−38.2	19 346.6
NA-SIS	201.6	−266.9	110 250.0
SA	165.8	−855.2	757 907.0

7 | SUMMARY

This study aims at efficiently estimating the quantile (or resistance level) for satisfying the required reliability level with stochastic black box computer models. The focus in reliability analysis is on rare events in the tail portion in the output density, which means that one does not have much information to start with and nor is it easier to get many relevant, valuable data points when one simply runs the simulator blindly. In the context of computationally expensive simulations especially, being able to select high-quality inputs can save tremendous computational resources in the quantile estimation. Our contribution is to extend the nonadaptive sampling structure of SIS (Choe et al., 2015) in order to informatively adjust the IS density with justification on convergence properties. Numerical evidence through numerical examples and a wind turbine case study shows that our proposed method elicits substantial computational improvements over the alternatives, which makes the resulting method much closer to being practical.

The proposed method requires the knowledge of the conditional POE. In this study we approximate it using a statistical metamodel. Building a metamodel incurs computational overhead, but it is needed to derive the simulation process effectively. Although our numerical studies indicate that the proposed approach is robust to the metamodel quality, building a high-quality metamodel can be of significant benefit. In the future, we plan to explore other metamodel techniques, depending on application contexts. For example, in our wind turbine case study, the nonhomogeneous GEV distribution provides a good fit (Choe et al., 2015). In the financial risk analysis, the delta-gamma approximation (Glasserman et al., 2000) and nonparametric approach (Hong et al., 2017) are shown to be effective. On the other hand, developing the general metamodeling methodology, or providing useful guidelines in the metamodel development, is needed. We will study the metamodeling techniques tailored to the IS

procedure, which can be generally applicable to a wide range of applications.

The IS scheme for high-dimensional problem is considered challenging in general. Our study was motivated by estimating the extreme load in a wind turbine application, which is a low-dimensional problem where our proposed scheme with the assumption of a good metamodel has merits. While our results for high-dimensional problems are promising, devising a good metamodel is challenging. For high-dimensional problems, some simple metamodels, such as quadratic (e.g., delta-gamma) or polynomial approximation, can be employed (Cannamela et al., 2008; Glasserman et al., 2000). We will further investigate other IS schemes, for example, cross-entropy method (Kurtz & Song, 2013), exponential twisting (Glasserman et al., 2000), IS with a mixture of densities (Owen & Zhou, 2000) or nonparametric densities (Hong et al., 2017; Morio, 2012; Neddermeyer, 2009; Zhang, 1996). We also plan to investigate more theoretical properties of our approach, for example, convergence rate, finite-time performance, in the future.

ACKNOWLEDGMENTS

The authors greatly appreciate editorial board members and anonymous reviewers for their thorough review and comments that helped improve the manuscript greatly. This work was partially supported by the National Science Foundation (Grant No. IIS-1741166, IIS-1849280, and CAREER CMMI-1834710) and the University of Michigan MCubed Grant. This research was also supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant No. NRF-2016R1D1A1B04933453).

ORCID

Qiyun Pan https://orcid.org/0000-0002-1988-2707
Eunshin Byon https://orcid.org/0000-0002-2506-1606
Young Myoung Ko https://orcid.org/0000-0003-0659-6688

REFERENCES

Ankenman, B., Nelson, B. L., & Staum, J. (2010). Stochastic kriging for simulation metamodeling. *Operations Research*, *58*, 371–382.

Asmussen, S., & Glynn, P. W. (2007). Stochastic simulation: Algorithms and analysis. Berlin: Springer Science and Business Media.

Au, S., & Beck, J. (1999). A new adaptive importance sampling scheme for reliability calculations. *Structural Safety*, *21*(2), 135–158.

Ba, S., & Joseph, V. R. (2012). Composite Gaussian process models for emulating expensive functions. *The Annals of Applied Statistics*, *6*, 1838–1860.

Balesdent, M., Morio, J., & Marzat, J. (2013). Kriging-based adaptive importance sampling algorithms for rare event estimation. *Structural Safety*, *44*, 1–10.

Bardou, O., Frikha, N., & Pages, G. (2009). Computing VaR and CVaR using stochastic approximation and adaptive unconstrained importance sampling. *Monte Carlo Methods and Applications*, *15*, 173–210.

Bastos, L. S., & O'Hagan, A. (2009). Diagnostics for Gaussian process emulators. *Technometrics*, *51*, 425–438.

Binois, M., Huang, J., Gramacy, R. B., & Ludkovski, M. (2019). Replication or exploration? Sequential design for stochastic simulation experiments. *Technometrics*, *61*(1), 7–23.

Broadie, M., Du, Y., & Moallemi, C. C. (2011). Efficient risk estimation via nested sequential simulation. *Management Science*, *57*(6), 1172–1194.

Bucklew, J. A. (2004). Introduction to rare event simulation. Berlin: Springer-Verlag.

Bugallo, M. F., Elvira, V., Martino, L., Luengo, D., Miguez, J., & Djuric, P. M. (2017). Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, *34*(4), 60–79.

Bulteau, S., & El Khadiri, M. (2002). A new importance sampling Monte Carlo method for a flow network reliability problem. *Naval Research Logistics*, *49*(2), 204–228.

Byon, E., Choe, Y., & Yampikulsakul, N. (2016, April). Adaptive learning in time-variant processes with application to wind power systems. *IEEE Transactions on Automation Science and Engineering*, *13*(2), 997–1007.

Cannamela, C., Garnier, J., & Iooss, B. (2008). Controlled stratification for quantile estimation. *The Annals of Applied Statistics*, *2*, 1554–1580.

Chen, X., Nelson, B. L., & Kim, K. (2012). *Stochastic kriging for conditional value-at-risk and its sensitivities*. In Proceedings of the 2012 winter simulation conference (pp. 1–12), Piscataway, New Jersey.

Choe, Y., Byon, E., & Chen, N. (2015). Importance sampling for reliability evaluation with stochastic simulation models. *Technometrics*, *57*, 351–361.

Choe, Y., Lam, H., & Byon, E. (2018). Uncertainty quantification of stochastic simulation for black-box computer experiments. *Methodology and Computing in Applied Probability*, *20*(4), 1155–1172.

Choe, Y., Pan, Q., & Byon, E. (2016). Computationally efficient uncertainty minimization in wind turbine extreme load assessments. *ASME Journal of Solar Energy Engineering: Including Wind Energy and Building Energy Conservation*, *138*, 041012: 1–8.

Chu, F., & Nakayama, M. K. (2012). Confidence intervals for quantiles when applying variance-reduction techniques. *ACM Trans. Model. Comput. Simul.*, *22*, 10:1–10:25.

Cornuet, J.-M., Marin, J.-M., Mirea, A., & Robert, C. P. (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, *39*(4), 798–812.

Egloff, D., & Leippold, M. (2010). Quantile estimation with adaptive importance sampling. *The Annals of Statistics*, *38*, 1244–1278.

Elvira, V., Martino, L., Luengo, D., & Bugallo, M. F. (2017). Improving population Monte Carlo: Alternative weighting and resampling schemes. *Signal Processing*, *131*, 77–91.

Elvira, V., Martino, L., Luengo, D., & Bugallo, M. F. (2019). Generalized multiple importance sampling. *Statistical Science*, *34*(1), 129–155.

Glasserman, P., Heidelberger, P., & Perwez, S. (1999). *Importance sampling and stratification for value-at-risk*. In Proceedings of the sixth international conference on computational finance (pp. 7–24). Cambridge, MA: MIT Press.

Glasserman, P., Heidelberger, P., & Shahabuddin, P. (2000). Variance reduction techniques for estimating value-at-risk. *Management Science*, *46*(10), 1349–1364.

Glynn, P. W. (1996). *Importance sampling for Monte Carlo estimation of quantiles*. In Mathematical methods in stochastic simulation and

experimental design: Proceedings of the 2nd St. Petersburg workshop on simulation, St Petersburg, Russia: Publishing House of Saint Petersburg University.

Gordy, M. B., & Juneja, S. (2010). Nested simulation in portfolio risk measurement. *Management Science*, *56*(10), 1833–1848.

Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, *37*, 185–194.

Hong, L. J., Juneja, S., & Liu, G. (2017). Kernel smoothing for nested estimation with application to portfolio risk measurement. *Operations Research*, *65*(3), 657–673.

International Electrotechnical Commission. (2005). *Wind turbines—Part 1: Design requirements*, IEC/TC88,61400-1 ed.3.

Jonkman, B. J. (2009). *Turbsim user's guide: Version 1.50*, National Renewable Energy Laboratory, Golden, CO, Technical Report No. NREL/TP-500-46198.

Jonkman, J. M., & Buhl, M. L. (2005). *Fast user's guide*. National Renewable Energy Laboratory, Golden, CO, Technical Report No. NREL/EL-500-38230.

Kohler, M., KrzyZak, A., & Walk, H. (2014). Nonparametric recursive quantile estimation. *Statistics and Probability Letters*, *93*, 102–107.

Kurtz, N., & Song, J. (2013). Cross-entropy-based adaptive importance sampling using Gaussian mixture. *Structural Safety*, *42*, 35–44.

