IISE Transactions

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uiie21

Optimal budget allocation for stochastic simulation with importance sampling: Exploration vs. replication

Young Myoung Ko & Eunshin Byon

To cite this article: Young Myoung Ko & Eunshin Byon (2022) Optimal budget allocation for stochastic simulation with importance sampling: Exploration vs. replication, IISE Transactions, 54:9, 881-893, DOI: 10.1080/24725854.2021.1953197

To link to this article: https://doi.org/10.1080/24725854.2021.1953197

IISE TRANSACTIONS
2022, VOL. 54, NO. 9, 881–893
https://doi.org/10.1080/24725854.2021.1953197

Optimal budget allocation for stochastic simulation with importance sampling: Exploration vs. replication

Young Myoung Ko[a] and Eunshin Byon[b]

[a]Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, Gyeongbuk, South Korea; [b]Department of Industrial & Operations Engineering, University of Michigan, Ann Arbor, MI, USA

ABSTRACT

This article investigates a budget allocation problem for optimally running stochastic simulation models with importance sampling in computer experiments. In particular, we consider a two-level (or nested) simulation to estimate the expectation of the simulation output, where the first-level draws random input samples and the second-level obtains the output given the input from the first-level. The two-level simulation faces the trade-off in allocating the computational budgets: exploring more inputs (exploration) or exploiting the stochastic response surface at a sampled point in more detail (replication). We study an appropriate computational budget allocation strategy that strikes a balance between exploration and replication to minimize the variance of the estimator when importance sampling is employed at the first-level simulation. Our analysis suggests that exploration can be beneficial than replication in many practical situations. We also conduct numerical experiments in a wide range of settings and wind turbine case study to investigate the trade-off.

ARTICLE HISTORY
Received 10 August 2020
Accepted 29 June 2021

KEYWORDS
Computer experiment; Monte Carlo sampling; reliability; variance reduction

1. Introduction

This article concerns a simulation budget allocation problem when estimating an expectation of a random quantity that is a function of random inputs and some unknown random effects. The unknown random effects make the function generate random outputs, given the realization of random inputs. Choe *et al.* (2015) call such simulation models (or computer models) the *stochastic* simulation models, in contrast with the *deterministic* simulation models where the randomness of the function only comes from input variables.

Simulation with stochastic computer models basically takes a two-level procedure; the first-level (referred to as outer simulation) collects input samples from their distribution, and the second-level (referred to as inner simulation) conducts simulation runs, given inputs from the first-level. A case in point of the two-level simulation is the wind turbine simulation. The International Electrotechnical Commission (IEC)'s design standard requires to assess the turbine reliability using stochastic simulations at the design stage (International Electrotechnical Commission, 2005). In response, the U.S. Department of Energy's National Renewable Energy Laboratory (NREL) developed aeroelastic simulators to assist wind turbine manufactures to design a reliable wind turbine operating under various wind conditions (Jonkman and Buhl Jr., 2005; Jonkman, 2009). In the reliability problem of wind turbines, the input wind condition is sampled at the first-level and then aeroelastic simulators generate stochastic load responses at the sampled wind condition at the second-level.

The stochastic simulation model is also called *nested simulation*, and it has been used to obtain financial portfolio risk measurements such as Value-at-Risk (VaR) and expected shortfall in the literature (Gordy and Juneja, 2010; Lan *et al.*, 2010; Broadie *et al.*, 2011). VaR is the quantile estimation of risk factors given the probability of loss. Its expected shortfall estimates the tail expectation that quantifies the actual loss amount when the large loss happens. In this way it complements a VaR that ignores the loss distribution beyond the quantile (Gordy and Juneja, 2010). The risk factors are drawn in the outer step and the loss is evaluated using the inner step simulation.

When we have a limited budget on the simulation runs, we need to optimize the allocation of the budget for both levels to accurately estimate the output of interest. Choe *et al.* (2015) provided a general framework for resource allocation at both levels when the first-level uses importance sampling. The objective of importance sampling is to take more samples from the *important* input region to reduce estimation variance with limited budgets. Choe *et al.* (2015) considered the so-called *stochastic black box model*, where the second-level simulation purely relies on a complicated black box computer model, such as a wind turbine simulator. In the two-level simulation framework, they jointly derived the importance sampling density for the first-level simulation and the optimal budget allocation for the second-level simulation; the importance sampling density affects the optimal budget allocation and *vice versa*. Their approach is called stochastic importance sampling and has

CONTACT Eunshin Byon ebyon@umich.edu
Supplemental data for this article can be accessed online at https://doi.org/10.1080/24725854.2021.1953197

been extensively studied in wind energy applications (Choe *et al.*, 2015; Choe *et al.*, 2016; Choe *et al.*, 2018; Cao and Choe, 2019; Pan *et al.*, 2020; Pan *et al.*, 2021).

Specifically, given the sample size M of the input random variable at the first-level simulation and the total number N_T of simulations in the second-level, Choe *et al.* (2015) derived the optimal importance sampling density to draw the input X_i, $i \in \{1, \ldots, M\}$, and the optimal budget allocation N_i in the second-level simulation for each input X_i. However, the trade-off in deciding the input sample size M has not yet been studied. Large M would provide better *exploration* of the response surface over the important input region, whereas a small M would allow better quantification of the variability of the stochastic response, that is, it provides better *exploitation* at the sampled inputs by *replication*. To be clear, in this article, exploration implies drawing more inputs with large M, whereas replication assigns more budgets to replicate the stochastic response at a smaller number of selected inputs.

In order to best utilize limited computational resources and accurately estimate the output of interest, it is needed to provide a guideline on how many input samples need to be drawn and how many replications are needed for each sampled input. We investigate the trade-off between the exploration and replication and derive the theoretically optimal input sample size M, given the limited budget N_T. Our analysis shows that setting the input sample size to be the same as the total budget, that is, $M = N_T$, is optimal when the positive integer restriction is not imposed on the budget values (N_i values) in the second-level simulations. This result implies that the exploration is better than replication. However, as the theoretical results do not account for the integer requirement, the budget value should be rounded to its nearest positive integer in practice. Therefore, with $M = N_T$, N_i should be one in practical implementation. We theoretically prove that the variance with the theoretical optimal allocation (that takes real-valued N_i values without rounding) is smaller than the variance with the allocation after rounding (i.e., $N_i = 1$). Our implementation results also suggest that rounding to $N_i = 1$ could lead to a non-negligible increase in the variance.

Therefore, with the integer condition on the budgets, the optimal input sample size M should lie between one and N_T, balancing exploration and replication. However, optimal M depends on the problem structure and is hard to obtain analytically or empirically. Having the fact that $M = N_T$ provides the theoretically optimal allocation, we consider another estimator that is designed to purely explore the input area without replicating the stochastic response. We refer to this estimator as the exploration-only estimator. This estimator allows only one simulation run in the second-level at each input drawn at the first-level. We prove that the theoretical variance of the exploration-only estimator is smaller than that of the original optimal estimator with rounding. We also empirically show that the exploration-only estimator provides consistently good performance in numerical examples and wind turbine case study.

Overall the contribution of this article can be summarized as follows:

- We prove that under the limited budget N_T, more first-level simulation runs (i.e., more exploration) theoretically reduces the variance of the estimator when imposing no integer constraints on the second-level budget values.
- We show that rounding the budget values loses optimality. That is, the resulting allocation is not optimal anymore when the practical implementation requires us to round the real-valued *theoretically* optimal allocation to the nearest natural numbers.
- We analytically prove that the exploration-only estimator has a smaller variance than the implementable version of the theoretically optimal estimator with rounding.
- Based on the theoretical analysis and empirical results, we show that the full exploration strategy provides a robust solution for the two-level simulation combined with importance sampling at the first-level.

The organization of this article is as follows: Section 2 reviews relevant studies in the two-level simulation and statistical literature. Section 3 describes the problem of interest. Section 4 studies the optimal resource allocation and its practical issues. Section 5 confirms the theoretical results using numerical examples and the wind turbine case study. Section 6 makes concluding remarks and discusses future work.

2. Literature review

Estimation and inference of systems using stochastic computer models have gained popularity recently. For computer models whose run-time is not negligible, several studies investigate the resource allocation problems in different contexts to understand systems better with limited computational budgets. In the statistical literature, adaptive sampling strategies for building accurate surrogate models that emulate stochastic computer models have been actively studied. The goal of these metamodeling studies is to obtain high quality metamodels by investigating the trade-off between exploration and replication. Sinha and Wiens (2002) develop a sequential design scheme for a nonlinear parametric regression model as a surrogate, when the fitted model is possibly incorrect. Recently, several studies develop a Gaussian Process (GP) as a surrogate model of a computationally expensive computer model (Wang *et al.*, 2020) and provide adaptive sampling approaches that sequentially determine design points in order to build an accurate GP emulator. Based on a Bayesian tree-based GP, Gramacy and Lee (2009) combine the classic design of experiments method with the active learning approach and propose a new adaptive sampling design strategy in supercomputer experiments. Binois *et al.* (2019) further generalize the approach and show that replication can be more beneficial, especially for heteroscedastic systems. They sequentially find a design point that minimizes the predictive uncertainty measured by the Integrated Mean Squared Error (IMSE).

