# Computationally Efficient Reliability Evaluation With Stochastic Simulation Models

by

Youngjun Choe

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2016

Doctoral Committee:

Assistant Professor Eunshin Byon, Chair
Professor Judy Jin
Professor Vijayan N. Nair
Assistant Professor Chinedum E. Okwudire

To Annie

# ACKNOWLEDGEMENTS

I cannot thank enough my dissertation committee members, Professors Eunshin Byon, Judy Jin, Vijay Nair, and Chinedum Okwudire, who generously spent their time to guide me through the journey of the Ph.D. program. I am especially indebted to the committee chair, Professor Byon, for her countless hours of guidance and generous support that made this dissertation possible.

# PREFACE

The body of this dissertation consists of five chapters. Chapter I introduces the reader to the background of this dissertation research that covers three main subjects in Chapters II–IV. Chapter II is based on an article (*Choe et al.*, 2015), which was originally published by the American Society for Quality (ASQ) and the American Statistical Association (ASA). ASQ/ASA granted to me the right to reproduce the manuscript in this dissertation. Chapters III and IV are based on two working manuscripts. Chapter V concludes the dissertation with a summary and future research directions.

# TABLE OF CONTENTS

# LIST OF FIGURES

**Figure**

# LIST OF TABLES

**Table**

# LIST OF APPENDICES

**Appendix**

# LIST OF ABBREVIATIONS

**AIC** Akaike information criterion

**BIC** Bayesian information criterion

**BIS** Benchmark Importance Sampling

**CE** cross-entropy

**CE-AIS-GM** cross-entropy-based adaptive IS using Gaussian mixture

**CI** confidence interval

**CIC** cross-entropy information criterion

**CLT** central limit theorem

**CMC** crude Monte Carlo

**DIS** IS for deterministic simulation models

**DOE** Department of Energy

**EM** expectation–maximization

**EMCE** EM-based cross-entropy

**GAM** generalized additive model

**GAMLSS** generalized additive model for location, scale and shape

**GEV** Generalized Extreme Value

**GLM** generalized linear model

**GMM** Gaussian mixture model

**IS** importance sampling

**KS** Kolmogorov-Smirnov

**MCE** minimum cross-entropy estimator

**MCMC** Markov chain Monte Carlo

**MLE** maximum likelihood estimator

**NREL** National Renewable Energy Laboratory

**POE** probability of exceedance

**SIS** Stochastic Importance Sampling

**SIS1** Stochastic Importance Sampling Method 1

**SIS2** Stochastic Importance Sampling Method 2

# ABSTRACT

Computationally Efficient Reliability Evaluation With Stochastic Simulation Models

by

Youngjun Choe

Chair: Eunshin Byon

Thanks to advanced computing and modeling technologies, computer simulations are becoming more widely used for the reliability evaluation of complex systems. Yet, as simulation models represent physical systems more accurately and utilize a large number of random variables to reflect various uncertainties, high computational costs remain a major challenge in analyzing the system reliability.

The objective of this dissertation research is to provide new solutions for saving computational time of simulation-based reliability evaluation that considers large uncertainties within the simulation. This dissertation develops (a) a variance reduction technique for stochastic simulation models, (b) an uncertainty quantification method for the variance reduction technique, and (c) an adaptive approach of the variance reduction technique.

First, among several variance reduction techniques, importance sampling has been widely used to improve the efficiency of simulation-based reliability evaluation using deterministic simulation models. In contrast to deterministic simulation models whose simulation output is uniquely determined given a fixed input, stochastic simulation models produce random outputs. We extend the theory of importance sampling to

efficiently estimate a system's reliability with stochastic simulation models.

Second, to quantify the uncertainty of the reliability estimation with stochastic simulation models, we can repeat the simulation experiment multiple times. It, however, multiplies computational burden. To overcome this, we establish the central limit theorem for the reliability estimator with stochastic simulation models, and construct an asymptotically valid confidence interval using data from a single simulation experiment.

Lastly, theoretically optimal importance sampling densities require approximations in practice. As a candidate density to approximate the optimal density, a mixture of parametric densities can be used in the cross-entropy method that aims to minimize the cross-entropy between the optimal density and the candidate density. We propose an information criterion to identify an appropriate number of mixture densities. This criterion enables us to adaptively find the importance sampling density as we gather data through an iterative procedure.

Case studies, using computationally intensive aeroelastic wind turbine simulators developed by the U.S. Department of Energy (DOE)'s National Renewable Energy Laboratory (NREL), demonstrate the superiority of the proposed approaches over alternative methods in estimating the system reliability using stochastic simulation models.

# CHAPTER I

# Introduction

Thanks to the advance of computing and modeling technologies, simulations are employed in many applications to understand and analyze complex system behaviors. Stochastic simulation models are especially of interest due to the increasing importance of uncertainties observed in real world operations. Reliability evaluation of systems that need high reliability typically requires many replications of stochastic simulations to generate rare events of system failures. However, as simulation models represent physical systems more accurately, each simulation replication takes significant computational resources. This computational challenge calls for sophisticated approach in simulation-based reliability evaluation.

The objective of this dissertation research is to provide new solutions for saving computational time of simulation-based reliability evaluation that considers large uncertainties within the simulation. This dissertation develops (a) a new variance reduction technique for stochastic simulation models, (b) an uncertainty quantification method for the variance reduction technique, and (c) an adaptive approach of the variance reduction technique, which utilizes a novel information criterion to guide simulation process.

First, among several variance reduction techniques, importance sampling has been widely used to improve the efficiency of simulations, but its application has been

limited to deterministic simulation models. In contrast to deterministic simulation models whose simulation output is uniquely determined given a fixed input, stochastic simulation models produce random outputs. We extend the theory of importance sampling to estimate a system's reliability with stochastic simulation models. Given a budget constraint on total simulation replications, we develop a new approach, called stochastic importance sampling (SIS), which efficiently uses stochastic simulation models with unknown output distribution. Specifically, we derive the optimal importance sampling density and simulation allocation procedure that minimize the variance of a reliability estimator.

Second, to quantify the estimation uncertainty, one possible approach is to repeat the simulation experiment multiple times and obtain the sample standard deviation of the estimation. Repeating the experiment, however, multiplies computational burden. We develop an uncertainty quantification approach that does not require multiple experiments. Specifically, we establish the central limit theorems for SIS-based reliability estimators and construct asymptotically valid confidence intervals using the data obtained from a single simulation experiment.