Kushner, H., & Yin, G. G. (2003). Stochastic approximation and recursive algorithms and applications. Berlin: Springer Science & Business Media.

Lee, G., Byon, E., Ntaimo, L., & Ding, Y. (2013). Bayesian spline method for assessing extreme loads on wind turbines. *Annals of Applied Statistics*, *7*(4), 2034–2061.

Manuel, L., Nguyen, H. H., & Barone, M. F. (2013). *On the use of a large database of simulated wind turbine loads to aid in assessing design standard provisions*. In Proceedings of the 51st AIAA aerospace sciences meeting including the new horizons forum and aerospace exposition. The conference was held in Grapevine, TX: American Institute of Aeronautics and Astronautics.

Mathworks. (2004). *An adventure of sorts-behind the scenes of a MATLAB upgrade*. Retrieved from http://www.mathworks.com/company/newsletters/articles/an-adventure-of-sortsbehind-the-scenes-of-a-matlab-upgrade.html.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Moriarty, P. (2008). Database for validation of design load extrapolation techniques. *Wind Energy*, *11*, 559–576.

Morio, J. (2012). Extreme quantile estimation with nonparametric adaptive importance sampling. *Simulation Modelling Practice and Theory*, *27*, 76–89.

Muresan, M. M. (2009). A concrete approach to classical analysis. New York: Springer.

Neddermeyer, J. C. (2009). Computationally efficient nonparametric importance sampling. *Journal of the American Statistical Association*, *104*, 788–802.

Oakley, J. (2004). Estimating percentiles of uncertain computer code outputs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*, 83–93.

Owen, A. (2013). *Monte carlo theory, methods and examples*. Retrieved from https://statweb.stanford.edu/owen/c/

Owen, A., & Zhou, Y. (2000). Safe and effective importance sampling. *Journal of the American Statistical Association*, *95*, 135–143.

Pan, Q., Byon, E., & Ko, Y. M. (2020). Uncertainty quantification for extreme quantile estimation with stochastic computer models. *IEEE Transactions on Reliability*, https://ieeexplore-ieee-org.proxy.lib.umich.edu/abstract/document/9067075?casa_token=a14_RxkAKysAAAAA:R2IupYTSI0vNUaolonGmViHW6mFLKGKZ813mWkjxwCHgO8G-9s80Vk32qh4NeqG7kS7-r_F59g. DOI: 10.1109/TR.2020.2980448.

Polyak, B. T. (1990). New stochastic approximation type procedures. *Automation and Remote Control*, *7*, 937–1008.

Ragan, P., & Manuel, L. (2008). Statistical extrapolation methods for estimating wind turbine extreme loads. *Journal of Solar Energy Engineering*, *130*, 031011: 1–15.

Ranjan, P., Bingham, D., & Michailidis, G. (2008). Sequential experiment design for contour estimation from complex computer codes. *Technometrics*, *50*, 527–541.

Rigby, R. A., & Stasinopoulos, M. D. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*, 507–554.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*, 400–407.

Ruppert, D. (1988). *Efficient estimations from a slowly convergent Robbins–Monro process* (Tech. Rep.). Cornell University Operations Research and Industrial Engineering.

Shi, W., & Chen, X. (2018). Efficient budget allocation strategies for elementary effects method in stochastic simulation. *Naval Research Logistics*, *65*(3), 218–241.

Sun, Y., Apley, D. W., & Staum, J. (2011). Efficient nested simulation for estimating the variance of a conditional expectation. *Operations Research*, *59*, 998–1007.

Wang, B., & Hu, J. (2015). *On the monotonic performance of stochastic kriging predictors*. In Proceedings of the 2015 winter simulation conference (pp. 3825–3833). Huntington Beach, CA.

Yampikulsakul, N., Byon, E., Huang, S., Sheng, S., & You, M. (2014). Condition monitoring of wind power system with nonparametric regression analysis. *IEEE Transactions on Energy Conversion*, *29*(2), 288–299.

Yang, F., Ankenman, B., & Nelson, B. L. (2007). Efficient generation of cycle time-throughput curves through simulation and metamodeling. *Naval Research Logistics*, *54*(1), 78–93.

Zhang, P. (1996). Nonparametric importance sampling. *Journal of the American Statistical Association*, *91*, 1245–1253.

How to cite this article: Pan Q, Byon E, Ko YM, Lam H. Adaptive importance sampling for extreme quantile estimation with stochastic black box computer models. *Naval Research Logistics* 2020;1–24. https://doi.org/10.1002/nav.21938

APPENDIX: PROOFS AND DERIVATIONS

A1 | Proof of Lemma 1

Recall that the combined POE estimator $\widehat{P}_{1:K}(y)$ is given by.

$$\widehat{P}_{1:K}(y) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{\widetilde{n}_{i,k}}\sum_{j=1}^{\widetilde{n}_{i,k}}\mathbb{I}(Y_{ij,k} > y)\right)\frac{p(\mathbf{X}_{i,k})}{\widetilde{q}(\mathbf{X}_{i,k};\theta_k)}\right),$$

where

$$\widetilde{q}(x;\theta_k) = \frac{1}{C_{\widetilde{q}}}p(x)\sqrt{\widetilde{s}(\mathbf{x};\theta_k)}\sqrt{\frac{1}{n_T} + \left(1 - \frac{1}{n_T}\right)\widetilde{s}(\mathbf{x};\theta_k)}.$$

We obtain the variance bounds of the individual and combined estimators.

(1) Bound of $Var[\widehat{P}_k(y)]$: From the fact that

$$\widetilde{s}(\mathbf{x};\theta_k) \geq \frac{\delta}{k^{\beta}},$$

$$\sqrt{\frac{1}{n_T} + \left(1 - \frac{1}{n_T}\right)\widetilde{s}(\mathbf{x};\theta_k)} \geq \sqrt{\frac{1}{n_T}},$$

$$C_{\widetilde{q}} = \int p(x)\frac{\sqrt{\widetilde{s}(\mathbf{x};\theta_k)}\sqrt{1 + (n_T - 1)\widetilde{s}(\mathbf{x};\theta_k)}}{\sqrt{n_T}}dx \leq 1,$$

we obtain the bound of the likelihood ratio as follows.

$$\begin{aligned}\frac{p(x)}{\widetilde{q}(x;\theta_k)} &= \frac{C_{\widetilde{q}}}{\sqrt{\widetilde{s}(\mathbf{x};\theta_k)}\sqrt{\frac{1}{n_T} + \left(1 - \frac{1}{n_T}\right)\widetilde{s}(\mathbf{x};\theta_k)}} \\ &\leq \frac{\sqrt{n_T}}{\sqrt{\delta}}k^{\frac{\beta}{2}} \\ &\leq \frac{\sqrt{n_T}}{\sqrt{\delta}}K^{\frac{\beta}{2}} \\ &= DK^{\frac{\beta}{2}}, \end{aligned} \tag{A1}$$

where $D = \sqrt{n_T}/\sqrt{\delta} < \infty$.

Using (A1), we now have a bound for $\widehat{P}_k(y)$ as follows:

$$\begin{aligned}\widehat{P}_k(y) &= \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{\widetilde{n}_{i,k}}\sum_{j=1}^{\widetilde{n}_{i,k}}\mathbb{I}(Y_{ij,k} > y)\right)\frac{p(\mathbf{X}_{i,k})}{\widetilde{q}(\mathbf{X}_{i,k};\theta_k)} \\ &\leq DK^{\frac{\beta}{2}}. \end{aligned} \tag{A2}$$

(2) Bound of $Var[\widehat{P}_{1:K}(y)]$: We first show that $\widehat{P}_k(y)$ is an unbiased estimator of $\mathbb{P}(Y > y)$ as follows:

$$\begin{aligned}E[\widehat{P}_k(y)] &= E_{\mathbf{X}_{1:k-1},\mathbf{Y}_{1:k-1}}E[\widehat{P}_k(y)|\mathbf{X}_{1:k-1},\mathbf{Y}_{1:k-1}] \\ &= E_{\mathbf{X}_{1:k-1},\mathbf{Y}_{1:k-1}}E[\widehat{P}_k(y)|\theta_k] \\ &= E_{\mathbf{X}_{1:k-1},\mathbf{Y}_{1:k-1}}[\mathbb{P}(Y > y)], \\ &= E_{\mathbf{X}_{1:k-1},\mathbf{Y}_{1:k-1}}[\mathbb{P}(Y > y)], \end{aligned}$$

where the subscript, $1:k-1$, implies the data samples obtained up to the $(k-1)^{th}$ iteration. Thus, we get

$$E[\widehat{P}_{1:K}(y)] = E\left[\frac{1}{K}\sum_{k=1}^{K}\widehat{P}_k(y)\right] = \mathbb{P}(Y > y).$$

Noting that $\widehat{P}_h(y) - \mathbb{P}(Y > y)$ has conditional mean 0, given all previous information (or equivalently given θ_h), we get $Cov[\widehat{P}_h(y), \widehat{P}_l(y)] = 0$. Specifically, for any $1 \leq h < l \leq K$, it holds.