Ankenman *et al.* (2010) extend the deterministic kriging method to the stochastic kriging method. At each design point, they derive the optimal number of replications for minimizing IMSE, which is proportional to the standard deviation of the intrinsic variance (and the square root of a function of the extrinsic covariance). Wang and Haaland (2019) also demonstrate how replication could help signal isolation in stochastic kriging. Xiong *et al.* (2013) present a sequential design scheme when both high-accuracy and low-accuracy computer models are available. In Goetz *et al.* (2018), active sampling schemes are presented to build a non-parametric tree-based metamodel. The primary objective of these surrogate studies is to build a globally accurate emulator over the entire input space. Although some of these studies investigate the trade-off between exploration and replication, their focus is to estimate the predictive distribution $Y|X$, where Y denotes the simulation output and X is a design point (not random variable).

In financial risk analysis, given the portfolio (the first-level), computational budget allocation (the second-level allocation) to each scenario is the focus of several studies (Gordy and Juneja, 2010; Broadie *et al.*, 2011). In these studies, the first-level simulation usually assumes a predetermined distribution (portfolio) and mostly the second-level simulation decides the optimal budget allocation. Broadie *et al.* (2011) propose a sequential approach that allocates more simulation budget to the inner simulation of the outer scenarios located close to the boundary of the tail probability, i.e., close to c for the estimator of $P(L > c)$, using the optimization problem that maximizes the probability of a sign change. For the resource allocation at both levels, Gordy and Juneja (2010) formulate an optimization problem that determines the first and second-level budgets to minimize the Mean Squared Error (MSE) of the estimator of risk measurements. Although they investigate the number of total outer and inner simulation numbers, their analysis focuses on the second-level budget allocation. They also do not consider the sampling distribution of input variables, such as importance sampling.

The trade-off between exploration and the replication problem has also been studied in the context of data-driven optimization. Following the increased popularity of GPs, Bayesian optimization has gained attention in the literature as one of the black box optimization techniques, typically when the objective function is continuous (Mockus, 1989; Snoek *et al.*, 2012). Bayesian optimization consists of two major components: a GP for modeling an objective function over a solution (or design) space and an acquisition function to choose the next design point (Frazier, 2018). It updates the posterior probability on the objective function using all available data and chooses the next sample point that maximizes the acquisition function. Acquisition functions, including the well-known expected improvement, are designed to explore new design points with high uncertainty while exploiting the estimated objective value.

Similarly, adaptive learning has been actively studied in multi-armed bandits in order to solve discrete sequential optimization problems (Gittins and Jones, 1979). One of the popular algorithms is Thomson sampling (Thompson, 1933; Chapelle and Li, 2011). Similar to Bayesian optimization, Thomson sampling updates the posterior and chooses the next action using the posterior. Multi-arm bandits have been applied in a wide range of online decision problems, such as revenue management, Internet advertising, recommendation systems, and hyperparameter tuning (Russo *et al.*, 2018). These optimization studies mainly focus on finding the best solution that optimizes the objective function, which is different from the problem context considered in this study.

In the aforementioned studies, inputs at the first-level are considered as design points or decision variables (not random variables), so the first-level budget allocation is not taken into consideration in general. One of the popular methods for the optimal budget allocation at the first-level is importance sampling. Most importance sampling studies in the literature consider deterministic computer models, so it only aims to optimize the first-level simulation. For example, Glynn and Iglehart (1989) study importance sampling in the simulation of stochastic processes. Glasserman *et al.* (2000) apply importance sampling to estimate the VaR in financial risk analysis.

In summary, existing studies, by and large, focus on the resource allocation at either first- or second-level simulation. This article investigates the optimal resource allocation at both levels in the importance sampling framework.

3. Problem description

Problems involving nested simulation estimate the expectation of a random variable. For example, estimating a tail probability – one of the popular topics in reliability analysis – can be regarded as estimating the expectation of an indicator function. In this study, we state the problem in a general form. Let X and Y denote random variables for the simulation input and output, respectively. Suppose we want to estimate the expectation of a random variable Z, i.e., $E[Z]$, where Z is a function of Y. For the estimation of a tail probability, we can set $Z = \mathbb{I}(Y > l)$, so that $E[Z]$ becomes the tail probability $P(Y > l)$. Then we can estimate $E[Z]$ using the law of total expectation as

$$E[Z] = E[E[Z|X]]. \tag{3.1}$$

To estimate $E[Z]$ in the two-level simulation framework, the first-level is to sample the input data X and given the sampled X, the second-level is to conduct the stochastic simulation and get Y (or Z). Broadly speaking, there are two major approaches to estimate $E[Z]$: sampling-based estimation and statistical surrogate-based estimation. In this study we take the former approach. Let $s(x)$ denote the conditional expectation, i.e., $s(x) = E[Z|X = x]$. Estimating $s(x)$ is important to determine the quality of the estimator of $E[Z]$ in (3.1). Thus, we would like to estimate $s(x)$ accurately in an important input region. This is the fundamental idea of importance sampling or variance reduction in a broader sense.

Let $\hat{Z}$ denote an estimator of $E[Z]$ and $\hat{s}(x)$ be an unbiased estimator of $s(x)$. Note that in the surrogate-based approach, biased estimators can be considered for $\hat{s}(x)$ and a good estimator is chosen with measures such as MSE or IMSE (Gordy and Juneja, 2010; Lan *et al.*, 2010; Broadie

et al., 2011; Binois *et al.*, 2019). However, when the estimation is made with the sampling-based procedure, unbiased estimators are typically employed. In particular, in the importance sampling literature, most studies limit their analysis to unbiased estimators (Glynn and Iglehart, 1989; Glasserman *et al.*, 1999). For example, to obtain $\hat{s}(\cdot)$, we can use the sample average of multiple replicates.

Importance sampling draws an input random sample of size M, i.e., $X_1, \ldots, X_M$, from a biased density q, instead of drawing inputs from their original distribution F (with density f). At each sampled X_i, we run simulator N_i times. With the limited total simulation budget of N_T, we define an estimator $\hat{Z}$ as follows. For given $M > 0$ and $N_T > 0$,

$$\begin{aligned} \hat{Z} &\equiv \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}\sum_{j=1}^{N_i}\hat{s}_j(X_i)\frac{f(X_i)}{q(X_i)}, \\ N_T &= \sum_{i=1}^{M} N_i, \end{aligned} \tag{3.2}$$

where $\hat{s}_j(\cdot)$ is the jth replication of $\hat{s}(\cdot)$, $q(\cdot)$ is an importance sampling density, N_T is the total simulation budget, and N_i is the allocated second-level simulation budget for X_i. We assume that $q(x) = 0$ implies $\hat{s}(x)f(x) = 0$ for all x so that $\hat{Z}$ becomes an unbiased estimator of $E[Z]$. The proof of the unbiasedness of $\hat{Z}$ is available in the online supplement.

In many applications, the *cost of the* first-level simulation is cheap or negligible, whereas the second-level simulation cost is expensive (Sun *et al.*, 2011; Choe *et al.*, 2015). To put this in our problem context, drawing X_i from $q(\cdot)$ (the first-level simulation) is negligible, but running the black box computer model to obtain $\hat{s}_j(\cdot)$ (the second-level simulation) is computationally intensive. The simulation budget, therefore, applies to the second-level simulation; the total budget N_T is the sum of N_i for $i \in \{1, \ldots, M\}$.

When designing an estimator with a budget constraint, the performance of an estimator is measured by minimizing the MSE (Gordy and Juneja, 2010; Lan *et al.*, 2010; Broadie *et al.*, 2011) or variance (Glasserman *et al.*, 2000; Choe *et al.*, 2015; Choe *et al.*, 2016; Pan *et al.*, 2020; Pan *et al.*, 2021). In this study, we assume the unbiasedness of $\hat{s}(x)$. Then, minimizing variance becomes the same as minimizing the MSE. Given the limited budget N_T, we study the optimal balance between exploration and replication. With larger M, we sample more inputs, allowing more exploration. On the other hand, smaller M, which leads to larger N_i values, puts more efforts for exploitation. We derive the theoretically optimal sample size M and budget allocation N_i for $i \in \{1, \ldots, M\}$ that can strike a balance to minimize the variance.

4. Optimal budget allocation

In Section 4.1 we revisit and generalize the method in Choe *et al.* (2015) for the optimal allocation when M and N_T are both given. Then, the optimal M, given N_T, are derived in Section 4.2. Section 4.3 explains the rounding issue of the budget allocation N_i for $i \in \{1, \ldots, M\}$ and investigates its effects on the optimality.

4.1. Theoretically optimal budget allocation given M

This section derives the optimal budget allocation for N_i for $i \in \{1, \ldots, M\}$, when the input sample size M is given. Building on the results in this section, we derive the optimal M and N_i in Section 4.2. We first review the results in Choe *et al.* (2015), where the two-level simulation is used for reliability analysis. Then we generalize the results to estimate an expectation of a random quantity in (3.2).

Choe *et al.* (2015) derived the *theoretically* optimal importance sampling density and the budget allocation for the estimation of the tail probability when M is given. They considered a tail probability estimator, called $\hat{P}_{SIS1}$, as follows:

$$\begin{aligned} \hat{P}_{SIS1} &= \frac{1}{M}\sum_{i=1}^{M}\hat{P}(Y > l|X_i)\frac{f(X_i)}{q(X_i)} \\ &= \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{I}\left(Y_j^{(i)} > l\right)\frac{f(X_i)}{q(X_i)}, \end{aligned} \tag{4.1}$$

where $\mathbb{I}(\cdot)$ denotes an indicator function and $Y_j^{(i)}$ is the jth replication of the simulation output, given X_i.