Lastly, theoretically optimal importance sampling densities require some approximations in practice, such as the cross-entropy method. The standard cross-entropy method uses a parametric density to approximate the optimal importance sampling density. To overcome the rigidity of using a single parametric density, a mixture of parametric densities can be used. The performance of the mixture model-based cross-entropy method depends on the number of mixture components, yet no rigorous approach to decide the number of mixture components is available in the literature. We derive a new information criterion that can identify an appropriate number of component densities. By choosing the component number that minimizes the proposed criterion, we obtain the mixture model that asymptotically approaches the optimal density.

Case studies, using computationally intensive aeroelastic wind turbine simulators developed by the U.S. National Renewable Energy Laboratory, demonstrates the superiority of the proposed approaches over alternative methods in estimating the system reliability using stochastic simulation models.

In this chapter, we discuss the backgrounds of the three main subjects outlined above. Section 1.1 presents how computer simulations are used to evaluate the reliability of stochastic systems, why computational costs remain a challenge to the reliability evaluation, and what approaches can potentially address the challenges. Computational costs are closely related to the accuracy of reliability evaluation because higher accuracy (or smaller uncertainty) of the evaluation often requires more computational efforts. Section 1.2 discusses why quantifying the uncertainty of reliability evaluation is important and how existing studies approach the problem. Section 1.3 reviews the cross-entropy method that is widely used to approximate the optimal importance sampling method in order to reduce the computational burden in the simulation-based reliability evaluation.

## 1.1 Reliability Evaluation Using Monte Carlo Simulations

With the rapid growth of computing power over the last decades, computer simulation modeling has become very popular in many applications where real experiments are expensive, difficult, or perhaps impossible. These simulation models are often used to evaluate the reliability of large-scale, complex systems. For example, safety evaluation of a nuclear power plant often employs complex computer simulations (*D'Auria et al.*, 2006). The U.S. DOE's NREL has developed aeroelastic simulation tools to help wind turbine manufacturers design reliable wind power systems (*Jonkman and Buhl Jr.*, 2005; *Jonkman*, 2009).

This dissertation is concerned with reliability evaluation of systems based on Monte Carlo simulations that use repeated random sampling to understand prob-

abilistic behaviors of the systems modeled by computer simulations. As a measure of reliability, we consider the failure probability of the system. Failure events of interest can be soft failures (e.g., structural/mechanical loads on a wind turbine exceed a design resistance level) or hard failures (e.g., a wind turbine experiences structural or mechanical failures). To estimate the failure probability of a system, we use the computer simulation model of the system. The model is built by experts in the domain to reflect the system reliability at the detailed level. We regard the model as a black box to which we supply inputs in order to observe simulated outputs.

The crude Monte Carlo (CMC) method (*Kroese et al.*, 2011) is one of the most widely used methods to estimate the failure probability. In CMC, we sample inputs to the simulation model from a known distribution. This input distribution is usually estimated from the field data or specified by domain experts. Using the sampled inputs, we run the computer simulation model (or simulator) to generate corresponding outputs. Each replication (or a single run of the simulator) can be computationally intensive and time demanding. For example, the NREL aeroelastic simulators take roughly 1-min wall-clock time (on Intel Xeon CPU E31230 3.20GHz, RAM 8GB) to simulate 10-min actual operation of a wind turbine. After observing all simulation outputs, we count the number of replications where the system failed. The failure probability estimator in CMC is the proportion of the failure observations out of the total number of observations.

When people build and use a computationally intensive simulator to understand the reliability of a system, they usually face a very high standard on the system reliability (e.g., nuclear power plant, passenger aircraft, utility-scale wind turbine, etc.) because the reliability is very important for such systems. To meet the high reliability standard, the failure event should occur rarely, if any. A computer simulation model that well represents the actual system will similarly make us observe failure events rarely even if we repeat running the simulator many times with many different

random inputs. Furthermore, observing a single failure incidence out of millions of replications hardly gives us an accurate picture of all potential failure mechanisms. To have more concrete understanding of the potential failures or more accurate estimation of the failure probability, we need to observe more failure events than just few, requiring even more replications in CMC. This is the computational challenge associated with understanding rare events like failures of highly reliable systems.

Due to the sheer amount of required computation, CMC for the reliability evaluation is commonly implemented with high performance computing in practice (*Graf et al.*, 2015). For example, to evaluate the reliability of a wind turbine, *Moriarty* (2008) used grid computing with 60 desktops at NREL for 5 weeks, and *Manuel et al.* (2013) used cluster computing with 1,024 cores at Sandia National laboratories. To reduce the computational burden while maintaining the evaluation accuracy, a more sophisticated approach than CMC is needed.

A class of statistical methods called variance reduction techniques (*Kroese et al.*, 2011) aims to reduce the variance of the estimator that is based on computer simulations, while keeping the unbiasedness of the estimator. With the fixed amount of computational budget, such techniques can lead to a more accurate estimator of failure probability. On the flip side, to meet a target estimation accuracy, we can use less computational resources by employing variance reduction techniques.

Among several variance reduction techniques, importance sampling (IS) has been regarded as one of the most efficient methods (*Kroese et al.*, 2011) because IS can reduce the estimator variance to zero in theory for the deterministic simulation model. The underlying idea is to change the input distribution for the simulation so that we observe outputs of interest (e.g., failures) more frequently. The optimal input distribution or IS distribution exists regardless of whether the input is discrete, continuous, or mixed. We hereafter confine ourselves to continuous inputs for ease of presentation, but the extension to other input types is straightforward.

The existing IS assumes a deterministic relationship between simulation input and output (*Kahn and Marshall*, 1953). This assumption does not hold for the stochastic simulation model that generates random outputs for the same input. Stochastic simulation models are increasingly important for us to understand the reliability of systems under large uncertainties, because these simulators can incorporate numerous random variables within the simulators. Chapter II discusses the extension of IS to the stochastic simulation model and provides the optimal IS densities and allocation sizes that minimize the reliability estimator variance.

## 1.2 Uncertainty Quantification of Reliability Evaluation

When evaluating the reliability of a system under uncertainty using computer simulations, any practical estimator of reliability or failure probability is also subject to uncertainty. Without quantifying the uncertainty, it is hard to justify a point estimator, which may or may not be close enough to the true failure probability.

As one of the most common measures of the uncertainty, we can consider the variance of the estimator. To estimate the variance, a possible approach is repeating the estimation of failure probability and obtaining multiple estimates to compute the sample variance of the estimates as an estimator of the true variance. This empirical approach is commonly used when repeating the reliability estimation is computationally inexpensive. On the other hand, when obtaining even a single estimate is computationally expensive, it is necessary to rely on the theoretical property of the estimator to quantify the estimation uncertainty without repeating the estimation.