$$E[\widehat{P}_h(y)\widehat{P}_l(y)] = E_{\mathbf{X}_{1:h},\mathbf{Y}_{1:h}}E_{\mathbf{X}_{1:l},\mathbf{Y}_{1:l}|\mathbf{X}_{1:h},\mathbf{Y}_{1:h}}[\widehat{P}_h(y)\widehat{P}_l(y)|\mathbf{X}_{1:h},\mathbf{Y}_{1:h}]$$

$$
\begin{aligned}
&= E_{\boldsymbol{X}_{1:h},\boldsymbol{Y}_{1:h}}[\widehat{P}_h(y)E_{\boldsymbol{X}_{1:l},\boldsymbol{Y}_{1:l}|\boldsymbol{X}_{1:h},\boldsymbol{Y}_{1:h}}[\widehat{P}_l(y)|\mathbf{X}_{1:h},\mathbf{Y}_{1:h}]]\\
&= E_{\mathbf{X}_{1:h},\boldsymbol{Y}_{1:h}}[\widehat{P}_h(y)E_{\boldsymbol{X}_{1:l},\boldsymbol{Y}_{1:l}|\boldsymbol{X}_{1:h},\boldsymbol{Y}_{1:h}}[\widehat{P}_l(y)|\theta_h]]\\
&= \mathbb{P}(Y>y)E_{\boldsymbol{X}_{1:h},\boldsymbol{Y}_{1:h}}[\widehat{P}_h(y)]\\
&= \mathbb{P}^2(Y>y),
\end{aligned}
\tag{A3}
$$

where the second equality holds because, given $\boldsymbol{X}_{1:h}$, $\boldsymbol{Y}_{1:h}$, $\widehat{P}_h(y)$ can be treated as a constant and the second last equality holds because of $E_{\boldsymbol{X}_{1:l},\boldsymbol{Y}_{1:l}|\boldsymbol{X}_{1:h},\boldsymbol{Y}_{1:h}}[\widehat{P}_l(y)|\theta_h]] = \mathbb{P}(Y>y)$. Then, (A3) implies $Cov[\widehat{P}_h(y),\widehat{P}_l(y)]=0$.

Having proved that $\widehat{P}_k(y)$ is unbiased and $Cov[\widehat{P}_h(y),\widehat{P}_l(y)]=0$, from (A2) we can obtain an upper bound of the variance of $\widehat{P}_{1:K}(y)$ as

$$
\begin{aligned}
Var[\widehat{P}_{1:K}(y)] &= Var\left[\frac{1}{K}\sum_{k=1}^{K}\widehat{P}_k(y)\right]\\
&\le D^2K^{\beta-1}.
\end{aligned}
$$

A2 | Proof of Theorem 1

(1) Proof of $\widehat{P}_{1:K}(y) \xrightarrow{P} \mathbb{P}(Y>y)$: Because $\widehat{P}_{1:K}(y)$ is the unbiased estimator for $\mathbb{P}(Y>y)$, we use Chebyshev's inequality to obtain

$$
\mathbb{P}(\widehat{P}_{1:K}(y)-\mathbb{P}(Y>y)|>\varepsilon) \le \frac{1}{\varepsilon^2}Var[\widehat{P}_{1:K}(y)] = \frac{D^2K^{\beta-1}}{\varepsilon^2}. \tag{A4}
$$

Therefore, for $0<\beta<1$, we attain $\widehat{P}_{1:K}(y) \xrightarrow{P} \mathbb{P}(Y>y)$, $\forall y\in\Omega_Y$.

(2) Proof of $\widehat{\mathbf{P}}_{\mathbf{1:K}}(\mathbf{y}) \xrightarrow{\mathbf{a.s.}} \mathbb{P}(\mathbf{Y}>\mathbf{y})$: Let $K=n^2$. Then, by the Chebyshev's inequality, we have

$$
\mathbb{P}(|\widehat{P}_{1:n^2}(y)-\mathbb{P}(Y>y)|>\varepsilon) \le \frac{D^2}{\varepsilon^2}n^{2\beta-2}. \tag{A5}
$$

For $0<\beta<1/2$, we know that the series consisting of (A5) converges, that is

$$
\sum_{n=1}^{\infty}\mathbb{P}(|\widehat{P}_{1:n^2}(y)-\mathbb{P}(Y>y)|>\varepsilon)<\infty.
$$

Then, by the Borel-Cantelli lemma, we have the following almost sure convergence result:

$$
\widehat{P}_{1:n^2}(y) \xrightarrow{a.s.} \mathbb{P}(Y>y),
$$

which implies that $\forall y\in\Omega_Y$, we attain

$$
\widehat{P}_{1:K}(y) \xrightarrow{a.s.} \mathbb{P}(Y>y).
$$

A3 | Proof of Corollary 1

Recall the definitions of $\widehat{y}_k^{\alpha}$ and $\widehat{y}_{k,\alpha}$:

$$
\begin{aligned}
\widehat{y}_k^{\alpha} &= \inf\{y : 0<\widehat{P}_{1:K}(y)\le\alpha\},\\
\widehat{y}_{k,\alpha} &= \sup\{y : \widehat{P}_{1:K}(y)\ge\alpha\}.
\end{aligned}
$$

From the above definitions, we get

$$
\widehat{P}_{1:K}(\widehat{y}_K^{\alpha}) \le \alpha \le \widehat{P}_{1:K}(\widehat{y}_{K,\alpha}).
$$

Note that the difference between $\widehat{P}_{1:K}(\widehat{y}_K^{\alpha})$ and $\widehat{P}_{1:K}(\widehat{y}_{K,\alpha})$ is at most one sample. We, therefore, have the following bound.

$$
\begin{aligned}
|\widehat{P}_{1:K}(\widehat{y}_K^{\alpha})-\alpha| &\le \widehat{P}_{1:K}(\widehat{y}_{K,\alpha})-\widehat{P}_{1:K}(\widehat{y}_K^{\alpha})\\
&\le \frac{1}{Km\widetilde{n}_{i,k_0}}\frac{p(\mathbf{X}_{i,k_0})}{\widetilde{q}(\mathbf{X}_{i,k_0};\theta_{k_0})}\\
&= \frac{C_{\widetilde{q}}}{Km\sqrt{n_T\widetilde{s}(\mathbf{X}_{i,k_0};\theta_{k_0})(1-\widetilde{s}(\mathbf{X}_{i,k_0};\theta_{k_0}))}}\sum_{i=1}^{m}\sqrt{\frac{1-\widetilde{s}(\mathbf{X}_{i,k_0};\theta_{k_0})}{1+(n_T-1)\widetilde{s}(\mathbf{X}_{i,k_0};\theta_{k_0})}}\\
&\le C_1K^{\beta-1},
\end{aligned}
$$

where k_0 and $\mathbf{X}_{i,k_0}$ denote the iteration index and input vector that generates $\widehat{y}_K^{\alpha}$, respectively, and $\widetilde{n}_{i,k_0}$ is the corresponding allocation size. The second last equation is obtained using (11) and (14). The last inequality holds because $\sum_{i=1}^{m}\sqrt{\frac{1-\widetilde{s}(\mathbf{X}_{i,k_0};\theta_{k_0})}{1+(n_T-1)\widetilde{s}(\mathbf{X}_{i,k_0};\theta_{k_0})}}$ and $C_{\widetilde{q}}$ are bounded, $\delta K^{-\beta} \le \delta k_0^{-\beta} \le 1-\widetilde{s}(\mathbf{X}_{i,k_0};\theta_{k_0})$ and $\delta K^{-\beta} \le \delta k_0^{-\beta} \le \widetilde{s}(\mathbf{X}_{i,k_0};\theta_{k_0})$. Therefore, for $0<\beta<1$, as $K\to\infty$, we attain

$$|\widehat{P}_{1:K}(\widehat{y}_K^{\alpha}) - \mathbb{P}(Y > y_{\alpha})| \to 0. \tag{A6}$$

On the other hand, by taking supremum on both sides of (A5), we have.

$$\sup_{y\in\Omega_Y} \mathbb{P}(|\widehat{P}_{1:K}(y) - \mathbb{P}(Y>y)| > \varepsilon) \le \frac{D^2}{\varepsilon^2} K^{\beta-1}. \tag{A7}$$

From (A7), we know that

$$\mathbb{P}(|\widehat{P}_{1:K}(\widehat{y}_K^{\alpha}) - \mathbb{P}(Y > \widehat{y}_K^{\alpha})| > \varepsilon) \le \frac{D^2}{\varepsilon^2} K^{\beta-1}, \tag{A8}$$

which implies the convergence in probability.

Based on Equations (A6) and (A8), we get $\mathbb{P}(Y>\widehat{y}_K^{\alpha}) \xrightarrow{P} \mathbb{P}(Y>y_{\alpha})$. Then from Assumption 1, it implies $\widehat{y}_K^{\alpha} \xrightarrow{P} y_{\alpha}$.

A4 | Proof of Corollary 2

Corollary 2 is obvious if Corollary 1 is true. Applying the similar procedure in Corollary 1, we get $\widehat{y}_{K,\alpha} \xrightarrow{P} y_{\alpha}$. The next density parameter θ_{k+1} is set to be $\widehat{y}_{k,\alpha}$, which impliess $\theta_K = \widehat{y}_{K-1,\alpha}$. As such we obtain $\theta_K \xrightarrow{P} y_{\alpha}$ as $K\to\infty$.