Given the importance sampling density $q(\cdot)$, Lemma 4.1 derives the optimal budget allocation N_i in terms of $q(\cdot)$ that minimizes $Var[\hat{P}_{SIS1}]$ in (4.1):

Lemma 4.1 (Choe *et al.* (2015)). *For a given $q(\cdot)$ in (4.1), the optimal budget allocation N_i for $i \in \{1, \ldots, M\}$ for minimizing $Var[\hat{P}_{SIS1}]$ is given by*

$$N_i = N_T \cdot \frac{\sqrt{s(x_i)(1-s(x_i))}f(x_i)/q(x_i)}{\sum_{j=1}^{M}\sqrt{s(x_j)(1-s(x_j))}f(x_j)/q(x_j)},$$

where $s(x)$ denote the conditional tail probability, that is, $s(x) = P(Y > l|X = x)$.

Using the result of Lemma 4.1, Theorem 4.2 jointly optimizes the importance sampling density and the budget allocation for minimizing $Var[\hat{P}_{SIS1}]$:

Theorem 4.2 (Choe *et al.* (2015)). *Given the estimator $\hat{P}_{SIS1}$, the optimal importance sampling density $q_{SIS1}(\cdot)$ and budget allocation N_i for $i \in \{1, \ldots, M\}$ for minimizing $Var[\hat{P}_{SIS1}]$ are*

$$q_{SIS1}(x) = \frac{1}{C_{q1}}f(x)\sqrt{\frac{1}{N_T}s(x)(1-s(x)) + s(x)^2},$$

$$N_i = N_T\frac{\sqrt{\frac{N_T(1-s(x_i))}{1+(N_T-1)s(x_i)}}}{\sum_{j=1}^{M}\sqrt{\frac{N_T(1-s(x_j))}{1+(N_T-1)s(x_j)}}}, \text{ for } i \in \{1, \ldots, M\},$$

where C_{q1} is a normalizing constant and $s(x) = P(Y > l|X = x)$.

We can generalize the result of Choe *et al.* (2015) to the estimation of the expectation of a random variable $\hat{Z}$ in (3.2). Lemma 4.3 and Theorem 4.4 provide the extension of Lemma 4.1 and Theorem 4.2, respectively. We omit the proofs of Lemma 4.3 and Theorem 4.4, because they can be easily obtained by extending the proofs of Lemma 4.1 and Theorem 4.2.

Lemma 4.3. *For the estimator $\hat{Z}$ of $E[Z]$ in (3.2), given an importance sampling density $q(\cdot)$, the optimal budget allocation N_i for $i \in \{1, ..., M\}$ for minimizing $Var[\hat{Z}]$ is*

$$N_i = N_T \cdot \frac{\sqrt{Var[\hat{s}(x_i)]}f(x_i)/q(x_i)}{\sum_{j=1}^{M}\sqrt{Var\left[\hat{s}(x_j)\right]}f(x_j)/q(x_j)}. \tag{4.2}$$

Theorem 4.4. *Provided the estimator $\hat{Z}$ of $E[Z]$ in (3.2), the optimal importance sampling density $q^*(\cdot)$ and the budget allocation N_i for $i \in \{1, ..., M\}$ for minimizing $Var[\hat{Z}]$ are given by*

$$q^*(x) = \frac{1}{C_{q^*}} f(x)\sqrt{\frac{1}{N_T}Var[\hat{s}(x)] + E[\hat{s}(x)]^2}, \tag{4.3}$$

$$\begin{aligned} N_i &= N_T \cdot \frac{\sqrt{Var[\hat{s}(x_i)]}f(x_i)/q^*(x_i)}{\sum_{j=1}^{M}\sqrt{Var\left[\hat{s}(x_j)\right]}f(x_j)/q^*(x_j)} \\ &= N_T \frac{\sqrt{\frac{N_T Var[\hat{s}(x_i)]}{Var[\hat{s}(x_i)]+N_T E[\hat{s}(x_i)]^2}}}{\sum_{j=1}^{M}\sqrt{\frac{N_T Var\left[\hat{s}(x_j)\right]}{Var\left[\hat{s}(x_j)\right]+N_T E\left[\hat{s}(x_j)\right]^2}}}, \end{aligned} \tag{4.4}$$

where C_{q^*} is a normalizing constant.

It should be noted that in this importance sampling scheme, inputs with larger $E[\hat{s}(x)]$ and $Var[\hat{s}(x)]$ are sampled from $q^*(x)$ in (4.3). Furthermore, more budgets are allotted to the inputs with larger variance in (4.4), if $E[\hat{s}(x)]$ is the same. Now we have the optimal importance sampling density $q^*(\cdot)$ and the corresponding budget allocation N_i for $i \in \{1, ..., M\}$ when the sample size M is given. We, however, still need to find the optimal M for minimizing $Var[\hat{Z}]$. Choe *et al.* (2015) provided numerical results for choosing M using different M/N_T ratios and discussed that the result is not sensitive to the ratio. The following section will theoretically investigate how different M values affect $Var[\hat{Z}]$.

4.2. Optimal sample size determination

In this section, we derive the optimal sample size M when the positive integer constraint is not imposed on the budget allocation. Note that the variance of $\hat{Z}$ can be expressed as follows:

$$\begin{aligned} Var[\hat{Z}] &= Var\left[\frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}\sum_{j=1}^{N_i}\hat{s}_j(X_i)\frac{f(X_i)}{q(X_i)}\right] \\ &= \frac{1}{M^2}\left[Var_q\left[E\left[\sum_{i=1}^{M}\frac{1}{N_i}\sum_{j=1}^{N_i}\hat{s}_j(X_i)\frac{f(X_i)}{q(X_i)}|X\right]\right]\right. \\ &\quad \left. + E_q\left[Var\left[\sum_{i=1}^{M}\frac{1}{N_i}\sum_{j=1}^{N_i}\hat{s}_j(X_i)\frac{f(X_i)}{q(X_i)}|X\right]\right]\right] \\ &= \frac{1}{M}Var_q\left[E[\hat{s}(X_1)|X]\frac{f(X_1)}{q(X_1)}\right] \\ &\quad + \frac{1}{M^2}E_q\left[\sum_{i=1}^{M}\frac{1}{N_i}Var[\hat{s}(X_i)|X]\frac{f(X_i)^2}{q(X_i)^2}\right]. \end{aligned} \tag{4.5}$$

We plug the optimal budget allocation in (4.2), given M, into (4.5) to obtain

$$\begin{aligned} Var[\hat{Z}] &= \frac{1}{MN_T}\left[E_f\left[Var[\hat{s}(X)|X]\frac{f(X)}{q(X)}\right]\right. \\ &\quad \left. +(M-1)E_f\left[\sqrt{Var[\hat{s}(X)|X]}\right]^2\right] \\ &\quad + \frac{1}{M}Var_q\left[E[\hat{s}(X)|X]\frac{f(X)}{q(X)}\right] \\ &= \frac{1}{MN_T}[\kappa_1 + (M-1)\kappa_2] + \frac{1}{M}\kappa_3, \end{aligned} \tag{4.6}$$

where we define

$$\kappa_1 \equiv E_f\left[Var[\hat{s}(X)|X]\frac{f(X)}{q(X)}\right], \kappa_2 \equiv E_f\left[\sqrt{Var[\hat{s}(X)|X]}\right]^2,$$
$$\kappa_3 \equiv Var_q\left[E[\hat{s}(X)|X]\frac{f(X)}{q(X)}\right].$$

The detailed derivation of (4.6) is available in the online supplement. We note that κ_1, κ_2, and κ_3 are strictly positive constants due to the randomness of X and the unknown randomness in $\hat{s}(\cdot)$. Theorem 4.5 shows that $Var[\hat{Z}]$ is decreasing over M for a sufficiently large N_T.

Theorem 4.5. *There exists $N_T^* \in \mathbb{N}$ such that for all $N_T \geq N_T^*$, $Var[\hat{Z}]$ decreases in M; $M = N_T$ is the optimal sample size for $N_T \geq N_T^*$. Furthermore, if $f(x) = q(x)$, $Var[\hat{Z}]$ decreases in M for all $N_T \in \mathbb{N}$.*

Proof. Taking the derivative of $Var[\hat{Z}]$ in (4.6) with respect to M, we get

$$\frac{d}{dM}Var[\hat{Z}] = \frac{\kappa_2 - \kappa_1 - N_T\kappa_3}{M^2 N_T}. \tag{4.7}$$

Since κ_1, κ_2, and κ_3 are positive constants, the derivative is either positive or negative for a given N_T. Then, we can find $N_T^* = \min\{N_T : \frac{d}{M}Var[\hat{Z}] < 0\}$ and $\frac{d}{dM}Var[\hat{Z}] < 0$ for $N_T \geq N_T^*$.