If an estimator is an average of random variables and generally well-behaved (e.g., the random variables being averaged have finite variances), then it is standard in the literature to establish the central limit theorem (CLT) of the estimator. CLT provides information on the distribution of the difference between the estimator and the true quantity being estimated. Specifically, CLT states that if the difference is

scaled up by the square root of the sample size (i.e., the total number of simulation replications), then the scaled difference tends to follow a normal distribution as the sample size grows (*Keener*, 2010).

Establishing CLT or knowing the asymptotic distributional property of the estimator allows us to quantify the estimation uncertainty more precisely than knowing only the (asymptotic) variance of the estimator. We can use the distribution information to build an asymptotic confidence interval (CI) that provides a confidence statement on our estimator of failure probability. When the sample size is large enough, the asymptotic CI will cover the true quantity being estimated with a high probability.

In the literature, CLT for the CMC estimator of the failure probability is well known (*Keener*, 2010). CLT for the IS estimator with deterministic simulation models is also well established (*Geweke*, 2005). However, the CLTs for the reliability estimators with stochastic simulation models have not been studied yet. Chapter III establishes CLTs for the estimators developed in Chapter II and quantifies their uncertainties using asymptotic CIs.

## 1.3 Cross-Entropy Method for Importance Sampling

The optimal IS distribution is practically not attainable because we need to know the explicit relationship between all possible simulation inputs and outputs a priori, whereas our essential assumption is that we can only learn the relationship incrementally by running the simulator. In practice, people use various approaches (*De Boer et al.*, 2005; *Dubourg et al.*, 2013) for approximating the optimal IS density and find them very effective in various application domains including structural reliability analysis (*Kurtz and Song*, 2013) and computational finance (*Wang and Zhou*, 2015).

A common approach is building a model of the simulation model, called metamodel or emulator, based on a small pilot sample of simulation data and using the metamodel to approximate the optimal IS density that focuses sampling efforts on important

input region (*Dubourg et al.*, 2013; *Choe et al.*, 2015). This metamodel-based approach works well in practice if we can build a good metamodel that captures the important region well. We however need an alternative approach if building a metamodel is difficult, for example, due to highly complicated simulation input-output relationship.

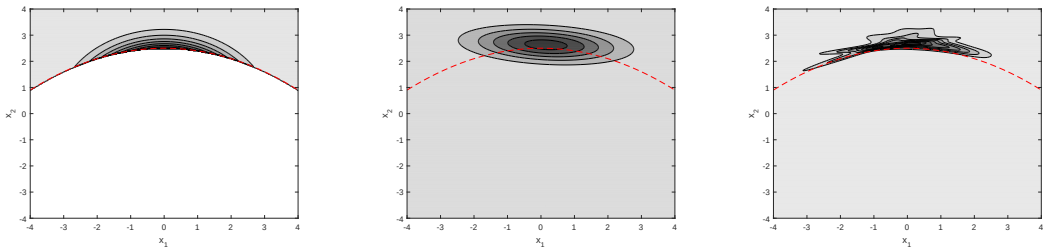
The cross-entropy (CE) method (*Rubinstein*, 1999) is a widely used alternative method in the literature. This method confines the candidate IS density to a parametric distribution family and updates the distribution parameter as we gather simulation data. The updating procedure aims to minimize the difference between the candidate IS density and the optimal IS density. The difference is measured in terms of CE, hence the name of the method.

Limiting the candidate density to a parametric distribution family makes the CE method convenient to use. Especially when the candidate density belongs to the natural exponential family (e.g., normal distribution with known variance, binomial distribution with known number of trials, Poisson distribution, etc.), updating the parameter of the candidate density reduces to evaluating analytical updating equations from numerically minimizing an estimate of the CE (*De Boer et al.*, 2005).

The convenience of using a parametric candidate density does not come without a price. If the optimal IS density takes a too complicated form to approximate with a parametric density, the standard CE method cannot achieve the full potential of using IS in terms of variance reduction and computational saving. For example, if the important input regions (or input conditions that lead to frequent failure events in reliability evaluation) are represented by two separate zones, a unimodal parametric density cannot exactly capture the important zones but may focus on either one of the two zones or diffuse the sampling efforts to cover both zones and the in-between area, which is not necessarily important for understanding the system reliability.

To overcome the rigidity of using a parametric density with a small number of parameters, recent studies propose using a mixture of multiple parametric densities

(*Botev et al.*, 2013; *Kurtz and Song*, 2013; *Wang and Zhou*, 2015). Yet, this approach encounters another problem of being potentially too flexible. Because the CE method minimizes an estimate of CE based on data at each updating step, allowing the candidate IS density to be too flexible tends to create data overfitting problem, which makes the IS density unstable. To illustrate, Figure 1.1 shows the 2-dimensional contour plots of (a) the optimal IS density (from the example with $b = 2.5$ in Section 4.4.1), (b) a Gaussian density from the standard CE method, and (c) the mixture of many Gaussian densities. We observe that a single Gaussian density generally captures the important region but fails to have the right parabolic shape. On the other hand, the mixture of many Gaussian densities overfits the data, having a too wiggly shape. It is an open problem in the literature to find the best number of component densities, which determines the flexibility of mixture density. Chapter IV devises a novel information criterion to find the best component number based on observed simulation data.



(a) Optimal IS Density (b) Single Parametric Density (c) Mixture of Many Parametric Densities

Figure 1.1: Comparison of the Optimal IS Density with Approximating Densities

# CHAPTER II

# Importance Sampling for Reliability Evaluation With Stochastic Simulation Models

## 2.1 Introduction

This study extends the theory of IS to estimate the reliability of systems using simulations. Various IS methods have been proposed for deterministic simulation models (*De Boer et al.*, 2005; *Cannamela et al.*, 2008; *Dubourg et al.*, 2013). However, conventional IS methods devised for deterministic simulation models are not applicable to stochastic simulation models (to be detailed in Section 2.2).