A5 | Proof of Theorem 2

We first show that the asymptotic variance of $\widehat{P}_k(y)$ is smaller than, or equal to, the CMC variance and extend the result to the asymptotic variance of $\widehat{P}_{1:K}(y)$.

(1) Proof of $\lim_{k\to\infty} Var[\widehat{P}_k(y)] \le \frac{\alpha(1-\alpha)}{n_T}$: We apply the results in Corollary 2 to the analytical form of $Var[\widehat{P}_k(y)]$. First, to obtain $Var[\widehat{P}_k(y)]$, we use the total law of variance to get.

$$\begin{aligned} Var[\widehat{P}_k(y)] &= \mathbb{E}_{\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}} Var[\widehat{P}_k(y)|\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}] + Var_{\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}} E[\widehat{P}_k(y)|\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}] \\ &= \mathbb{E}_{\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}} Var[\widehat{P}_k(y)|\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}], \end{aligned} \tag{A9}$$

for $k>1$, where the second equality holds because the second term in (A9) vanishes because $E[\widehat{P}_k(y)|\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}] = E[\widehat{P}_k(y)|\theta_k] = \mathbb{P}(Y>y)$, which is constant. By applying the total law of variance again, we have

$$\begin{aligned} &Var[\widehat{P}_k(y)|\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}] \\ &= \mathbb{E}_{\boldsymbol{X}_k|\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}} Var_{Y_k|\boldsymbol{X}_{1:k},\boldsymbol{Y}_{1:k-1}} \left(\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{\widetilde{n}_{i,k}}\sum_{j=1}^{\widetilde{n}_{i,k}} \mathbb{I}(Y_{ij,k}>y)\right)\frac{p(\boldsymbol{X}_{i,k})}{\widetilde{q}(\boldsymbol{X}_{i,k};\theta_k)} |\boldsymbol{X}_{1:k},\boldsymbol{Y}_{1:k-1}\right) \\ &\quad + Var_{\boldsymbol{X}_k|\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}} \mathbb{E}_{Y_k|\boldsymbol{X}_{1:k},\boldsymbol{Y}_{1:k-1}} \left(\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{\widetilde{n}_{i,k}}\sum_{j=1}^{\widetilde{n}_{i,k}} \mathbb{I}(Y_{ij,k}>y)\right)\frac{p(\boldsymbol{X}_{i,k})}{\widetilde{q}(\boldsymbol{X}_{i,k};\theta_k)} |\boldsymbol{X}_{1:k},\boldsymbol{Y}_{1:k-1}\right) \\ &= \mathbb{E}_{\boldsymbol{X}_k|\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}} \left(\frac{1}{m^2} Var_{Y_k|\boldsymbol{X}_{1:k},\boldsymbol{Y}_{1:k-1}} \left(\sum_{i=1}^{m}\left(\frac{1}{\widetilde{n}_{i,k}}\sum_{j=1}^{\widetilde{n}_{i,k}} \mathbb{I}(Y_{ij,k}>y)\right)\frac{p(\boldsymbol{X}_{i,k})}{\widetilde{q}(\boldsymbol{X}_{i,k};\theta_k)} |\boldsymbol{X}_k,\theta_k\right)\right) \\ &\quad + Var_{\boldsymbol{X}_k|\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}} \left(\frac{1}{m}\sum_{i=1}^{m}\frac{s(\boldsymbol{X}_{i,k};y)p(\boldsymbol{X}_{i,k})}{\widetilde{q}(\boldsymbol{X}_{i,k};\theta_k)}\right) \\ &= \frac{1}{m^2}\sum_{i=1}^{m}\left(\mathbb{E}_{\boldsymbol{X}_k|\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}} \left(\frac{s(\boldsymbol{X}_{i,k};y)(1-s(\boldsymbol{X}_{i,k};y))}{\widetilde{n}_{i,k}}\frac{p^2(\boldsymbol{X}_{i,k})}{\widetilde{q}^2(\boldsymbol{X}_{i,k};\theta_k)}\right)\right. \\ &\quad \left. + Var_{\boldsymbol{X}_k|\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}} \left(\frac{s(\boldsymbol{X}_{i,k};y)p(\boldsymbol{X}_{i,k})}{\widetilde{q}(\boldsymbol{X}_{i,k};\theta_k)}\right)\right). \end{aligned} \tag{A10}$$

Here, the second equality holds because (1) given $\boldsymbol{X}_{1:k-1}$, $\boldsymbol{Y}_{1:k-1}$, θ_k is determined; (2) given θ_k and $\boldsymbol{X}_k$, $Y_{ij,k}$'s are i.i.d. Bernoulli random variables; and (3) the mean of $\mathbb{I}(Y_{ij,k} > y)$ is $s(\mathbf{X}_{i,k}; y)$. The last equality holds because $\boldsymbol{X}_{i,k}$'s are independently drawn from $\widetilde{q}(\mathbf{X}; \theta_k)$ and the variance of $\mathbb{I}(Y_{ij,k} > y)$ is $s(\mathbf{X}_{i,k}; y)(1 - s(\mathbf{X}_{i,k}; y))$.

Next, we calculate the two terms inside the innermost parentheses in (A10). Note that.

$$\frac{p(\boldsymbol{x})}{\widetilde{q}(\boldsymbol{x}; \theta_k)} = \frac{C_{\widetilde{q}}\sqrt{n_T}}{\sqrt{\widetilde{s}(\boldsymbol{x}; \theta_k)} \cdot \sqrt{1 + (n_T - 1)\widetilde{s}(\boldsymbol{x}; \theta_k)}}. \tag{A11}$$

From $\widetilde{n}_{i,k}$ in (14), we also have

$$\frac{1}{\widetilde{n}_{i,k}}\sqrt{\frac{1 - \widetilde{s}(\mathbf{X}_{i,k}; \theta_k)}{1 + (n_T - 1)\widetilde{s}(\mathbf{X}_{i,k}; \theta_k)}} = \frac{1}{n_T}\sum_{j=1}^{m}\sqrt{\frac{1 - \widetilde{s}(\mathbf{X}_{j,k}; \theta_k)}{1 + (n_T - 1)\widetilde{s}(\mathbf{X}_{j,k}; \theta_k)}}. \tag{A12}$$

From (A11) and (A12), we obtain

$$\begin{aligned}\frac{\sqrt{\widetilde{s}(\mathbf{X}_{i,k}; \theta_k)(1 - \widetilde{s}(\mathbf{X}_{i,k}; \theta_k))}p(\mathbf{X}_{i,k})}{\widetilde{n}_{i,k}\widetilde{q}(\mathbf{X}_{i,k}; \theta_k)} &= \frac{C_{\widetilde{q}}\sqrt{n_T}}{\widetilde{n}_{i,k}}\sqrt{\frac{1 - \widetilde{s}(\mathbf{X}_{i,k}; \theta_k)}{1 + (n_T - 1)\widetilde{s}(\mathbf{X}_{i,k}; \theta_k)}} \\ &= \frac{C_{\widetilde{q}}}{\sqrt{n_T}}\sum_{j=1}^{m}\sqrt{\frac{1 - \widetilde{s}(\mathbf{X}_{j,k}; \theta_k)}{1 + (n_T - 1)\widetilde{s}(\mathbf{X}_{j,k}; \theta_k)}},\end{aligned}$$

and thus, we get

$$\frac{p(\mathbf{X}_{i,k})}{\widetilde{n}_{i,k}\widetilde{q}(\mathbf{X}_{i,k}; \theta_k)} = \frac{C_{\widetilde{q}}}{\sqrt{n_T\widetilde{s}(\mathbf{X}_{i,k}; \theta_k)(1 - \widetilde{s}(\mathbf{X}_{i,k}; \theta_k))}}\sum_{j=1}^{m}\sqrt{\frac{1 - \widetilde{s}(\mathbf{X}_{j,k}; \theta_k)}{1 + (n_T - 1)\widetilde{s}(\mathbf{X}_{j,k}; \theta_k)}}. \tag{A13}$$

Then, plugging (A11) and (A13) into the quantity inside the expectation of the first term in (A10), it follows

$$\begin{aligned}&\frac{s(\mathbf{X}_{i,k}; y)(1 - s(\mathbf{X}_{i,k}; y))}{\widetilde{n}_{i,k}}\frac{p^2(\mathbf{X}_{i,k})}{\widetilde{q}^2(\mathbf{X}_{i,k}; \theta_k)} \\ &= \frac{C_{\widetilde{q}}^2 s(\mathbf{X}_{i,k}; y)(1 - s(\mathbf{X}_{i,k}; y))}{\widetilde{s}(\mathbf{X}_{i,k}; \theta_k)\sqrt{[1 + (n_T - 1)\widetilde{s}(\mathbf{X}_{i,k}; \theta_k)](1 - \widetilde{s}(\mathbf{X}_{i,k}; \theta_k))}}\sum_{j=1}^{m}\sqrt{\frac{1 - \widetilde{s}(\mathbf{X}_{j,k}; \theta_k)}{1 + (n_T - 1)\widetilde{s}(\mathbf{X}_{j,k}; \theta_k)}}.\end{aligned} \tag{A14}$$