If $f(x) = q(x)$, we obtain $\kappa_2 \leq \kappa_1$ by Jensen's inequality. Therefore, $\frac{d}{dM}Var[\hat{Z}] < 0$ holds for all $N_T \in \mathbb{N}$. □

The result in Theorem 4.5 states that $M = N_T$ is optimal, when N_T is sufficiently large. Since $M \leq N_T$, we know that $M \leq N_T$ as $M \to \infty$. Hence, $Var[\hat{Z}]$ decreases to zero eventually as N_T increases. It, however, is not *mathematically* clear that $Var[\hat{Z}]$ is decreasing over M for *any* fixed N_T. We prove that $Var[\hat{Z}]$ is a decreasing function of M for any fixed $N_T \in \mathbb{N}$ when $f(x) = q(x)$. We conjecture that $Var[\hat{Z}]$ also decreases for *practically almost all* $N_T \in \mathbb{N}$ even if $f(x) \neq q(x)$. This is because the derivative of $Var[\hat{Z}]$ in (4.7) is either positive or negative, implying that $Var[\hat{Z}]$ is either increasing or decreasing in M. If N_T is sufficiently large such that the numerator in (4.7) is negative, the derivative becomes negative. Otherwise, suppose that the derivative is positive. Then $M = 1$ becomes optimal, implying that we only need to take one input sample X_1 at the first-level simulation and assign all simulation budget N_T to X_1, which

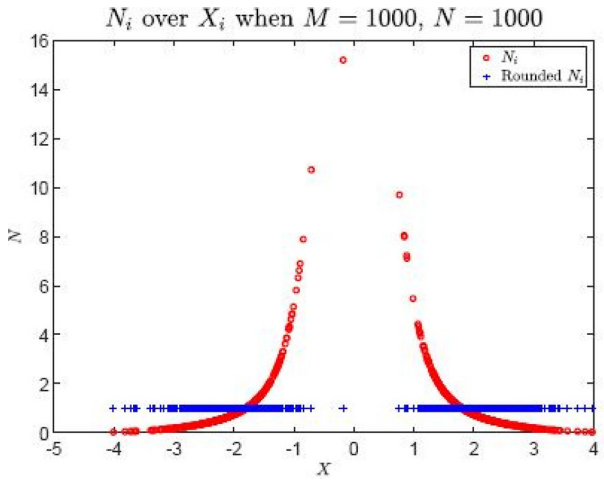


Figure 4.1. Example of N_i over X_i with $N_T = M = 1000$.

seems unreasonable. Therefore, we believe the derivative is negative and thus, $M = N_T$ is optimal in most cases.

Theorem 4.5 can be interpreted as the optimality at the full exploration. However, the resulting optimal N_i values likely take real values. A problem, hence, arises when we actually implement the simulation with $M = N_T$. To maintain the unbiasedness of $\hat{Z}$, each X_i for $i \in \{1, ..., M\}$ should have at least one instance of the second-level simulation (Choe *et al.*, 2015). In other words, we should have $N_i \geq 1$. This implies that with $M = N_T$, we have to assign $N_i = 1$ budget to each X_i in the second-level simulation. To illustrate, Figure 4.1 depicts the optimal allocation (see the red circles) under $M = N_T$ using the numerical example in Section 5.1. In this specific example, $|x_i|$ values around $|x| = 2$ are mostly sampled from $q^*(x)$ in (4.3). On the other hand, the optimal allocation N_i for each sampled x_i assigns more budgets to small $|x_i|$. This may look counter-intuitive, but it is not. Large second-level budget allocation around $|x| = 0$ is due to small $E[\hat{s}(x_i)]$, increasing N_i in (4.4). Furthermore, the budget allocation is *jointly* determined by importance sampling density and the second-level allocation. That is, as seen in Figure 4.1, importance sampling takes more samples in regions having large $E[\hat{s}(x)]$ and $Var[\hat{s}(x)]$ around $|X| = 2$. Thus, the sum of N_i values in those regions is larger than that in regions with few x_i values (around $|X| = 0$). Here, the key point is that theoretically optimal N_i values are different among sampled inputs under $M = N_T$. However, the implementable version of the theoretically optimal allocation assigns only one replication for each input (i.e., $N_i = 1$) (see the blue dots in Figure 4.1). Doing so loses the optimality. The following section shows that such rounding affects the optimality.

4.3. Comparison of different exploration-only strategies

The fact that the theoretical optimality is achieved at $M = N_T$ leads us to additionally consider an alternative estimator, denoted by $\hat{Z}_2$, which is intentionally designed to explore the input area without exploitation. Hence, before investigating how rounding of the real-valued optimal N_i values affects the optimality in $\hat{Z}$, we investigate the *exploration-only* estimator $\hat{Z}_2$ (Choe *et al.*, 2015):

$$\hat{Z}_2 \equiv \frac{1}{N_T}\sum_{i=1}^{N_T} \hat{s}(X_i)\frac{f(X_i)}{q(X_i)}. \tag{4.8}$$

This estimator runs the simulator once at each sampled X_i, that is, no replication at X_i. Thus, it does not exploit the stochastic response surface, rather it permits exploration only.

One might think that the estimator $\hat{Z}_2$ in (4.8) can be regarded as a special case of the original estimator $\hat{Z}$ when $M = N_T$ and $N_i = 1$ for all $i \in \{1, ..., N_T\}$. It, however, is not true because the optimal importance sampling density for $\hat{Z}$ and that for $\hat{Z}_2$ are different. Let $q_2^*(\cdot)$ denote the optimal importance sampling density that minimizes $Var[\hat{Z}_2]$. Chen and Choe (2019) derived the optimal $q_2^*(\cdot)$ as follows.

Theorem 4.6 (Chen and Choe, 2019). *For the estimator $\hat{Z}_2$ in (4.8), the optimal importance sampling density $q_2^*(\cdot)$ that minimizes $Var[\hat{Z}_2]$ is given by*

$$q_2^*(x) = \frac{1}{C_{q_2^*}} f(x)\sqrt{E\left[\hat{s}(x)^2\right]}, \tag{4.9}$$

where $C_{q_2^*}$ is a normalizing constant. Let $\hat{Z}_2^*$ be $\hat{Z}_2$ with the optimal importance sampling density $q_2^*(\cdot)$. Then, $Var[\hat{Z}_2^*]$ is as follows:

$$Var\left[\hat{Z}_2^*\right] = \frac{1}{M}\left[E_f\left[\sqrt{E\left[\hat{s}(X)^2|X\right]}\right]^2 - E_f[E[\hat{s}(X)|X]]^2\right].$$

We can easily notice the difference between $q^*(\cdot)$ in (4.3) and $q_2^*(\cdot)$ in (4.9). Therefore, the exploration-only estimator $\hat{Z}_2$ in (4.8) cannot be viewed as a special case of the original estimator $\hat{Z}$ when $M = N_T$. Moreover, we would like to point out that, whereas $\hat{Z}_2$ does not allow replication at each sampled X_i, it naturally accounts for heterogeneous noise over the input space. Considering that $E\left[\hat{s}(x)^2\right] = Var[\hat{s}(x)] + E[\hat{s}(x)]^2$, $q_2^*(\cdot)$ samples more inputs in regions with greater variance and expectation, striking a balance between exploration and exploitation.

Noting that both the original estimator $\hat{Z}$ with theoretically optimal input size $M = N_T$ and the exploration-only estimator $\hat{Z}_2$ suggest the full exploration strategy, we compare their variances. Furthermore, we also consider the implementable version of the original optimal estimator by setting $N_i = 1$. Specifically, let us consider three different cases as follows:

- Case 1 represents the theoretically optimal solution. It uses the original estimator $\hat{Z}$ with $M = N_T$ and employs the optimal importance sampling density $q^*(\cdot)$ in (4.3). The optimal N_i values are obtained from (4.4). Let $\hat{Z}_1^*$ denote the resulting theoretically optimal estimator.
- Case 2 employs the optimal exploration-only estimator $\hat{Z}_2^*$ with the optimal importance sampling density $q_2^*(\cdot)$.
- Case 3 uses $\hat{Z}$ with $q^*(\cdot)$ when $M = N_T$. The difference from Case 1 is to set $N_i = 1$, because $N_i \geq 1$ for all

$i \in \{1, \ldots, M\}$ is needed for *actual implementation* as mentioned at the end of Section 4.2. Let $\hat{Z}_3^*$ denote the implementable version of $\hat{Z}_1^*$.

Mathematically, we can write the estimators for three cases as follows:

$$\hat{Z}_1^* \equiv \frac{1}{N_T}\sum_{i=1}^{N_T}\frac{1}{N_i^*}\sum_{j=1}^{N_i^*}\hat{s}_j(X_i)\frac{f(X_i)}{q^*(X_i)},$$

$$\hat{Z}_2^* \equiv \frac{1}{N_T}\sum_{i=1}^{N_T}\hat{s}(X_i)\frac{f(X_i)}{q_2^*(X_i)}, \hat{Z}_3^* \equiv \frac{1}{N_T}\sum_{i=1}^{N_T}\hat{s}(X_i)\frac{f(X_i)}{q^*(X_i)},$$

where $\hat{Z}_j^*$ is the estimator of Case j for $j \in \{1,2,3\}$ and N_i^* in $\hat{Z}_1^*$ is the theoretically optimal budget allocation in (4.4) for $i \in \{1, ..., N_T\}$. Note that $\hat{Z}_1^*$ is the same as $\hat{Z}$ with $M = N_T$. We summarize the estimators in Table 4.1.

In Case 1, the resulting N_i values are likely real-valued numbers, so we cannot conduct actual experiments. We can, however, still obtain the theoretically optimal variance using (4.6). Theorem 4.7 compares the variance of $\hat{Z}_1^*, \hat{Z}_2^*$, and $\hat{Z}_3^*$.

Theorem 4.7.

$$Var\left[\hat{Z}_1^*\right] \le Var\left[\hat{Z}_2^*\right] \le Var\left[\hat{Z}_3^*\right].$$

Proof. The proof is available in the online supplement. □

Theorem 4.7 provides important implications. First, given N_T, the variance ($Var[\hat{Z}_3^*]$) of Case 3, which is Case 1's implementable version, is larger than Case 1's variance. The gap between $Var[\hat{Z}_1^*]$ and $Var[\hat{Z}_3^*]$ comes from the rounding error of N_i values to make them into integers and keep the unbiasedness of the estimator. Second, the theoretically optimal variance ($Var[\hat{Z}_1^*]$) in Case 1 is smaller than the optimal variance ($Var[\hat{Z}_2^*]$) in Case 2. However, the variance ($Var[\hat{Z}_3^*]$) of Case 3 is larger than Case 2's variance. In summary, although Case 1 with the original estimator $\hat{Z}_1^*$ theoretically provides better performance than the estimator $\hat{Z}_2^*$, Case 1's implementable version (Case 3) performs worse than Case 2.