This study develops a new approach, which we call Stochastic Importance Sampling (SIS), that efficiently uses stochastic simulations with unknown output distribution. We propose two methods to estimate a failure probability. First, we use a failure probability estimator that allows multiple simulation replications at each input and derive the optimal IS density and allocation of simulation efforts at each sampled input for minimizing the estimator variance. Second, we propose another estimator that allows one simulation replication at each sampled input and derive the optimal IS density. Both methods use variance decomposition (*Kroese et al.*, 2011) to account for different sources of output variability and find the optimal IS densities using functional minimization (*Courant and Hilbert*, 1989). We demonstrate

the proposed methods using the NREL simulators to estimate wind turbine failure probabilities. The implementation results suggest that the SIS approach can produce estimates with smaller variances compared to alternative approaches when the total simulation budget is fixed.

## 2.2 Background and Literature Review

We first give an overview of IS for deterministic simulation models (DIS). Let $\mathbf{X}$, an input vector, denote a random vector following a known density, $f$. Given $\mathbf{X}$, a simulator generates an output, $Y = g(\mathbf{X})$, via a deterministic performance function, $g(\cdot)$. The function, $g(\cdot)$, is not explicitly known, but we can evaluate it with a simulation model. In reliability analysis with a deterministic simulation model, the failure probability is $P(Y > l) = E[\mathbb{I}(g(\mathbf{X}) > l)]$, where $l$ denotes the system's resistance level.

The CMC method is one of the simplest methods to estimate the failure probability. In CMC, we independently draw $\mathbf{X}_i$, $i = 1, 2, \cdots, N_T$, from its density, $f$, and unbiasedly estimate the failure probability by

$$\hat{P}_{CMC} = \frac{1}{N_T} \sum_{i=1}^{N_T} \mathbb{I}(g(\mathbf{X}_i) > l), \tag{2.1}$$

where $N_T$ is the total number of simulation replications.

Alternatively, DIS uses the following estimator,

$$\hat{P}_{DIS}(Y > l) = \frac{1}{N_T} \sum_{i=1}^{N_T} \mathbb{I}(g(\mathbf{X}_i) > l) \frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}, \tag{2.2}$$

where $\mathbf{X}_i, i = 1, 2, \cdots, N_T$, is independently sampled from $q$, called an IS density. Since $\mathbf{X}_i$ is sampled from $q$, we multiply the likelihood ratio, $f(\mathbf{X}_i)/q(\mathbf{X}_i)$, in (2.2) to obtain an unbiased estimator of $P(Y > l)$. Note that $\hat{P}_{DIS}$ in (2.2) is unbiased

under the condition that $q(\mathbf{x}) = 0$ implies that $\mathbb{I}(g(\mathbf{x}) > l) f(\mathbf{x}) = 0$ for any $\mathbf{x}$. An appropriately selected IS density reduces the estimator variance. It is well-known that the following IS density renders $Var[\hat{P}_{DIS}]$ zero (*Kroese et al.*, 2011):

$$q_{DIS}(\mathbf{x}) = \frac{\mathbb{I}(g(\mathbf{x}) > l) f(\mathbf{x})}{P(Y > l)}. \tag{2.3}$$

Here, $q_{DIS}(\mathbf{x})$ can be interpreted as the conditional density of $\mathbf{X}$, given that the failure event occurs. Since the denominator in (2.3) is the target quantity one wants to estimate and $\mathbb{I}(g(\mathbf{x}) > l)$ is unknown, $q_{DIS}(\mathbf{x})$ is not implementable in practice. Several approximations have been developed, including the cross-entropy method (*De Boer et al.*, 2005) and metamodel-based approximations (*Dubourg et al.*, 2013). These methods aim to find good IS densities that focus sampling efforts on the failure event region.

Existing IS studies consider the deterministic performance function, $g(\cdot)$. That is, for a fixed input, $\mathbf{x}$, the observed output, $Y = g(\mathbf{x})$, is always the same. This case corresponds to the simulation with a deterministic simulation model where the same input generates the same output. On the other hand, when a stochastic simulation model is used, the simulation output is random even at the same input. We can express the random output as $Y = g(\mathbf{X}, \boldsymbol{\epsilon})$, where $\boldsymbol{\epsilon}$ collectively denotes the uncontrollable randomness inside the simulator and $\mathbf{X}$ denotes a controllable random vector with its known density, $f$.

One might claim that in any simulations, both variables, $\mathbf{X}$ and $\boldsymbol{\epsilon}$, are controllable because some sampling distributions are specified for both variables in order to run the simulation. However, there are some cases where the DIS approach cannot be applied. First, to use DIS, the joint distribution function of $\mathbf{X}$ and $\boldsymbol{\epsilon}$, which needs to be biased in the IS method, should be explicitly defined. In many realistic simulations, the relationships among the elements of $\boldsymbol{\epsilon}$ (or between $\mathbf{X}$ and $\boldsymbol{\epsilon}$) are governed

by physical rules and constraints, and thus finding an explicit form of the joint distribution function can be intractable. Second, even if we know the joint distribution function of $\mathbf{X}$ and $\boldsymbol{\epsilon}$ explicitly, when the dimension of $\boldsymbol{\epsilon}$ is extremely high, applying DIS becomes very difficult due to the curse of dimensionality (*Au and Beck*, 2003). In addition, some third-party simulation software may not allow access and control for $\boldsymbol{\epsilon}$.

For example, with the specification we adopted from *Moriarty* (2008), the NREL simulators use over 8 million random variables for each simulation run to generate a three dimensional stochastic wind profile at multiple grid points via the inverse Fourier transform (*Jonkman*, 2009). The relationship of $\mathbf{X}$, the input wind condition, with $\boldsymbol{\epsilon}$, which collectively represents the 8 million *plus* random variables, is highly complicated due to the spatial and temporal dependence coupled with the inverse Fourier transform. Consequently, one cannot find the explicit joint distribution of $\mathbf{X}$ and $\boldsymbol{\epsilon}$. Even if one were to find it, applying the DIS approach jointly to $\mathbf{X}$ and $\boldsymbol{\epsilon}$ is virtually impossible due to the curse of dimensionality as previously mentioned. In fact, this difficulty is typical for many realistic simulations of actual stochastic systems with high degrees of freedom.

Therefore, for the stochastic simulation models where we effectively do not have control over $\boldsymbol{\epsilon}$, the DIS density in (2.3), $q_{DIS}$, can no longer be optimal. In fact, $q_{DIS}$ cannot be applied to the stochastic simulation model because given $\mathbf{x}$, $\mathbb{I}\left(g(\mathbf{x}) > l\right)$ in (2.3) is random.