Plugging (A14) into (A10), we obtain

$$\begin{aligned}&E_{\mathbf{X}_k|\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}}\left(\frac{s(\mathbf{X}_{i,k}; y)(1 - s(\mathbf{X}_{i,k}; y))}{\widetilde{n}_{i,k}}\frac{p^2(\mathbf{X}_{i,k})}{\widetilde{q}^2(\mathbf{X}_{i,k}; \theta_k)}\right) \\ &= E_{\mathbf{X}_k|\boldsymbol{X}_{1:k-1},\boldsymbol{Y}_{1:k-1}}\left(\frac{C_{\widetilde{q}}^2 s(\mathbf{X}_{i,k}; y)(1 - s(\mathbf{X}_{i,k}; y))}{\widetilde{s}(\mathbf{X}_{i,k}; \theta_k)\sqrt{[1 + (n_T - 1)\widetilde{s}(\mathbf{X}_{i,k}; \theta_k)](1 - \widetilde{s}(\mathbf{X}_{i,k}; \theta_k))}} \cdot \sum_{j=1}^{m}\sqrt{\frac{1 - \widetilde{s}(\mathbf{X}_{j,k}; \theta_k)}{1 + (n_T - 1)\widetilde{s}(\mathbf{X}_{j,k}; \theta_k)}}\right) \\ &= C_{\widetilde{q}}^2\left[\int \frac{s(\mathbf{x}; y)(1 - s(\mathbf{x}; y))\widetilde{q}(\mathbf{x}; \theta_k)}{\widetilde{s}(\mathbf{x}; \theta_k)\sqrt{[1 + (n_T - 1)\widetilde{s}(\mathbf{x}; \theta_k)](1 - \widetilde{s}(\mathbf{x}; \theta_k))}}\sqrt{\frac{1 - \widetilde{s}(\mathbf{x}; \theta_k)}{1 + (n_T - 1)\widetilde{s}(\mathbf{x}; \theta_k)}}d\mathbf{x}\right.\end{aligned} \tag{A15}$$

$$\left.+(m-1)\int \frac{s(\mathbf{x}; y)(1 - s(\mathbf{x}; y))\widetilde{q}(\mathbf{x}; \theta_k)}{\widetilde{s}(\mathbf{x}; \theta_k)\sqrt{[1 + (n_T - 1)\widetilde{s}(\mathbf{x}; \theta_k)](1 - \widetilde{s}(\mathbf{x}; \theta_k))}}d\mathbf{x}\int \sqrt{\frac{1 - \widetilde{s}(\mathbf{x}; \theta_k)}{1 + (n_T - 1)\widetilde{s}(\mathbf{x}; \theta_k)}}\widetilde{q}(\mathbf{x}; \theta_k)d\mathbf{x}\right]$$

$$\begin{aligned}&= \frac{C_{\widetilde{q}}}{\sqrt{n_T}}\int \frac{s(\mathbf{x}; y)(1 - s(\mathbf{x}; y))p(\mathbf{x})}{\sqrt{\widetilde{s}(\mathbf{x}; \theta_k)}}\sqrt{\frac{1}{1 + (n_T - 1)\widetilde{s}(\mathbf{x}; \theta_k)}}d\mathbf{x} \\ &\quad + \frac{m-1}{n_T}\int \frac{s(\mathbf{x}; y)(1 - s(\mathbf{x}; y))p(\mathbf{x})}{\sqrt{\widetilde{s}(\mathbf{x}; \theta_k)(1 - \widetilde{s}(\mathbf{x}; \theta_k))}}d\mathbf{x}\int \sqrt{(1 - \widetilde{s}(\mathbf{x}; \theta_k))\widetilde{s}(\mathbf{x}; \theta_k)}p(\mathbf{x})d\mathbf{x},\end{aligned} \tag{A16}$$

where (A15) is obtained because $\mathbf{X}_{j,k}$'s are independently sampled from $\widetilde{q}(\mathbf{x}; \theta_k)$ in (11) and (A16) follows by plugging $\widetilde{q}(\mathbf{x}; \theta_k)$ into (A15).Similarly, we get

$$\begin{aligned}&Var_{\mathbf{X}_k|\mathbf{X}_{1:k-1},\mathbf{Y}_{1:k-1}}\left(\frac{s(\mathbf{X}_{i,k}; y)p(\mathbf{X}_{i,k})}{\widetilde{q}(\mathbf{X}_{i,k}; \theta_k)}\right) \\ &= \int \left(\frac{s(\mathbf{x}; y)p(\mathbf{x})}{\widetilde{q}(\mathbf{x}; \theta_k)}\right)^2 \widetilde{q}(\mathbf{x}; \theta_k)d\mathbf{x} - \left(\int \frac{s(\mathbf{x}; y)p(\mathbf{x})}{\widetilde{q}(\mathbf{x}; \theta_k)}\widetilde{q}(\mathbf{x}; \theta_k)d\mathbf{x}\right)^2\end{aligned}$$

$$= C_{\widetilde{q}} \int \frac{s^2(\mathbf{x}; y)p(\mathbf{x})\sqrt{n_T}}{\sqrt{\widetilde{s}(\mathbf{x}; \theta_k)}\sqrt{1 + (n_T - 1)\widetilde{s}(\mathbf{x}; \theta_k)}} d\mathbf{x} - \left(\int s(\mathbf{x}; y)p(\mathbf{x})d\mathbf{x} \right)^2$$

$$= C_{\widetilde{q}} \int \frac{s^2(\mathbf{x}; y)p(\mathbf{x})\sqrt{n_T}}{\sqrt{\widetilde{s}(\mathbf{x}; \theta_k)}\sqrt{1 + (n_T - 1)\widetilde{s}(\mathbf{x}; \theta_k)}} d\mathbf{x} - \mathbb{P}^2(Y > y). \tag{A17}$$

Applying the results in (A10), (A16) and (A17) to (A9), $Var[\widehat{P}_k(y)]$ becomes

$$Var[\widehat{P}_k(y)] = \frac{1}{m}\mathbb{E}_{\mathbf{X}_{1:k-1}, \mathbf{Y}_{1:k-1}} \left[C_{\widetilde{q}}\sqrt{n_T} \int \frac{s(\mathbf{x}; y)p(\mathbf{x}) \left[\frac{1}{n_T}(1 + (n_T - 1)s(\mathbf{x}; y)) \right]}{\sqrt{\widetilde{s}(\mathbf{x}; \theta_k)}\sqrt{1 + (n_T - 1)\widetilde{s}(\mathbf{x}; \theta_k)}} d\mathbf{x} \right.$$

$$\left. + \frac{m-1}{n_T} \int \frac{s(\mathbf{x}; y)(1 - s(\mathbf{x}; y))p(\mathbf{x})}{\sqrt{\widetilde{s}(\mathbf{x}; \theta_k)(1 - \widetilde{s}(\mathbf{x}; \theta_k))}} d\mathbf{x} \int \sqrt{(1 - \widetilde{s}(\mathbf{x}; \theta_k))\widetilde{s}(\mathbf{x}; \theta_k)}p(\mathbf{x})d\mathbf{x} \right] - \frac{\mathbb{P}^2(Y > y)}{m}. \tag{A18}$$

where the normalizing constant, $C_{\widetilde{q}}$, is given by

$$C_{\widetilde{q}} = \int p(\mathbf{x}) \frac{\sqrt{\widetilde{s}(\mathbf{x}; \theta_k)} \cdot \sqrt{1 + (n_T - 1)\widetilde{s}(\mathbf{x}; \theta_k)}}{\sqrt{n_T}} d\mathbf{x}. \tag{A19}$$

Finally, by plugging $C_{\widetilde{q}}$ in (A19) into (A18), $Var[\widehat{P}_k(y)]$ becomes

$$Var[\widehat{P}_k(y)]$$

$$= \frac{1}{m}\mathbb{E}_{\mathbf{X}_{1:k-1}, \mathbf{Y}_{1:k-1}} \left[\int p(\mathbf{x})\sqrt{\widetilde{s}(\mathbf{x}; \theta_k)}\sqrt{1 + (n_T - 1)\widetilde{s}(\mathbf{x}; \theta_k)}d\mathbf{x} \int \frac{p(\mathbf{x})s(\mathbf{x}; y) \left[\frac{1}{n_T}(1 + (n_T - 1)s(\mathbf{x}; y)) \right]}{\sqrt{\widetilde{s}(\mathbf{x}; \theta_k)}\sqrt{1 + (n_T - 1)\widetilde{s}(\mathbf{x}; \theta_k)}} d\mathbf{x} \right.$$

$$\left. + \frac{m-1}{n_T} \int \frac{s(\mathbf{x}; y)(1 - s(\mathbf{x}; y))p(\mathbf{x})}{\sqrt{\widetilde{s}(\mathbf{x}; \theta_k)(1 - \widetilde{s}(\mathbf{x}; \theta_k))}} d\mathbf{x} \int \sqrt{(1 - \widetilde{s}(\mathbf{x}; \theta_k))\widetilde{s}(\mathbf{x}; \theta_k)}p(\mathbf{x})d\mathbf{x} \right] - \frac{\mathbb{P}^2(Y > y)}{m}. \tag{A20}$$