Note that all three cases take the *exploration-only* strategy, and there is no exploitation. Theorem 4.7 indicates that among the exploration-only options, $\hat{Z}_2^*$, which is intentionally designed to explore only, is better than the original estimator that implements the exploration with rounding.

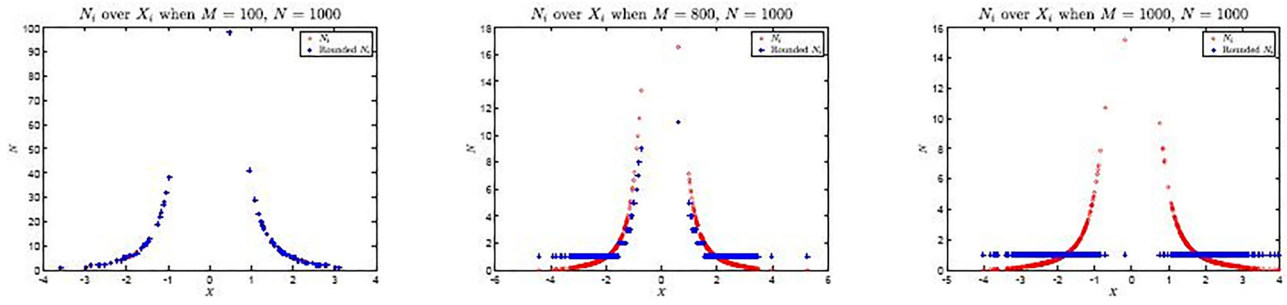
On the other hand, because we lose optimality due to rounding when the original estimator $\hat{Z}$ is used, the optimal M that considers rounding could lie between one and N_T. If we can find such an optimal M, $\hat{Z}$ could outperform $\hat{Z}_2^*$. In our implementation in Section 5, we actually observe that $\hat{Z}$ with $M < N_T$ generates a smaller variance than $\hat{Z}_2^*$ in some cases. However, the optimal M depends on the problem structure, and finding optimal M, either analytically or empirically, is not straightforward. In Section 5, we empirically study when the exploration-replication is better than the exploration-only strategy.

Before moving to the numerical studies, it is worthwhile to look into the effect of rounding to determine N_i. As M gets closer to N_T, we sample more inputs and thus, each input gets less budgets, so the rounding error becomes larger. To illustrate, Figure 4.2 depicts the optimal allocations of $N_T = 1000$ without integer constraints in red circles and the rounded integer allocation in blue circles in the one-dimensional example in Section 5.1. When M is 100, the difference is insignificant. However, as M gets larger, the difference becomes obvious. When $M = 1000$, the practical allocation $N_T = 1$ is substantially different from the optimal allocation.

Finally, as a remark, our importance sampling scheme is different from an importance sampling procedure used in Bayesian inference or Monte Carlo integration. We derive the optimal importance sampling density $q(\cdot)$ that can minimize the estimation variance. The density $q(\cdot)$ with a stochastic computer model is oftentimes complicated, so one cannot directly draw samples from $q(\cdot)$. Thus, we employ the rejection sampling to get independent and identically distributed samples in our implementation.

Table 4.1. Summary of different estimators.

Estimator	*Explanation*
$\hat{Z}$	original estimator with the input sample size M
$\hat{Z}_1^*$	$\hat{Z}$ with q^* and N_i^* for $i \in \{1, ..., N_T\}$
$\hat{Z}_2^*$	$\hat{Z}_2$ with q_2^*
$\hat{Z}_3^*$	$\hat{Z}$ with q^* and $N_i = 1$ for $i \in \{1, ..., N_T\}$



(a) $M = 100$ (b) $M = 800$ (c) $M = 1000$

Figure 4.2. N_i over X_i with different M values, $N_T = 1000$.

5. Numerical experiments

We conduct experiments with a one-dimensional numerical example for estimating the tail probability $P(Y > l)$ in Section 5.1. Section 5.2 estimates $E[Y^2]$ to confirm our findings in general settings other than the tail probability. In Section 5.3, a case study for a wind turbine simulator in Choe *et al.* (2015) is presented in our analysis framework. Based on these experiments, Section 5.4 discusses our findings. We note that more numerical experiments—the tail probability with three-dimensional inputs and the expected shortfall $E[Y\mathbb{I}_{\{Y>l\}}]$—are available in the online supplement.

5.1. One-dimensional example for estimating P(Y>l)

We first use the following example that estimates the tail probability of a random variable, $P(Y > l)$ (Choe *et al.*, 2015):

$$X \sim N(0,1), \quad Y|X = x \sim N\big(\mu(x), \sigma^2(x)\big), \tag{5.1}$$

where $\mu(x)$ and $\sigma(x)$ denote the true mean and standard deviation of $Y|X = x$, respectively, given by

$$\begin{aligned} \mu(x) &= 0.95x^2(1 + 0.5\cos(10x) + 0.5\cos(20x)), \\ \sigma(x) &= 1 + 0.7|x| + 0.4\cos(x) + 0.3\cos(14x). \end{aligned} \tag{5.2}$$

The total simulation budget N_T is 1000. We consider $\alpha = P(Y > l) = 0.05$.

Note that $\mu(x)$ and $\sigma(x)$ are used to find $E[\hat{s}(x)]$ and $Var[\hat{s}(x)]$ in the importance sampling densities. In practice, when the second-level simulation uses a black box computer model, $\mu(x)$ and $\sigma(x)$ are unknown (and thus, $E[\hat{s}(x)]$ and $Var[\hat{s}(x)]$ are unknown). This problem commonly arises in the nested simulation literature. Thus, the analysis assumes *pre-experiments* to build a (rough) surrogate model estimating the response surface from existing data or a small pilot sample (Choe *et al.*, 2015). To see the effect of an inaccurate surrogate, we consider the following estimates of $\mu(x)$ and $\sigma(x)$, which multiply cosine terms in (5.2) by a constant ρ, as in Choe *et al.* (2015):

$$\begin{aligned} \hat{\mu}(x) &= 0.95x^2(1 + 0.5\rho\cos(10x) + 0.5\rho\cos(20x)), \\ \hat{\sigma}(x) &= 1 + 0.7|x| + 0.4\rho\cos(x) + 0.3\rho\cos(14x). \end{aligned} \tag{5.3}$$

We will investigate how the performance of the estimators changes when we vary ρ from one to zero. Please note that using a single parameter ρ is just one way of controlling the surrogate accuracy in this example.

First, we compare the variance of the three estimators, $\hat{Z}_1^*, \hat{Z}_2^*$, and $\hat{Z}_3^*$, when the estimation of $\mu(x)$ and $\sigma(x)$ is exact ($\rho = 1.0$) in Figure 5.1. Here, $\sigma[\hat{Z}_1^*]$ is the theoretical standard deviation, and $\sigma[\hat{Z}_2^*]$ and $\sigma[\hat{Z}_3^*]$ are the sample standard deviations, each obtained from 1000 experiments. The results agree with Theorem 4.7, i.e., $Var[\hat{Z}_1^*] = \sigma^2[\hat{Z}_1^*] \le Var[\hat{Z}_2^*] \le Var[\hat{Z}_3^*]$. Moreover, the performance of $\hat{Z}_2^*$ is comparable to the theoretically optimal estimator $\hat{Z}_1^*$; $\sigma[\hat{Z}_1^*]$ is close to $\sigma[\hat{Z}_2^*]$. On the other hand, the difference between $\sigma[\hat{Z}_1^*]$ and $\sigma[\hat{Z}_3^*]$ is not negligible, demonstrating that the rounding could affect the optimality significantly when $M = N_T$.

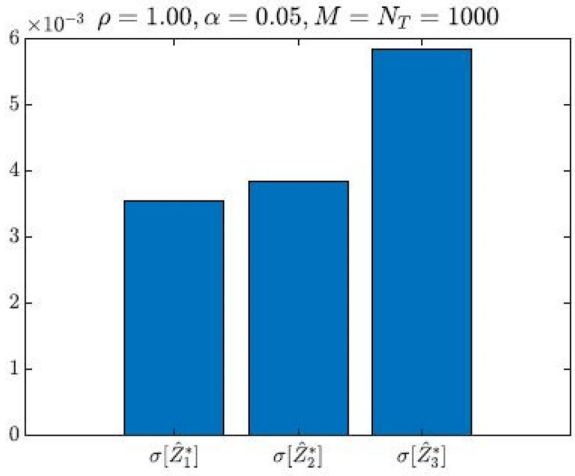


Figure 5.1. Comparison of $\sigma[\hat{Z}_1^*], \sigma[\hat{Z}_2^*]$ and $\sigma[\hat{Z}_1^*]$ with $\rho = 1.0, \alpha = 0.05$, and $N_T = 1000$.