Recently, stochastic simulation models that consider stochastic outputs given an input condition have also been studied in the literature (*Huang et al.*, 2006; *Ankenman et al.*, 2010). *Ankenman et al.* (2010) consider a queueing system simulation as an example of stochastic simulation models, where the arrival rate is the input, $x$, and the average number of customers in the system during specific time units, $T$, is an output, $Y$. Here, $\boldsymbol{\epsilon}$ collectively denotes the customer inter-arrival times and the service times.

*Huang et al.* (2006) also consider stochastic simulation models and use the inventory system simulation where the output, a total cost per month, is stochastic, given the input including a reorder point and a maximal holding quantity. Even though these studies account for the intrinsic uncertainty in outputs, their focuses are different from our study's. For example, *Ankenman et al.* (2010) develop stochastic simulation metamodeling, extending the kriging methodology (*Joseph*, 2006), and estimate an unknown quantity based on a metamodel. We note that this metamodeling-based approach is useful for estimating a mean response. However, this approach usually smooths a response function so that it loses its estimation accuracy in a tail probability estimation, as discussed in *Cannamela et al.* (2008).

Another well-known approach is "IS for stochastic simulations" which has been extensively studied (*Heidelberger*, 1995) after the seminal paper by *Glynn and Iglehart* (1989). This approach is proven effective if we can control stochastic processes inside a simulation. However, when a simulator involves complicated processes (e.g., wind turbine simulators), controlling these processes can be difficult, if not impossible. Therefore, our proposed approach treats a simulator as a black box model, and thus differs from the existing approach.

## 2.3 Methodology

This section devises optimal SIS methods for stochastic simulation models. We include the detailed derivations and proofs in Appendix A.

### 2.3.1 Failure probability estimators

A stochastic simulation model generates a random variable, $Y$, given a realization of the input, $\mathbf{X} \in \mathbb{R}^p$. In this context, the failure probability is

$$P(Y > l) = E_f\left[P(Y > l \mid \mathbf{X})\right] = \int_{\mathcal{X}_f} P(Y > l \mid \mathbf{X} = \mathbf{x})\, f(\mathbf{x})\, \mathrm{d}\mathbf{x}, \tag{2.4}$$

where $f$ is the density of $\mathbf{X}$ with the support of $\mathcal{X}_f$, and the subscript $f$ appended to the expectation operator in (2.4) indicates that the expectation is taken with respect to $f$. We call an estimator of $P(Y > l)$, $\hat{P}(Y > l)$, a probability of exceedance (POE) estimator.

A simple Monte Carlo estimator for $P(Y > l)$ in (2.4) is

$$\hat{P}_{MC} = \frac{1}{M}\sum_{i=1}^{M} \hat{P}(Y > l \mid \mathbf{X}_i) = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N_i}\sum_{j=1}^{N_i} \mathbb{I}\left(Y_j^{(i)} > l\right)\right), \tag{2.5}$$

where $\mathbf{X}_i$, $i = 1, 2, \cdots, M$, is independently sampled from $f$. The number of sampled inputs, $M$, is called an input sample size. At each $\mathbf{X}_i$, we run simulations $N_i$ times to obtain $N_i$ outputs, $Y_j^{(i)}$, $j = 1, 2, \cdots, N_i$, where $Y_j^{(i)}$ denotes the output obtained in the $j^{\text{th}}$ replication. Note that the estimator in (2.5) allows multiple replications at each $\mathbf{X}_i$ to account for the stochastic outputs at the same input. We call the number of simulation replications at each $\mathbf{X}_i$, $N_i$, an allocation size. In (2.5), we call $\hat{P}(Y > l \mid \mathbf{X}_i)$ a conditional POE estimator. The total number of replications is $N_T = \sum_{i=1}^{M} N_i$. With deterministic simulation models, multiple replications at the same input are not necessary because the outcome is conclusively determined at the given input.

In the spirit of IS methods, we propose the following SIS estimator:

$$\hat{P}_{SIS1} = \frac{1}{M}\sum_{i=1}^{M} \hat{P}(Y > l \mid \mathbf{X}_i)\,\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{I}\left(Y_j^{(i)} > l\right)\right)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}, \quad (2.6)$$

where $\mathbf{X}_i$ is drawn from $q$. $\hat{P}_{SIS1}$ is unbiased if $q(\mathbf{x}) = 0$ implies $\hat{P}(Y > l|\mathbf{X} = \mathbf{x})\, f(\mathbf{x}) = 0$ for any $\mathbf{x}$. We assume that the total simulation budget, $N_T$, and the input sample size, $M$, are given. Note that since we treat the stochastic elements inside the simulator as an uncontrollable input, we apply the underlying idea of IS only to $X$ and use the sample mean to estimate the conditional POE. Here, the conditional POE can be viewed as the success probability parameter in the binomial distribution, and the sample mean is the unique uniformly minimum-variance unbiased estimator for the binomial distribution (*Casella and Berger*, 2002).

We also propose an alternative estimator that restricts $N_i$ to be one:

$$\hat{P}_{SIS2} = \frac{1}{N_T}\sum_{i=1}^{N_T}\mathbb{I}\left(Y_i > l\right)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}, \quad (2.7)$$

where $Y_i$ is an output at $\mathbf{X}_i$, $i = 1, 2, \cdots, N_T$. $\hat{P}_{SIS2}$ is also an unbiased estimator of $P(Y > l)$ if $q(\mathbf{x}) = 0$ implies $\mathbb{I}\left(Y > l\right) f(\mathbf{x}) = 0$ for any $\mathbf{x}$. In the sequel, Sections 2.3.2 and 2.3.3 derive the minimum-variance solutions for the estimators in (2.6) and (2.7), respectively.