From Corollary 2, we have $\theta_k \xrightarrow{P} y_\alpha$, as $k \to \infty$. Therefore, as $k \to \infty$, the quantities in (A20) converge to their corresponding values as

$$p(\mathbf{x})\sqrt{\widetilde{s}(\mathbf{x}; \theta_k)}\sqrt{1 + (n_T - 1)\widetilde{s}(\mathbf{x}; \theta_k)} \xrightarrow{P} p(\mathbf{x})\sqrt{s(\mathbf{x}; y_\alpha)}\sqrt{1 + (n_T - 1)s(\mathbf{x}; y_\alpha)}, \tag{A21}$$

$$\frac{p(\mathbf{x})s(\mathbf{x}; y_\alpha)[1 + (n_T - 1)s(\mathbf{x}; y_\alpha)]}{\sqrt{\widetilde{s}(\mathbf{x}; \theta_k)}\sqrt{1 + (n_T - 1)\widetilde{s}(\mathbf{x}; \theta_k)}} \xrightarrow{P} p(\mathbf{x})\sqrt{s(\mathbf{x}; y_\alpha)}\sqrt{1 + (n_T - 1)s(\mathbf{x}; y_\alpha)}, \tag{A22}$$

$$\frac{s(\mathbf{x}; y_\alpha)(1 - s(\mathbf{x}; y_\alpha))p(\mathbf{x})}{\sqrt{\widetilde{s}(\mathbf{x}; \theta_k)(1 - \widetilde{s}(\mathbf{x}; \theta_k))}} \xrightarrow{P} \sqrt{s(\mathbf{x}; y_\alpha)(1 - s(\mathbf{x}; y_\alpha))}p(\mathbf{x}), \tag{A23}$$

$$\sqrt{(1 - \widetilde{s}(\mathbf{x}; \theta_k))\widetilde{s}(\mathbf{x}; \theta_k)}p(\mathbf{x}) \xrightarrow{P} \sqrt{s(\mathbf{x}; y_\alpha)(1 - s(\mathbf{x}; y_\alpha))}p(\mathbf{x}). \tag{A24}$$

Using the above convergence results, we have

$$\lim_{k \to \infty} Var[\widehat{P}_k(y_\alpha)]$$

$$= \frac{1}{n_T} \left[\frac{1}{m} \int p(\mathbf{x})\sqrt{s(\mathbf{x}; y_\alpha)}\sqrt{1 + (n_T - 1)s(\mathbf{x}; y_\alpha)}d\mathbf{x} \int p(\mathbf{x})\sqrt{s(\mathbf{x}; y_\alpha)}\sqrt{1 + (n_T - 1)s(\mathbf{x}; y_\alpha)}d\mathbf{x} \right.$$

$$\left. + \frac{m-1}{m} \int \sqrt{s(\mathbf{x}; y_\alpha)(1 - s(\mathbf{x}; y_\alpha))}p(\mathbf{x})d\mathbf{x} \int \sqrt{s(\mathbf{x}; y)(1 - s(\mathbf{x}; y_\alpha))}p(\mathbf{x})d\mathbf{x} \right] - \frac{\mathbb{P}^2(Y > y_\alpha)}{m}$$

$$= \frac{1}{n_T} \left[\frac{1}{m} \left(\int \sqrt{p(\mathbf{x})s(\mathbf{x}; y_\alpha)}\sqrt{p(\mathbf{x})[1 + (n_T - 1)s(\mathbf{x}; y_\alpha)]}d\mathbf{x} \right)^2 \right.$$

$$\left. + \frac{m-1}{m} \left(\int \sqrt{p(\mathbf{x})s(\mathbf{x}; y_\alpha)}\sqrt{p(\mathbf{x})(1 - s(\mathbf{x}; y_\alpha))}d\mathbf{x} \right)^2 \right] - \frac{\mathbb{P}^2(Y > y_\alpha)}{m}$$

$$\leq \frac{1}{n_T} \left[\frac{1}{m} \left(\int p(\mathbf{x})s(\mathbf{x}; y_\alpha)d\mathbf{x} \right) \left(\int p(\mathbf{x})(1 + (n_T - 1)s(\mathbf{x}; y_\alpha))d\mathbf{x} \right) \right.$$

$$\left. + \frac{m-1}{m} \left(\int p(\mathbf{x})s(\mathbf{x}; y_\alpha)d\mathbf{x} \right) \left(\int p(\mathbf{x})(1 - s(\mathbf{x}; y_\alpha))d\mathbf{x} \right) \right] - \frac{\mathbb{P}^2(Y > y_\alpha)}{m}$$

$$= \frac{1}{n_T}\left[\frac{\mathbb{P}(Y > y_\alpha)(1 + (n_T - 1)\mathbb{P}(Y > y_\alpha))}{m} + \frac{(m-1)\mathbb{P}(Y > y_\alpha)(1 - \mathbb{P}(Y > y_\alpha))}{m}\right] - \frac{\mathbb{P}^2(Y > y_\alpha)}{m}$$
$$= \frac{1}{n_T}\alpha(1-\alpha), \tag{A25}$$

where (A25) holds from Cauchy inequality. Therefore, we get

$$\lim_{k\to\infty} Var[\widehat{P}_k(y_\alpha)] \le \frac{\alpha(1-\alpha)}{n_T}. \tag{A26}$$

The equality in (A26) holds if and only if $\exists c \ge 0$, s.t. $p(\mathbf{x})s(\mathbf{x}; y_\alpha) = cp(\mathbf{x})(1 + (n_T - 1)s(\mathbf{x}; y_\alpha))$ and $\exists c' \ge 0$, s.t. $p(\mathbf{x})s(\mathbf{x}; y_\alpha) = c' p(\mathbf{x})(1 - s(\mathbf{x}; y_\alpha))$. These conditions hold when $s(\mathbf{x}; y_\alpha)$ is constant with respect to $\mathbf{x} \in \Omega_{\mathbf{x}}$.

(2) Proof of $\lim_{K\to\infty} Var[\sqrt{K}\widehat{P}_{1:K}(y_\alpha)] \le \frac{\alpha(1-\alpha)}{n_T}$: Let

$$a_K = \sum_{k=1}^{K} Var[\widehat{P}_k(y_\alpha)],$$
$$b_K = K.$$

Then, we have

$$\lim_{K\to\infty} Var[\sqrt{K}\widehat{P}_{1:K}(y_\alpha)] = \lim_{K\to\infty} \frac{1}{K} Var\left[\sum_{k=1}^{K} \widehat{P}_k(y_\alpha)\right]$$
$$= \lim_{K\to\infty} \frac{a_K}{b_K}$$
$$= \lim_{K\to\infty} \frac{a_{K+1} - a_K}{b_{K+1} - b_K} \tag{A27}$$

$$= \lim_{K\to\infty} Var[\widehat{P}_{K+1}(y_\alpha)]$$
$$\le \frac{\alpha(1-\alpha)}{n_T}, \tag{A28}$$

where (A27) holds due to Stolz Cesàro theorem (Muresan, 2009) and the last inequality in (A28) is from (A26). In other words, we have

$$\lim_{K\to\infty} Var[\sqrt{Kn_T}\widehat{P}_{1:K}(y_\alpha)] \le \alpha(1-\alpha)$$

A6 | Proof of Theorem 3

Let $s^a(\mathbf{x}; \theta_k)$ denote an estimation for $\widetilde{s}(\mathbf{x}; \theta_k)$ and let $\widetilde{s}^a(\mathbf{x}; \theta_k) \coloneqq \left(1 - \frac{2\delta}{k^\beta}\right) s^a(\mathbf{x}; \theta_k) + \frac{\delta}{k^\beta}$. From the proof of Theorem 2, we only need to show that (A21)–(A24) hold with $\widetilde{s}^a(\mathbf{x}; \theta_k)$ if $\|s^a(\mathbf{x}; \theta_k) - s(\mathbf{x}; \theta_k)\| = o(k^{-\beta})$ is satisfied. Let $h \coloneqq \min_{\mathbf{x}\in\Omega_{\mathbf{x}}} |s(\mathbf{x}; y)|$. The difference between $\widetilde{s}^a(\mathbf{x}; \theta_k)$ and $s(\mathbf{x}; y_\alpha)$ becomes

$$|\widetilde{s}^a(\mathbf{x}; \theta_k) - s(\mathbf{x}; y_\alpha)|$$
$$= \left|\left(1 - \frac{2\delta}{k^\beta}\right) s^a(\mathbf{x}; \theta_k) + \frac{\delta}{k^\beta} - s(\mathbf{x}; y_\alpha)\right|$$
$$= \left|\left(1 - \frac{2\delta}{k^\beta}\right)(s^a(\mathbf{x}; \theta_k) - s(\mathbf{x}; \theta_k)) + \left(1 - \frac{2\delta}{k^\beta}\right)(s(\mathbf{x}; \theta_k) - s(\mathbf{x}; y_\alpha)) + \frac{\delta}{k^\beta}(1 - 2s(\mathbf{x}; y_\alpha))\right|$$
$$\le \left(1 - \frac{2\delta}{k^\beta}\right)\|s^a(\mathbf{x}; \theta_k) - s(\mathbf{x}; \theta_k)\| + \left(1 - \frac{2\delta}{k^\beta}\right)|s(\mathbf{x}; \theta_k) - s(\mathbf{x}; y_\alpha)| + \frac{\delta}{k^\beta}(1 - 2s(\mathbf{x}; y_\alpha)).$$