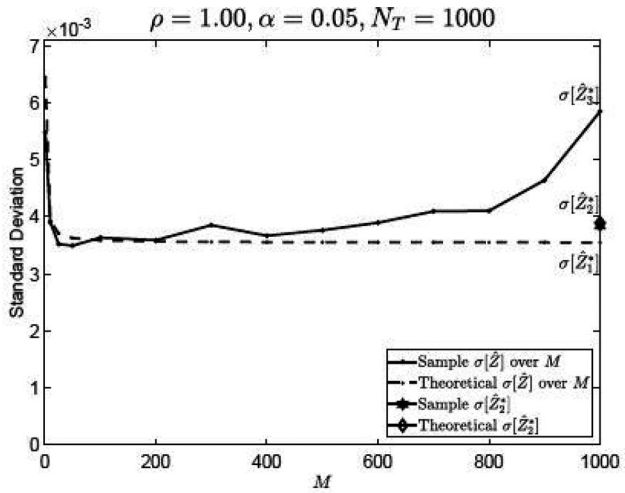


Figure 5.2. $\sigma[\hat{Z}]$ over M and $\sigma[\hat{Z}_2^*]$ with $\rho = 1.0, \alpha = 0.05$, and $N_T = 1000$.

We further investigate how the performance of $\hat{Z}$ changes with different M. Figure 5.2 illustrates $\sigma[\hat{Z}]$ over M. The dotted line denotes the theoretical standard deviation of $\hat{Z}$ over M. The solid line represents the sample standard deviation of $\hat{Z}$ with 1000 experiments for each M, where each real-valued N_i is rounded to its nearest natural number. We also include $\sigma[\hat{Z}_1^*], \sigma[\hat{Z}_2^*]$, and $\sigma[\hat{Z}_3^*]$ in the right-most vertical line at $M = 1000$; the diamond and star markers indicate the theoretical and sample standard deviation $\sigma[\hat{Z}_2^*]$, respectively. As $\hat{Z}_2$ does not involve the rounding issue, the sample standard deviation is very close to its corresponding theoretical sample standard deviation, unlike $\hat{Z}_1^*$.

In Figure 5.2, we observe that theoretical standard deviation of $\hat{Z}$ decreases over M, as shown in Theorem 4.5. The actual sample standard deviation decreases in the beginning, but it starts to increase because the rounding error becomes exacerbated as M gets close to N_T, i.e., as we assign less budget to each X_i. In the end, $\sigma[\hat{Z}_3^*]$ is much larger than

Table 5.1. $\sigma[\hat{Z}]$ over M and $\sigma[\hat{Z}_2^*]$ with exact estimation of $E[\hat{s}(x)]$ and $Var[\hat{s}(x)]$ for $N_T = 1000, \rho = 1.0,$ and $\alpha = 0.05$.

	$\sigma[\hat{Z}]$							
M	1	50	100	300	500	700	1000	$\sigma[\hat{Z}_2^*]$
Sample	0.0055	0.0035	0.0036	0.0038	0.0038	0.0041	0.0058 ($\sigma[\hat{Z}_3^*]$)	0.0038
Theoretical	0.0064	0.0036	0.0036	0.0036	0.0035	0.0035	0.0035 ($\sigma[\hat{Z}_1^*]$)	0.0039

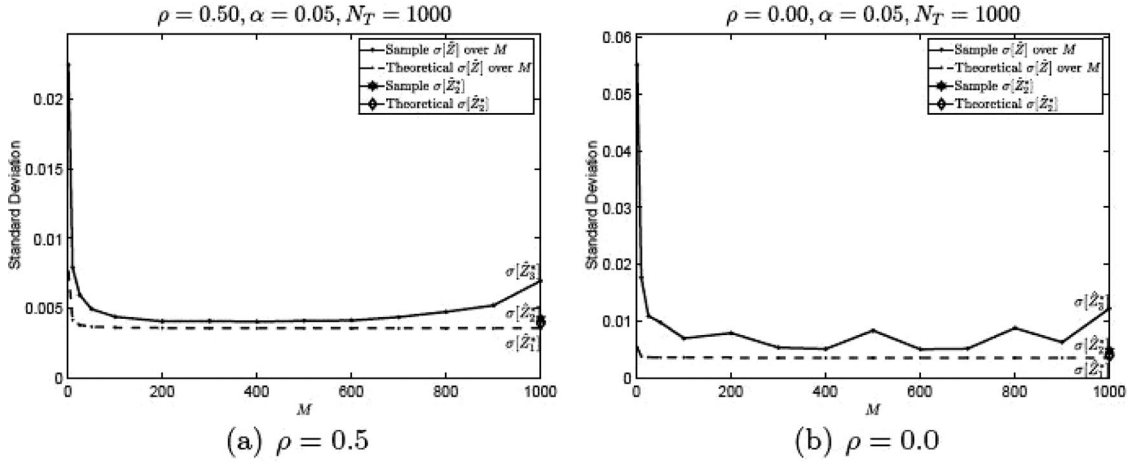


Figure 5.3. $\sigma[\hat{Z}]$ over M and $\sigma[\hat{Z}_2^*]$ with $\rho \in \{0.0, 0.5\}$.

$\sigma[\hat{Z}_1^*]$. The detailed values of standard deviations are summarized in Table 5.1.

One interesting aspect is that the decreasing rate of $\sigma[\hat{Z}]$ over M is quite fast. This is because the derivative of $Var[\hat{Z}]$ in (4.7) decays at a rate of $1/M^2$, given N_T. Similarly, $\sigma[\hat{Z}_3^*]$ also decreases fast when M is small. Considering that the theoretical standard deviation decreases fast when the rounding error increases along with M, we may consider choosing small M. For example, from Figure 5.2 and Table 5.1, we get the smallest sample standard deviation for $\hat{Z}$ when M is around 50. We observe similar patterns with different N_T and α values. And, $\sigma[\hat{Z}]$ with $M = 50$ is slightly smaller than $\sigma[\hat{Z}_3^*]$.

Then, can we conclude that exploration-exploitation with a small sample size M is more beneficial than the exploration-only strategy? Note that the optimal importance sampling density functions, $q^*(\cdot)$ and $q_2^*(\cdot)$, for the original and exploration-only estimators include the mean and variance of $\hat{s}(x)$, which we assume to know so far with $\rho = 1.0$ in (5.3), but they should be estimated in reality (Chen and Choe, 2019). The estimation of the mean and variance of $\hat{s}(x)$ definitely affects the quality of the importance sampling density. We look into this issue in detail with different ρ values.

Figure 5.3 shows $\sigma[\hat{Z}]$ over M and $\sigma[\hat{Z}_2^*]$ with inexact estimation of $E[\hat{s}(x)]$ and $Var[\hat{s}(x)]$ ($\rho = 0.5$ and $\rho = 0.0$). The detailed values are also reported in Table 5.2. We notice that unlike the previous result with $\rho = 1$, the best M value becomes larger as the estimation becomes less accurate – smaller ρ values. It is around 300 and 500 when ρ is 0.5, whereas $M = 600$ to 700 yields small $\sigma[\hat{Z}]$ for $\rho = 0.0$. Recall that the optimal M was around 50 when $\rho = 1.0$. When the estimation is inaccurate, the importance sampler draws inputs from the unimportant input area. With small M, $\hat{Z}$ unnecessarily exploits the response surface at unimportant X_i values. Therefore, inaccurate estimation of $E[\hat{s}(x)]$ and $Var[\hat{s}(x)]$ requires more exploration (large M) than exploitation to reduce the variance, but doing so inevitably increases the rounding error in $\hat{Z}$. Without the quantitative measure for evaluating the estimation accuracy, it is not straightforward to find the optimal M value in $\hat{Z}$.

Notably, we observe that $\sigma[\hat{Z}_2^*]$ is robust to the estimation quality. Let us compare $\sigma[\hat{Z}_2^*]$ in the last column of Tables 5.1 and 5.2. With different values of ρ, we obtain similar results in $\hat{Z}_2^*$. On the contrary, the performance of $\hat{Z}$ and $\hat{Z}_3^*$ appears to be substantially affected by the estimation accuracy. Moreover, when $\rho = 0.5, \sigma[\hat{Z}_2^*]$ (0.0042) is close to the smallest value (0.0041) of $\sigma[\hat{Z}]$. Interestingly, when $\rho = 0.0, \sigma[\hat{Z}_2^*]$ is smaller than $\sigma[\hat{Z}]$ for any M, demonstrating that $\hat{Z}_2^*$ is more robust to the estimation accuracy.

In summary, the exploration-only estimator $\hat{Z}_2^*$ with the importance sampling density q_2^* attracts our attention. This estimator is free from the rounding error, and thus, the sample standard deviation nearly coincides with the theoretical standard deviation. Its resulting standard deviation is close to the optimal one in $\hat{Z}$ that considers both exploration and exploitation. Furthermore, Figure 5.3 and Table 5.2 show that it performs well, even when the estimation of $E[\hat{s}(x)]$ and $Var[\hat{s}(x)]$ is inaccurate. We conduct additional experiments with a wide range of parameters (α and N_T) and observe similar results.