### 2.3.2 Stochastic Importance Sampling Method 1

We want to find the optimal allocation sizes and the optimal IS density that minimize the variance of the failure probability estimator in (2.6). Considering the two sources of randomness, i.e., stochastic inputs and stochastic elements inside the stochastic simulation model, we decompose the estimator variance into two compo-

nents as

$$Var\left[\hat{P}_{SIS1}\right] = Var\left[\frac{1}{M}\sum_{i=1}^{M}\hat{P}(Y>l \mid \mathbf{X}_i)\,\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}\right]$$
$$= \frac{1}{M^2}E_q\left[Var\left[\sum_{i=1}^{M}\hat{P}(Y>l \mid \mathbf{X}_i)\,\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}\;\middle|\;\mathbf{X}_1,\cdots,\mathbf{X}_M\right]\right]$$
$$+ \frac{1}{M^2}Var_q\left[E\left[\sum_{i=1}^{M}\hat{P}(Y>l \mid \mathbf{X}_i)\,\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}\;\middle|\;\mathbf{X}_1,\cdots,\mathbf{X}_M\right]\right]. \quad (2.8)$$

Let $s(\mathbf{X})$ denote the conditional POE, $P(Y>l \mid \mathbf{X})$. Using the fact that $\mathbf{X}_i \overset{i.i.d}{\sim} q$ for $i = 1, 2, \cdots, M$, we simplify $Var\left[\hat{P}_{SIS1}\right]$ in (2.8) to

$$Var\left[\hat{P}_{SIS1}\right] = \frac{1}{M^2}E_q\left[\sum_{i=1}^{M}\frac{1}{N_i}s(\mathbf{X}_i)\,(1-s(\mathbf{X}_i))\frac{f(\mathbf{X}_i)^2}{q(\mathbf{X}_i)^2}\right] + \frac{1}{M}Var_q\left[s(\mathbf{X})\,\frac{f(\mathbf{X})}{q(\mathbf{X})}\right]. \quad (2.9)$$

To find the optimal allocation size and the optimal IS density function, we first profile out $N_i$ and express the variance in (2.9) in terms of $q(\mathbf{X})$. Lemma II.1 presents the optimal assignment of simulation replications to each $\mathbf{X}_i$ for any given $q$.

**Lemma II.1.** *Given $q$, the variance in (2.9) is minimized if and only if*

$$N_i = \frac{\sqrt{s(\mathbf{X}_i)\,(1-s(\mathbf{X}_i))}f(\mathbf{X}_i)\,/q(\mathbf{X}_i)}{\sum_{j=1}^{M}\sqrt{s(\mathbf{X}_j)\,(1-s(\mathbf{X}_j))}f(\mathbf{X}_j)\,/q(\mathbf{X}_j)}\cdot N_T \quad \textit{for } i = 1, 2, \cdots, M. \quad (2.10)$$

Next, we use the optimal allocation size in Lemma 1 to derive the optimal IS density for the estimator in (2.6). Plugging the $N_i$'s in (2.10) into the estimator variance in (2.9) gives

$$Var\left[\hat{P}_{SIS1}\right] = \frac{1}{M}\frac{1}{N_T}\left(E_f\left[s(\mathbf{X})\,(1-s(\mathbf{X}))\frac{f(\mathbf{X})}{q(\mathbf{X})}\right] + (M-1)\left(E_f\left[\sqrt{s(\mathbf{X})\,(1-s(\mathbf{X}))}\right]\right)^2\right)$$
$$+ \frac{1}{M}\left(E_f\left[s(\mathbf{X})^2\,\frac{f(\mathbf{X})}{q(\mathbf{X})}\right] - P\,(Y>l)^2\right). \quad (2.11)$$

We minimize the functional in (2.11) using the principles of the calculus of variations (*Courant and Hilbert*, 1989) and find the optimal IS density, $q_{SIS1}$. We also plug $q_{SIS1}$ into (2.10) to attain the optimal allocation size, which leads to Theorem II.2.

**Theorem II.2.** *(a) The variance of the estimator in* (2.6) *is minimized if the following IS density and the allocation size are used.*

$$q_{SIS1}(\mathbf{x}) = \frac{1}{C_{q1}} f(\mathbf{x}) \sqrt{\frac{1}{N_T} s(\mathbf{x}) \left(1 - s(\mathbf{x})\right) + s(\mathbf{x})^2}, \tag{2.12}$$

$$N_i = N_T \frac{\sqrt{\frac{N_T(1-s(\mathbf{x}_i))}{1+(N_T-1)s(\mathbf{x}_i)}}}{\sum_{j=1}^{M} \sqrt{\frac{N_T(1-s(\mathbf{x}_j))}{1+(N_T-1)s(\mathbf{x}_j)}}}, \quad i = 1, 2, \cdots, M, \tag{2.13}$$

*where* $C_{q1}$ *is* $\int_{\mathcal{X}_f} f(\mathbf{x}) \sqrt{\frac{1}{N_T} s(\mathbf{x}) \cdot (1 - s(\mathbf{x})) + s(\mathbf{x})^2}\, \mathrm{d}\mathbf{x}$ *and* $s(\mathbf{x})$ *is* $P(Y > l | \mathbf{X} = \mathbf{x})$. *(b) With* $q_{SIS1}$ *and* $N_i, i = 1, 2, \cdots, M$, *the estimator in* (2.6) *is unbiased.*

We call this approach Stochastic Importance Sampling Method 1 (SIS1). The optimal SIS1 density in (2.12) focuses its sampling efforts on the region where the failure event of interest likely occurs. On the other hand, the input condition, $\mathbf{x}_i$, with a smaller $s(\mathbf{x}_i)$ needs a larger accompanying $N_i$. In other words, among the important input conditions under which a system can possibly fail (that is, the conditions that $q_{SIS1}$ samples), SIS1 balances the simulation efforts by allocating a larger (smaller) number of replications in the area with a relatively small (large) $s(\mathbf{x})$.

We note that when applied to a deterministic simulation model, the proposed SIS1 method reduces to the DIS method with $q_{DIS}$ in (2.3). Using $s(\mathbf{x}) = \mathbb{I}\left(Y > l\right), \forall \mathbf{x} \in \mathcal{X}_f$, in a deterministic simulation model where $Y = g(\mathbf{x})$ is the deterministic output of the simulator at an input, $\mathbf{x}$, we can see that $q_{SIS1}$ in (2.12) is reduced to $q_{DIS}$. Also, when $s(\mathbf{x})$ is an indicator function, the first term in the variance in (2.9) vanishes, implying that we do not need the allocation step for SIS1 as we do not for DIS.

### 2.3.3 Stochastic Importance Sampling Method 2

This section derives the optimal IS density minimizing the variance of the failure probability estimator in (2.7), which restricts the allocation size to be one at each sampled input. This approach does not require the allocation of $N_i$. The estimator in (2.7) essentially takes a similar form in (2.2) used for a deterministic simulation model. However, it is not possible to use $q_{DIS}$ in (2.3) for stochastic simulation models since $Y$ is not a deterministic function of $\mathbf{X}$. Theorem II.3 presents the optimal IS density for the estimator in (2.7) with a stochastic simulation model.