When $\beta < 0.5$, from (A6), (A8) and assumptions in Theorem 3, we have $|\theta_k - y_\alpha| / k^{-\beta} \le O(k^{\beta-1+\beta}) \to 0$, as $k \to \infty$. As such, we get $|\theta_k - y_\alpha| = o(k^{-\beta})$. Because $s(\mathbf{x}; y)$ is assumed to be locally Lipschitz continuous at y_α, $|s(\mathbf{x}; \theta_k) - s(\mathbf{x}; y_\alpha)| = o(k^{-\beta})$. Consequently,

$$|\widetilde{s}^a(\mathbf{x}; \theta_k) - s(\mathbf{x}; y_\alpha)| \le \delta k^{-\beta}(1 - 2h) + o(k^{-\beta}). \tag{A29}$$

Note that $\widetilde{s}^a(\mathbf{x};\theta_k) \geq \delta k^{-\beta}$, because of $0 \leq s^a(\mathbf{x};\theta_k) \leq 1$. Therefore, using (A29), it holds

$$\left|1 - \frac{s(\mathbf{x};y_\alpha)}{\widetilde{s}^a(\mathbf{x};\theta_k)}\right| \leq 1 - 2h + o(1),$$
$$2h + o(1) \leq \frac{s(\mathbf{x};y_\alpha)}{\widetilde{s}^a(\mathbf{x};\theta_k)} \leq 2 - 2h + o(1). \tag{A30}$$

Similarly, we have

$$|(1 - \widetilde{s}^a(\mathbf{x};\theta_k)) - (1 - s(\mathbf{x};y_\alpha))| \leq \delta k^{-\beta}(1-2h) + o(k^{-\beta}),$$

and $1 - \widetilde{s}^a(\mathbf{x};\theta_k) \geq \delta k^{-\beta}$, it holds

$$\left|1 - \frac{1 - s(\mathbf{x};y_\alpha)}{1 - \widetilde{s}^a(\mathbf{x};\theta_k)}\right| \leq 1 - 2h + o(1),$$
$$2h + o(1) \leq \frac{1 - s(\mathbf{x};y_\alpha)}{1 - \widetilde{s}^a(\mathbf{x};\theta_k)} \leq 2 - 2h + o(1). \tag{A31}$$

Next, we show that (A21) to (A24) hold with $\widetilde{s}^a(\mathbf{x};\theta_k)$. First,

$$\begin{aligned}
&|p(\mathbf{x})\sqrt{\widetilde{s}^a(\mathbf{x};\theta_k)}\sqrt{1 + (n_T-1)\widetilde{s}^a(\mathbf{x};\theta_k)} - p(\mathbf{x})\sqrt{s(\mathbf{x};y_\alpha)}\sqrt{1 + (n_T-1)s(\mathbf{x};y_\alpha)}| \\
&= p(\mathbf{x})\frac{[\widetilde{s}^a(\mathbf{x};\theta_k) - s(\mathbf{x};y_\alpha)|(1 + (n_T-1)(\widetilde{s}^a(\mathbf{x};\theta_k) + s(\mathbf{x};y_\alpha)))}{\sqrt{\widetilde{s}^a(\mathbf{x};\theta_k)}\sqrt{1 + (n_T-1)\widetilde{s}^a(\mathbf{x};\theta_k)} + \sqrt{s(\mathbf{x};y_\alpha)}\sqrt{1 + (n_T-1)s(\mathbf{x};y_\alpha)}} \\
&\leq p_{max}(\delta k^{-\beta}(1-2h) + o(k^{-\beta}))\frac{(1 + 2(n_T-1))k^{\beta/2}}{\sqrt{\delta}} \\
&\leq o(1),
\end{aligned} \tag{A32}$$

where (A32) holds by plugging the result in (A29), $\widetilde{s}^a(\mathbf{x};\theta_k) \leq 1$ and $s(\mathbf{x};y_\alpha) \leq 1$ in the numerator. In the denominator, we use $\widetilde{s}^a(\mathbf{x};\theta_k) \geq \delta k^{-\beta}$ and $(n_T - 1)\widetilde{s}^a(\mathbf{x};\theta_k) \geq 0$ in the first term and $\sqrt{s(\mathbf{x};y_\alpha)}\sqrt{1 + (n_T-1)s(\mathbf{x};y_\alpha)} > 0$ in the second term. This result implies that

$$p(\mathbf{x})\sqrt{\widetilde{s}^a(\mathbf{x};\theta_k)}\sqrt{1 + (n_T-1)\widetilde{s}^a(\mathbf{x};\theta_k)} \xrightarrow{P} p(\mathbf{x})\sqrt{s(\mathbf{x};y_\alpha)}\sqrt{1 + (n_T-1)s(\mathbf{x};y_\alpha)}. \tag{A33}$$

Next, corresponding to (A22), we have

$$\begin{aligned}
&\left|\frac{p(\mathbf{x})s(\mathbf{x};y_\alpha)[1 + (n_T-1)s(\mathbf{x};y_\alpha)]}{\sqrt{\widetilde{s}^a(\mathbf{x};\theta_k)}\sqrt{1 + (n_T-1)\widetilde{s}^a(\mathbf{x};\theta_k)}} - p(\mathbf{x})\sqrt{s(\mathbf{x};y_\alpha)}\sqrt{1 + (n_T-1)s(\mathbf{x};y_\alpha)}\right| \\
&= p(\mathbf{x})\sqrt{s(\mathbf{x};y_\alpha)}[1 + (n_T-1)s(\mathbf{x};y_\alpha)] \\
&\quad \cdot \left|\frac{\sqrt{\widetilde{s}^a(\mathbf{x};\theta_k)}\sqrt{1 + (n_T-1)\widetilde{s}^a(\mathbf{x};\theta_k)} - \sqrt{s(\mathbf{x};y_\alpha)}\sqrt{1 + (n_T-1)s(\mathbf{x};y_\alpha)}}{\sqrt{\widetilde{s}^a(\mathbf{x};\theta_k)}\sqrt{1 + (n_T-1)\widetilde{s}^a(\mathbf{x};\theta_k)}\sqrt{1 + (n_T-1)s(\mathbf{x};y_\alpha)}}\right| \\
&\leq p_{max}(\delta k^{-\beta}(1-2h) + o(k^{-\beta}))\frac{(1 + 2(n_T-1))k^{\beta/2}}{\sqrt{\delta}}\frac{\sqrt{s(\mathbf{x};y_\alpha)}}{\sqrt{\widetilde{s}^a(\mathbf{x};\theta_k)}}\frac{\sqrt{1 + (n_T-1)s(\mathbf{x};y_\alpha)}}{\sqrt{1 + (n_T-1)\widetilde{s}^a(\mathbf{x};\theta_k)}} \qquad \text{(A34)} \\
&\leq o(1), \qquad \text{(A35)}
\end{aligned}$$

where (A34) holds by using (A32) and (A35) holds because of (A30) and the fact that the last factor is bounded by a constant. As such, we have

$$\frac{p(\mathbf{x})s(\mathbf{x};y_\alpha)[1 + (n_T-1)s(\mathbf{x};y_\alpha)]}{\sqrt{\widetilde{s}^a(\mathbf{x};\theta_k)}\sqrt{1 + (n_T-1)\widetilde{s}^a(\mathbf{x};\theta_k)}} \xrightarrow{P} p(\mathbf{x})\sqrt{s(\mathbf{x};y_\alpha)}\sqrt{1 + (n_T-1)s(\mathbf{x};y_\alpha)}. \tag{A36}$$