5.2. Expectation of $Z = Y^2$

Importance sampling can be more effective for problems where important input regions are narrow. The expectation involving tail regions, such as tail probability (the

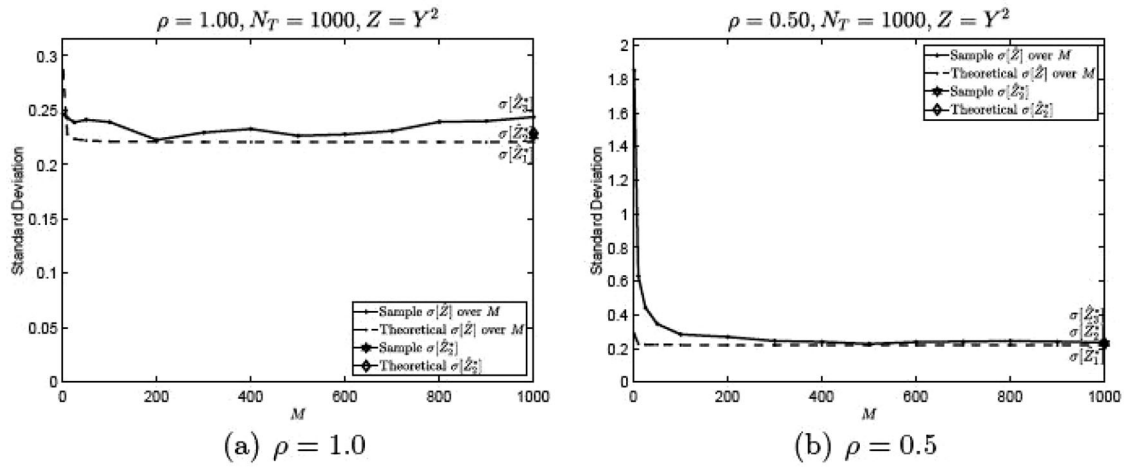


Figure 5.4. $Z = Y^2$ example.

expectation of an indicator function) in the previous section and the expected shortfall available in the online supplement, are such examples for effectively applying the importance sampling scheme. Theoretically, however, our approach should work in general cases, for example, the expectation of a polynomial function of a random variable. This section conducts an experiment when the random variable Z is the square of Y, i.e., $Z = Y^2$.

Figure 5.4 plots the standard deviation over M with the same distributions of X and $Y|X = x$ presented in Section 5.3. In this example, the effects of rounding errors are not as obvious as in the previous examples. This is because the importance sampling density covers the whole input region, and thus, we may not take a full advantage of the importance sampling principle. However, we still obtain consistent results as in the previous examples. That is, the estimator $\hat{Z}_2^*$ performs excellently and its standard deviation is close to the theoretically optimal standard deviation in both cases. The results demonstrate the general applicability of our approach.

5.3. Case study

We employ the NREL wind turbine simulator (Jonkman and Buhl Jr., 2005; Jonkman, 2009). The NREL wind turbine simulator generates various load responses as simulation outputs. Among the load responses, we consider two load types – edgewise and flapwise bending moments which represent parallel and perpendicular load responses to the blade rotor plane, respectively (Byon *et al.*, 2016; Ding, 2019). These two bending moments are important load responses in wind turbine reliability (Moriarty, 2008).

Specifically, the IEC design standard, IEC 61400-1 (International Electrotechnical Commission, 2005), specifies several Design Load Cases (DLCs). Among them, estimating the failure probability (or, probability of exceedance (POE)) with Y being the maximum load response during a specific interval (e.g., 10 minutes) is required in DLC 1.1. Following the design standard, we consider a maximum response (flapwise and edgewise moments) during 10-minute turbine operation as the response variable, and 10-minute average wind speed as the input variable. In this case study, we employ the truncated Rayleigh distribution between 3 m/s and 25 m/s as the wind speed density, as in Moriarty (2008).

In our analysis, $l = 8600$ kNm and 13,800 kNm are used as resistance levels for edgewise and flapwise moments, respectively. With these resistance levels, the estimated failure probability $P(Y > l)$ is around 0.05 in both load types. To estimate $P(Y > l)$, one can use the Crude Monte Carlo (CMC) sampling that samples wind speed from its original density function. However, to estimate the POE with high accuracy, CMC requires large computational budgets. The importance sampling scheme discussed in our study allows us to improve the estimation accuracy with limited budgets by reweighting the sampling efforts to observe exceedance events more frequently.

To implement the importance sampling, $E[\hat{s}(x)]$ and $Var[\hat{s}(x)]$ are approximated with the nonhomogeneous Generalized Extreme Value (GEV) distribution where the location and scale parameters are formulated as spline functions of wind speeds (Lee *et al.*, 2013; You *et al.*, 2017). To fit the GEV distribution, a pilot sample that consists of 600 observations of (X, Y) is used. The detailed simulation setting can be found in Choe *et al.* (2015).

Table 5.3 summarizes the sample standard deviations reported in Choe *et al.* (2015), obtained from 50 experiments for each case. For the total simulation budget, $N_T = 1000$ and 2000 are, respectively, used for the edgewise and flapwise bending moment in each experiment. The results include the sample standard deviations of the original estimator $\hat{Z}$ with four different M/N_T ratios (M/N_T =10%, 30%, 50%, and 80%) and of the exploration-only estimator $\hat{Z}_2^*$. Among the four different values, $M/N_T = 10\%$ generates the smallest sample standard deviation for edgewise moments, whereas 30% appears to perform best for flapwise moments. As M increases, the performance of $\hat{Z}$ becomes deteriorated.

From these results, we can conclude that it is not straightforward to determine the practically optimal M before trying different M values. However, the exhaustive search for M adds significant computational burden, which contrasts with the fundamental goal of importance sampling to expedite the simulation process. On the other hand, the

Table 5.2. $\sigma[\hat{Z}]$ over M and $\sigma[\hat{Z}_2^*]$ with inexact estimation of $E[\hat{s}(x)]$ and $Var[\hat{s}(x)]$ for $N_T = 1000$ and $\alpha = 0.05$.

	$\sigma[\hat{Z}]$							
M	1	50	100	300	500	700	1000 ($\sigma[\hat{Z}_3^*]$)	$\sigma[\hat{Z}_2^*]$
$\rho = 0.5$	0.0225	0.0049	0.0044	0.0041	0.0041	0.0044	0.0070	0.0042
$\rho = 0.0$	0.0552	0.0098	0.0070	0.0053	0.0083	0.0051	0.0122	0.0048

Table 5.3. $\sigma[\hat{Z}]$ over M and $\sigma[\hat{Z}_2^*]$ for the wind turbine case study (Choe *et al.*, 2015).

	$\sigma[\hat{Z}]$				
M/N_T	10%	30%	50%	80%	$\sigma[\hat{Z}_2^*]$
Edgewise	0.0016	0.0018	0.0022	0.0022	0.0020
Flapwise	0.0034	0.0028	0.0032	0.0033	0.0032

exploration-only estimator in the last column in Table 5.3 provides reasonably good results in both output types.

5.4. Observations and discussion

Recall that the two-level simulation with stochastic response faces the trade-off in estimating a random quantity: whether to explore more inputs or exploit the response surface at sampled points in more detail through replication. Based on the theoretical analysis in Section 4 and empirical studies in this section, we summarize our observations as follows:

1. When $E[\hat{s}(x)]$ and $Var[\hat{s}(x)]$ are well-estimated, the original estimator $\hat{Z}$ with relatively small M and the exploration-only estimator $\hat{Z}_2^*$ provide comparably good performance.
2. However, when the estimation of $E[\hat{s}(x)]$ and $Var[\hat{s}(x)]$ is inaccurate, $\hat{Z}_2^*$ provides more robust and better performance than $\hat{Z}$.

Recall that the goal of importance sampling is to focus the efforts on narrow input regions that really matter. With inaccurate estimates of $E[\hat{s}(x)]$ and $Var[\hat{s}(x)]$, importance sampling miss-guides sampling efforts to less important regions. Even worse, the original estimator $\hat{Z}$ replicates in those regions. This is why the exploration-only estimator performs better with an inaccurate surrogate.

In nested simulation, a surrogate model is typically assumed to exist or it is constructed from a small-scale pilot experiment. As such, surrogate models are likely inaccurate. Although the sampling-based importance sampling approach provides an unbiased estimator under certain conditions even with an inaccurate surrogate, the surrogate's quality affects the estimation performance. In practice, it is not straightforward to determine the surrogate's accuracy. As such, the exploration-only strategy with $\hat{Z}_2^*$, which provides consistently reliable performance, appears to be an adequate choice in most cases. This implication is also supported by the fact that the theoretical optimality suggests $M = N_T$ when $\hat{Z}$ is used as the estimator. We note that our analysis has been conducted in a general setting without restrictive assumptions, except the continuous sample space of the input X, as explained in Section 4.2. Thus, we believe our conclusion can be applied to a wide range of applications.

As a final remark, our advocacy for the exploration-only estimator may sound contradictory to the recommendations from the literature, but it is not. For example, in the study by Binois *et al.* (2019), which aims to build a globally accurate GP emulator, replication turns out to be a better choice, when the variance of the response surface is high. Recall that the optimal importance sampling density $q_2^*(x)$ in (4.9) of the exploration-only estimator draws more inputs in regions with greater variance and expectation. Therefore, even without replication, it considers the second-level variance to balance the trade-off between exploration and exploitation, which aligns with the result in Binois *et al.* (2019). Furthermore, we would like to point out that the fundamental idea of importance sampling is to focus on the narrow important input region of X. Thus, the underlying premise is that the first-level stochasticity is larger than the second-level noise. Thus, exploration over the important input region, characterized by the importance sampling density, has merits.

6. Conclusion

This article studies a simulation budget allocation problem under a two-level simulation framework when importance sampling is employed at the first-level. Importance sampling has been widely used in rare event analysis, such as reliability problems and financial risk analysis. Most importance sampling studies consider deterministic computer models where the optimal allocation does not need replication and thus, they aim to solely optimize the first-level simulation. With the increased popularity of stochastic computer models, how to balance the trade-off of exploration vs. replication at both levels becomes an important problem. Although importance sampling schemes for stochastic computer models have been studied in the literature (Choe *et al.*, 2015), no guidelines are provided to address such a trade-off. This study provides theoretical justification and practical guidelines on how to allocate sampling budgets, gain insights on the stochastic importance sampling schemes, and suggest an effective sampling strategy. To the best of our knowledge, this article is the first study to optimize the resource allocation at both levels in the importance sampling framework.