**Theorem II.3.** *(a) The variance of the estimator in (2.7) is minimized with the density,*

$$q_{SIS2}(\mathbf{x}) = \frac{1}{C_{q2}} \sqrt{s(\mathbf{x})} f(\mathbf{x}) \,, \tag{2.14}$$

*where* $C_{q2}$ *is* $\int_{\mathcal{X}_f} \sqrt{s(\mathbf{x})} f(\mathbf{x}) \, \mathrm{d}\mathbf{x}$ *and* $s(\mathbf{x})$ *is* $P(Y > l | \mathbf{X} = \mathbf{x})$.
*(b) With* $q_{SIS2}$*, the estimator in (2.7) is unbiased.*

We call this approach Stochastic Importance Sampling Method 2 (SIS2). Similar to SIS1, SIS2 focuses its sampling efforts on the input conditions under which the failure event likely occurs with a high probability, $s(\mathbf{x})$. Also, when applied to deterministic simulation models, $q_{SIS2}$ in (2.14) is reduced to $q_{DIS}$ in (2.3).

### 2.3.4 Implementation guidelines

In implementing SIS1, we use rounded $N_i$. If the rounding yields zero, we use one to ensure the unbiasedness. Note that $q_{SIS1}$, $N_i$'s and $q_{SIS2}$ require the conditional POE, $s(\mathbf{x})$, which is unknown. Therefore, the optimal solutions in (2.12)-(2.14) are theoretically optimal, but not implementable, which is a common problem encountered in any IS methods. In our implementation, we approximate the conditional

POE using a parametric regression model (or metamodel). The estimators in (2.6) and (2.7) are still unbiased with this approximation.

We can consider several methods to approximate the conditional POE. In many studies, Gaussian regression or its variants have been used to approximate the simulation model (*Seber and Lee*, 2003; *Cannamela et al.*, 2008; *Ankenman et al.*, 2010). In particular, when the output, $Y$, is the average of the quantities generated from a stochastic process or system, Gaussian regression or its variants would provide good approximation. More generally, when $Y$ tends to follow a distribution in the exponential family, generalized linear model (GLM)(*Green and Silverman*, 1994) or generalized additive model (GAM) (*Hastie and Tibshirani*, 1990) could be employed. When the distribution belongs to a non-exponential family, generalized additive model for location, scale and shape (GAMLSS) (*Rigby and Stasinopoulos*, 2005) will provide a flexible modeling framework. For example, if $Y$ is the maximum or minimum of the quantities during a specific time interval (e.g., maximum stress during 10-minute operations), the Generalized Extreme Value (GEV) distribution (*Coles*, 2001) can be employed for fitting the conditional distribution with the GAMLSS framework (to be detailed in Section 2.6).

While general regression models focus on capturing input-to-output relationships and are relatively straightforward to check the model accuracy, determining the metamodel accuracy for conditional POE imposes more challenges because not only is the regression relationship important, but selecting the appropriate distribution is also crucial. If the distribution fitting is not carefully conducted, the approximated POE might not help achieving the full potential of the proposed method. Provided that the primary purpose of the metamodel is to approximate the conditional POE, we recommend using goodness-of-fit tests for checking the metamodel accuracy (*Stephens*, 1974). Different tests have their own pros and cons depending on the hypothesized distribution; thus, it is advisable to decide on the specific test based on the distri-

bution of interest. Extensive studies have been conducted on the tests for specific distributions (e.g., *Choulakian and Stephens*, 2001).

We summarize SIS1 and SIS2 procedures as follows:

Step 1. Approximate the conditional POE, $s(\mathbf{x})$, with a metamodel.

Step 2. Sample $\mathbf{x}_i$, $i = 1, \cdots, M$, from $q_{SIS1}$ in (2.12) for SIS1 or $q_{SIS2}$ in (2.14) for SIS2 (Note that $M = N_T$ for SIS2).

Step 3. Determine the allocation size, $N_i$ for each $\mathbf{x}_i$, $i = 1, \cdots, M$, using (2.13) for SIS1 or set $N_i = 1$, $i = 1, \cdots, M$, for SIS2.

Step 4. Run simulation $N_i$ times at each $\mathbf{x}_i$, $i = 1, \cdots, M$.

Step 5. Estimate the failure probability using (2.6) for SIS1 or (2.7) for SIS2.

## 2.4 Benchmark Methods

We compare our two methods, SIS1 and SIS2, with two benchmark methods. First, we use the CMC estimator in (2.1), which is an unbiased estimator of the failure probability even if the simulation model is stochastic. The variance is known as $P(Y > l)(1 - P(Y > l))/N_T$.

Second, we introduce a new IS density, $q_{BIS}$, that mimics $q_{DIS}$ in (2.3). Recalling that it is not possible to use the IS density in (2.3) for stochastic simulation models, we simply replace the failure indicator function in (2.3), $\mathbb{I}(Y > l)$, with the conditional POE, $s(\mathbf{x})$, to obtain

$$q_{BIS}(\mathbf{x}) = \frac{s(\mathbf{x})f(\mathbf{x})}{P(Y > l)}. \tag{2.15}$$

With $q_{BIS}(\mathbf{x})$, we use the failure probability estimator in (2.7). We call this approach Benchmark Importance Sampling (BIS), since it emulates DIS.

## 2.5 Numerical Examples

We investigate the performances of the SIS methods using numerical examples with various settings. We take a deterministic simulation example in *Cannamela et al.* (2008) and modify it to have stochastic elements. Specifically, we use the following data generating structure:

$$X \sim N(0,1)\,, \qquad Y|X \sim N\big(\mu(X)\,, \sigma^2(X)\big)\,, \tag{2.16}$$

where the mean, $\mu(X)$, and the standard deviation, $\sigma(X)$, of the normal distribution are

$$\begin{aligned} \mu(X) &= 0.95\delta X^2\left(1 + 0.5\cos(5X) + 0.5\cos(10X)\right), \\ \sigma(X) &= 1 + 0.7\,|X| + 0.4\cos(X) + 0.3\cos(14X). \end{aligned} \tag{2.17}$$

In practice, we do not know the conditional distribution for $Y|X$; thus, as a metamodel, we use the normal distribution with the following mean and standard deviation:

$$\begin{aligned} \hat{\mu}(X) &= 0.95\delta X^2\left(1 + 0.5\rho\cos(5X) + 0.5\rho\cos(10X)\right), \\ \hat{\sigma}(X) &= 1 + 0.7\,|X| + 0.4\rho\cos(X) + 0.3\rho\cos(14X). \end{aligned} \tag{2.18}$$

Here, we include the parameters $\delta$ and $\rho$ to control the similarity of the IS density to the original input density and the metamodeling accuracy, respectively. We set $N_T = 1,000$ (with $M = 300$ for SIS1) and repeat the experiment 500 times to obtain the sample average and the standard error of each method's POE estimation. We use the following setup as a baseline and vary each parameter to see its effect on the performances of the proposed methods: $P_T = 0.01$, $\delta = 1$, and $\rho = 1$. We explain

each parameter and summarize the experiment results as follows:

- $P_T$, the magnitude of target failure probability: We study how the proposed methods perform at different levels of $P_T = P(Y > l)$. The computational efficiency of each method is evaluated using the standard error or equivalently the relative ratio, $N_T/N_T^{(CMC)}$, where $N_T^{(CMC)}$ is the number of CMC simulation replications needed to achieve the same standard error of each method. Table 2.1 suggests that the computational gains of SIS1 and SIS2 against CMC generally increase as $P_T$ gets smaller. Also, SIS1 and SIS2 always outperform BIS, providing more accurate estimates with lower standard errors.