Similarly, when $k \geq 2^{1/\beta}$, it holds

$$\begin{aligned}
&|\sqrt{(1 - \widetilde{s}^a(\mathbf{x};\theta_k))\widetilde{s}^a(\mathbf{x};\theta_k)}p(\mathbf{x}) - \sqrt{s(\mathbf{x};y_\alpha)(1 - s(\mathbf{x};y_\alpha))}p(\mathbf{x})| \\
&= p(\mathbf{x})\frac{[\widetilde{s}^a(\mathbf{x};\theta_k) - s(\mathbf{x};y_\alpha)|(1 + \widetilde{s}^a(\mathbf{x};\theta_k) + s(\mathbf{x};y_\alpha))}{\sqrt{(1 - \widetilde{s}^a(\mathbf{x};\theta_k))\widetilde{s}^a(\mathbf{x};\theta_k)} + \sqrt{s(\mathbf{x};y_\alpha)(1 - s(\mathbf{x};y_\alpha))}} \\
&\leq \frac{3p_{max}(\delta k^{-\beta}(1-2h) + o(k^{-\beta}))}{\sqrt{(1 - \delta k^{-\beta})\delta k^{-\beta}}} \qquad \text{(A37)} \\
&\leq \frac{3p_{max}(\delta k^{-\beta}(1-2h) + o(k^{-\beta}))}{\sqrt{\delta k^{-\beta} - 1/2\delta k^{-\beta}}} \qquad \text{(A38)} \\
&\leq o(1), \qquad \text{(A39)}
\end{aligned}$$

where (A37) holds by using the result in (A29), $\widetilde{s}^a(\mathbf{x};\theta_k) \le 1$ and $s(\mathbf{x};y_\alpha) \le 1$ in the numerator. In the denominator, the first term reaches its minimum when $\widetilde{s}^a(\mathbf{x};\theta_k) = \delta k^{-\beta}$ or $1-\delta k^{-\beta}$ and the second positive term can be dropped. (A38) holds because of $k^\beta \ge 2$ for $k \ge 2^{1/\beta}$ and thus, $-\delta k^{-2\beta} \ge -1/2\delta k^{-\beta}$. Thus, we have

$$\sqrt{(1-\widetilde{s}^a(\mathbf{x};\theta_k))\widetilde{s}^a(\mathbf{x};\theta_k)}p(\mathbf{x}) \xrightarrow{P} \sqrt{s(\mathbf{x};y_\alpha)(1-s(\mathbf{x};y_\alpha))}p(\mathbf{x}) \tag{A40}$$

Lastly,

$$\begin{aligned}
&\left| \frac{s(\mathbf{x};y_\alpha)(1-s(\mathbf{x};y_\alpha))p(\mathbf{x})}{\sqrt{\widetilde{s}^a(\mathbf{x};\theta_k)(1-\widetilde{s}^a(\mathbf{x};\theta_k))}} - \sqrt{s(\mathbf{x};y_\alpha)(1-s(\mathbf{x};y_\alpha))}p(\mathbf{x}) \right| \\
&= p(\mathbf{x})\sqrt{s(\mathbf{x};y_\alpha)}(1-s(\mathbf{x};y_\alpha)) \frac{|\sqrt{s(\mathbf{x};y_\alpha)(1-s(\mathbf{x};y_\alpha))} - \sqrt{\widetilde{s}^a(\mathbf{x};\theta_k)(1-\widetilde{s}^a(\mathbf{x};\theta_k))}|}{\sqrt{\widetilde{s}^a(\mathbf{x};\theta_k)(1-\widetilde{s}^a(\mathbf{x};\theta_k))}\sqrt{(1-s(\mathbf{x};y_\alpha)}} \\
&= p(\mathbf{x})|\sqrt{s(\mathbf{x};y_\alpha)(1-s(\mathbf{x};y_\alpha))} - \sqrt{\widetilde{s}^a(\mathbf{x};\theta_k)(1-\widetilde{s}^a(\mathbf{x};\theta_k))}| \frac{\sqrt{s(\mathbf{x};y_\alpha)}}{\sqrt{\widetilde{s}^a(\mathbf{x};\theta_k)}} \frac{\sqrt{1-s(\mathbf{x};y_\alpha)}}{\sqrt{1-\widetilde{s}^a(\mathbf{x};\theta_k)}} \\
&\le o(1),
\end{aligned} \tag{A41}$$

(A41) holds by plugging in (A39), (A30) and (A31). Therefore,

$$\frac{s(\mathbf{x};y_\alpha)(1-s(\mathbf{x};y_\alpha))p(\mathbf{x})}{\sqrt{\widetilde{s}^a(\mathbf{x};\theta_k)(1-\widetilde{s}^a(\mathbf{x};\theta_k))}} \xrightarrow{P} \sqrt{s(\mathbf{x};y_\alpha)(1-s(\mathbf{x};y_\alpha))}p(\mathbf{x}). \tag{A42}$$

Note that the convergence results in (A33), (A36), (A40) and (A42) correspond to the results in (A21)–(A24). Then the variance reduction property follows by using the similar procedure in the proof of Theorem 2.

A7 | Discussion on and derivation of the variance of CMC2's POE estimator

Recall the CMC2's POE estimator as

$$\widehat{P}_{CMC2}(y_\alpha) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{I}(Y_{ij} > y_\alpha)\right).$$

We obtain the optimal n_i which minimizes the variance of the CMC2's POE estimator as follows:

$$\begin{aligned}
Var[\widehat{P}_{CMC2}(y_\alpha)] &= Var\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{I}(Y_{ij} > y_\alpha)\right)\right] \\
&= \frac{1}{m^2}E\left[Var\left[\frac{1}{n_i}\sum_{i=1}^{m}\mathbb{I}(Y_{ij} > y_\alpha)|\mathbf{X}_1, \dots, \mathbf{X}_m\right]\right] + \frac{1}{m^2}Var\left[E\left[\frac{1}{n_i}\sum_{i=1}^{m}\mathbb{I}(Y_{ij} > y_\alpha)|\mathbf{X}_1, \dots, \mathbf{X}_m\right]\right] \\
&= \frac{1}{m^2}E\left[\sum_{i=1}^{m}\frac{1}{n_i}s(\mathbf{X}_i, y_\alpha)(1-s(\mathbf{X}_i, y_\alpha))\right] + \frac{1}{m}Var[s(\mathbf{X};y_\alpha)],
\end{aligned} \tag{A43}$$

where the second term in the last equation is obtained from the fact that $\mathbf{X}_i$'s are iid.

In (A43), the second term does not include n_i. To find n_i that minimizes $Var[\widehat{P}_{CMC2}(y_\alpha)]$, we minimize the first term. We let the allocation size N_i at $\mathbf{X}_i$ as a function of $\mathbf{X}_i$:

$$n_i = n_T \cdot \frac{c(\mathbf{X}_i)}{\sum_{j=1}^{M} c(\mathbf{X}_j)}, \quad i = 1, 2, \dots, m,$$

where $c(\mathbf{X})$ is a nonnegative function. Then, following the the procedure in Choe et al. (2015) (see the proof of Lemma 1 therein), the optimal n_i is given by

$$n_i = n_T \cdot \frac{\sqrt{s(\mathbf{X}_i)(1-s(\mathbf{X}_i))}}{\sum_{j=1}^{M}\sqrt{s(\mathbf{X}_j)(1-s(\mathbf{X}_j))}} \quad \text{for } i = 1, 2, \dots, m. \tag{A44}$$

There are several issues concerning the optimal form of n_i in (A44). First, it needs the information of conditional POE $s(\mathbf{X}_i)$. The CMC procedure, by definition, uses the input density $p(\mathbf{x})$ only, ignoring the geometric structure of response surface. So,

if we use the optimal n_i in (A44), this procedure is not essentially CMC. Second, let us compare CMC2 with the original SIS procedure that uses the following POE:

$$\widehat{P}_{SIS}(y) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{I}(y_{ij} > y)\right)\frac{p(\mathbf{X}_i)}{q(\mathbf{X}_i;\theta)}.$$

The SIS procedure optimizes $q(\mathbf{x}_i;\theta)$ and n_i, $i = 1, \ldots, m$, together. On the contrary, the aforementioned CMC2 optimizes n_i only, while fixing the input sampling density at $p(\mathbf{x})$. Therefore, CMC2 (even though n_i is optimized, assuming $s(\mathbf{X})$ is known) is suboptimal, compared to SIS. In fact, CMC2 is just a special case with $q(\mathbf{x};\theta) = p(\mathbf{x})$.

Next, let us consider the equal sample size allocation. Given the total computational resource Kn_T, we can set $n_i = (Kn_T)/m$. Then we obtain

$$\begin{aligned}
&Var[\widehat{P}_{CMC2}(y_\alpha)]\\
&= \frac{1}{m^2}\cdot m\cdot\frac{m}{Kn_T}\int s(\mathbf{x};y_\alpha)(1-s(\mathbf{x};y_\alpha))p(\mathbf{x})d\mathbf{x} + \frac{1}{m}\int [s(\mathbf{x};y_\alpha)-\alpha]^2 p(\mathbf{x})d\mathbf{x}\\
&= \frac{1}{Kn_T}\alpha + \left(\frac{1}{m}-\frac{1}{Kn_T}\right)\int s(\mathbf{x};y_\alpha)^2 p(\mathbf{x})d\mathbf{x} - \frac{1}{m}\alpha^2\\
&= \frac{1}{Kn_T}\alpha(1-\alpha) + \left(\frac{1}{m}-\frac{1}{Kn_T}\right)\left(\int s(\mathbf{x};y_\alpha)^2 p(\mathbf{x})d\mathbf{x} - \alpha^2\right)\\
&= \frac{1}{Kn_T}\alpha(1-\alpha) + \left(\frac{1}{m}-\frac{1}{Kn_T}\right)(E[s^2(\mathbf{x};y_\alpha)] - E[s(\mathbf{x};y_\alpha)]^2)\\
&\geq \frac{1}{Kn_T}\alpha(1-\alpha)
\end{aligned}$$

Noting that the right-hand side is the variance of the original CMC that runs simulation once at each $\mathbf{X}_i$, we can see that allowing multiple replicates is not beneficial in the CMC procedure.