We plan to make several extensions for our future work. First, we assume that the random input vector of the first-level simulation has a continuous density function. The nested simulation literature in financial engineering often concerns discrete portfolios for the first-level simulation. We will extend our analysis in the discrete setting at the first-level simulation. Second, we observe that the estimation accuracy of $E[\hat{s}(x)]$ and $Var[\hat{s}(x)]$ plays an important role in the two-level simulation. In the future, we plan to employ the adaptive surrogate modeling strategy (Binois *et al.*, 2019) to estimate them and incorporate it into the importance sampling framework, so that the estimation accuracy and computational efficiency can be further improved in practical implementation.

Next, it is known that importance sampling is not effective for high-dimensional problems in general, due to several challenges. The importance sampling density $q(\cdot)$ involves a normalizing constant that requires integration over input variables. When the input dimension is high, integration over multiple variables causes a critical computational issue. To avoid this issue, one can use a *self-normalized estimator* that does not require a normalizing constant (Owen, 2013). The self-normalized estimator, however, is biased with a small-size sample while it is a consistent estimator. Furthermore, we assume that the cost of sampling input variables is negligible relative to that of a target variable. For the high-dimensional problems where the cost of sampling input variables is considerable, one can consider a Markov chain Monte Carlo approach as an alternative to the rejection sampling, method; however, theoretical properties need to be re-investigated, due to the dependency of sampled points. On the other hand, we believe the importance sampling scheme studied in this aritcle has the potential to handle high-dimensional problems. Due to the parsimonious principal in typical engineering systems, not all input variables are equally important. Instead, a small number of selected input variables mainly affect the system response. Therefore, we can regard those important variables as main inputs and apply the importance sampling principal solely to them, while treating others as stochastic noise. The extension of this work for high-dimensional problems remains a subject for future study.

Lastly, for real-world applications, a guideline to select the total budget N_T would be beneficial and important for practitioners. We plan to adaptively increase the sample size until some criterion is satisfied. One such criterion could be a Coefficient Of Variation (COV). For example, if COV is smaller than a pre-specified threshold, we can sequentially add more samples. We would like to mention that the exploration-only estimator provides a better platform for adaptively deciding N_T. The original estimator allocates budgets to all sampled inputs at once, and thus, it is less appropriate in this sequential sampling. Although we can add a batch of samples and allocate budgets in each batch with the allocation rule in (4.4), rounding error would be exacerbated when the batch size is small. The exploration-only estimator does not face such issue, since it does not permit replication. We hope to extend our framework for further improving the budget determination and analyzing theoretical and practical properties with adaptive sample sizes in our future study.

Acknowledgments

We would like to thank the Editor, Department Editor, Associate Editor and anonymous reviewers for their constructive comments on various aspects of this work.

Funding

This work was partially supported by the Basic Science Research Program through National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016R1D1A1B04933453) and the U.S. National Science Foundation (Grant/Award number: IIS-1741166).

Notes on contributors

Young Myoung Ko received BS and MS degrees in Industrial Engineering from Seoul National University, Seoul, South Korea, and a PhD degree in industrial engineering from Texas A&M University, College Station, TX, USA, in 1998, 2000, and 2011 respectively. He is currently an associate professor with the Department of Industrial and Management Engineering, Pohang University of Science and Technology (POSTECH), Pohang, South Korea, where he focuses on simulation and optimization of stochastic systems, such as telecommunication networks, ICT infrastructure, and renewable energy systems.

Eunshin Byon is an Associate Professor in the Department of Industrial and Operations Engineering at the University of Michigan. She received her BS and MS in industrial and systems engineering from the Korea Advanced Institute of Science and Technology (KAIST) and a PhD in industrial and systems engineering from Texas A&M University. Her research interests include optimizing operations of renewable systems, data science, quality and reliability engineering, and sustainability. She is a member of IISE, INFORMS, and IEEE.

References

Ankenman, B., Nelson, B.L. and Staum, J. (2010) Stochastic kriging for simulation metamodeling. *Operations Research*, **58**(2), 371–382.

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

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

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

Cao, Q.D. and Choe, Y. (2019) Cross-entropy based importance sampling for stochastic simulation models. *Reliability Engineering & System Safety*, **191**, 106526.

Chapelle, O. and Li, L. (2011) An empirical evaluation of Thompson sampling. *Advances in Neural Information Processing Systems*, **24**, 2249–2257.

Chen, Y.C. and Choe, Y. (2019) Importance sampling and its optimality for stochastic simulation models. *Electronic Journal of Statistics*, **13**(2), 3386–3423.

Choe, Y., Byon, E. and Chen, N. (2015) Importance sampling for reliability evaluation with stochastic simulation models. *Technometrics*, **57**(3), 351–361.

Choe, Y., Lam, H. and 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. and Byon, E. (2016) Computationally efficient uncertainty minimization in wind turbine extreme load assessments. *Journal of Solar Energy Engineering*, **138**(4), 041012.

Ding, Y. (2019) *Data Science for Wind Energy*, CRC Press, Boca Raton, FL.

Frazier, P.I. (2018) A tutorial on Bayesian optimization. Available at: https://arxiv.org/pdf/1807.02811.pdf. (accessed 10 August 2020).

Gittins, J.C. and Jones, D.M. (1979) A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, **66**(3), 561–565.

Glasserman, P., Heidelberger, P. and Perwez, S. (1999) Importance sampling and stratification for value-at-risk, in *Proceedings of the Sixth International Conference on Computational Finance*, MIT Press, New York, NY, pp. 7–24.

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

Glynn, P.W. and Iglehart, D.L. (1989) Importance sampling for stochastic simulations. *Management Science*, **35**(11), 1367–1392.

Goetz, J., Tewari, A. and Zimmerman, P. (2018) Active learning for non-parametric regression using purely random trees. *Advances in Neural Information Processing Systems*, **31**, 2537–2546.

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

Gramacy, R.B. and Lee, H.K.H. (2009) Adaptive design and analysis of supercomputer experiments. *Technometrics*, **51**(2), 130–145.

International Electrotechnical Commission (2005) Wind Turbines - Part 1: Design Requirements, IEC/TC88, 61400–1 ed. 3, Geneva, Switzerland.

Jonkman, B.J. (2009) TurbSim user's guide: version 1.50, Technical Report NREL/TP-500-46198, National Renewable Energy Laboratory, Golden, CO.

Jonkman, J.M. and Buhl Jr., M.L. (2005) FAST User's Guide, Technical Report NREL/EL-500-38230, National Renewable Energy Laboratory, Golden, CO.

Lan, H., Nelson, B.L. and Staum, J. (2010) A confidence interval procedure for expected shortfall risk measurement via two-level simulation. *Operations Research*, **58**(5), 1481–1490.

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

Mockus, J. (1989) *Bayesian Approach to Global Optimization*, Springer, Dordrect, The Netherlands.

Moriarty, P. (2008) Database for validation of design load extrapolation techniques. *Wind Energy*, **11**(6), 559–576.

Owen, A.B. (2013) *Monte Carlo Theory, Methods and Examples*. Available at https://statweb.stanford.edu/~owen/mc/. (accessed 10 August 2020).

Pan, Q., Byon, E., Ko, Y. and Lam, H. (2020) Adaptive importance sampling for extreme quantile estimation with stochastic black box computer models. *Naval Research Logistics*, **67**, 524–547.

Pan, Q., Ko, Y.M. and Byon, E. (2021) Uncertainty quantification for extreme quantile estimation with stochastic computer models. *IEEE Transactions on Reliability*, **70**, 134–145.

Russo, D.J., Roy, B.V., Kazerouni, A., Osband, I. and Wen, Z. (2018) A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, **11**(1), 1–96.

Sinha, S. and Wiens, D.P. (2002) Robust sequential designs for nonlinear regression. *The Canadian Journal of Statistics*, **30**(4), 601–618.

Snoek, J., Larochelle, H. and Adams, R.P. (2012) Practical Bayesian optimization of machine learning algorithms, in *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, Curran Associates Inc., Lake Tahoe, NV, pp. 2951–2959. https://nips.cc/Conferences/2012.

Sun, Y., Apley, D.W. and Staum, J. (2011) Efficient nested simulation for estimating the variance of a conditional expectation. *Operations Research*, **59**(4), 998–1007.

Thompson, W.R. (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**(3/4), 285–294.

Wang, L., Chen, X., Kang, S., Deng, X. and Jin, R. (2020) Meta-modeling of high-fidelity FEA simulation for efficient product and process design in additive manufacturing. *Additive Manufacturing*, **35**, 101211.

Wang, W. and Haaland, B. (2019) Controlling sources of inaccuracy in stochastic kriging. *Technometrics*, **61**(3), 309–321.

Xiong, S., Qian, P.Z.G. and Wu, C.F.J. (2013) Sequential design and analysis of high-accuracy and low-accuracy computer codes. *Technometrics*, **55**(1), 37–46.

You, M., Byon, E., Jin, J. and Lee, G. (2017) When wind travels through turbines: A new statistical approach for characterizing heterogeneous wake effects in multi-turbine wind farms. *IISE Transactions*, **49**(1), 84–95.