Table 2.1: POE estimation results with different $\delta$ and $P_T$ ($\rho = 1$)

| | | $\delta = 1$ | | | $\delta = -1$ | | |
|---|---|---|---|---|---|---|---|
| | | $P_T$ | | | $P_T$ | | |
| | | 0.10 | 0.05 | 0.01 | 0.10 | 0.05 | 0.01 |
| SIS1 | Sample Average | 0.1004 | 0.0502 | 0.0100 | 0.1001 | 0.0500 | 0.0100 |
| | Standard Error | 0.0068 | 0.0039 | 0.0005 | 0.0090 | 0.0062 | 0.0026 |
| | Relative Ratio | 51% | 32% | 2.5% | 90% | 81% | 68% |
| SIS2 | Sample Average | 0.0999 | 0.0501 | 0.0100 | 0.1001 | 0.0500 | 0.0099 |
| | Standard Error | 0.0069 | 0.0042 | 0.0006 | 0.0086 | 0.0064 | 0.0028 |
| | Relative Ratio | 53% | 37% | 3.6% | 82% | 86% | 79% |
| BIS | Sample Average | 0.1002 | 0.0505 | 0.0101 | 0.1009 | 0.0503 | 0.0101 |
| | Standard Error | 0.0089 | 0.0068 | 0.0014 | 0.0095 | 0.0067 | 0.0031 |
| | Relative Ratio | 88% | 97% | 20% | 100% | 95% | 97% |
| CMC | Sample Average | 0.1005 | 0.0506 | 0.0100 | 0.1005 | 0.0498 | 0.0100 |
| | Standard Error | 0.0092 | 0.0070 | 0.0030 | 0.0096 | 0.0071 | 0.0031 |

*Note:* The 'Relative Ratio' is $N_T/N_T^{(CMC)}$, where $N_T^{(CMC)} = P_T(1 - P_T)/(S.E.)^2$. $S.E.$ denotes the standard error.

- $\delta$, the difference between the original input density and the optimal IS density: We consider $\delta$ of 1 or $-1$. The densities, $f$ and $q_{SIS1}$ (or $q_{SIS2}$), are more different from each other when $\delta = 1$ than when $\delta = -1$. Table 2.1 suggests that the computational gains of SIS1 and SIS2 are much more significant when $\delta = 1$ than when $\delta = -1$. Interestingly, when $\delta = -1$, BIS shows no advantage over CMC, whereas the proposed methods still lead to lower standard errors

than CMC.

- $\rho$, the metamodeling accuracy: We vary $\rho$ in $\hat{\mu}(X)$ and $\hat{\sigma}(X)$ in (2.18) to control the quality of the metamodel. Table 2.2 shows that the standard errors of all IS estimators increase as $\rho$ decreases. However, the standard errors of both SIS1 and SIS2 increase more slowly than BIS. The fact that the increment of the SIS2's standard error is minimal indicates that SIS2 is less sensitive to the metamodel quality than SIS1. The performance of BIS differs significantly depending on the metamodel quality, and BIS generates an even higher standard error than CMC when $\rho = 0$.

Table 2.2: POE estimation results with different $\rho$ ($\delta = 1$)

| | | $\rho$ | | |
|---|---|---|---|---|
| | | 1.00 | 0.50 | 0 |
| SIS1 | Sample Average | 0.0100 | 0.0100 | 0.0101 |
| | Standard Error | 0.0005 | 0.0008 | 0.0017 |
| SIS2 | Sample Average | 0.0100 | 0.0101 | 0.0100 |
| | Standard Error | 0.0006 | 0.0007 | 0.0010 |
| BIS | Sample Average | 0.0101 | 0.0100 | 0.0102 |
| | Standard Error | 0.0014 | 0.0018 | 0.0063 |
| CMC | Sample Average | 0.0099 | 0.0099 | 0.0099 |
| | Standard Error | 0.0030 | 0.0030 | 0.0030 |

Next, we investigate the impact of the variation of the randomness inside simulations. In Section 2.3, we noted that SIS1 and SIS2 are reduced to DIS when they are applied to a deterministic simulation model. Thus, we expect that if the uncontrollable randomness represented by $\epsilon$ has a small level of variation, the standard errors of SIS1 and SIS2 will be close to zero. To illustrate, we consider the same data generating structure in (2.16) and (2.17), but with a constant variance, $\sigma^2(X) = \tau^2$. We use the optimal IS densities for SIS1 and SIS2 in simulations. Table 2.3 shows that as $\tau$ gets close to zero, so do the standard errors of SIS1 and SIS2. That is, the proposed methods practically reduce to DIS.

Table 2.3: POE estimation results with different $\tau$ ($\delta = 1$)

| | | $\tau$ | | | | |
|---|---|---|---|---|---|---|
| | | 0.50 | 1.00 | 2.00 | 4.00 | 8.00 |
| SIS1 | Sample Average | 0.0102 | 0.0101 | 0.0101 | 0.0102 | 0.0100 |
| | Standard Error | 0.0001 | 0.0001 | 0.0005 | 0.0021 | 0.0028 |
| SIS2 | Sample Average | 0.0102 | 0.0101 | 0.0101 | 0.0104 | 0.0100 |
| | Standard Error | 0.0001 | 0.0002 | 0.0006 | 0.0023 | 0.0028 |

*Note:* SIS1's standard errors for $\tau = 0.50$ and $\tau = 1.00$ are .00007 and .00013, respectively, in more digits.

We conduct additional experiments with other parameter settings, which are detailed in Appendix A: (a) experiment results with different $M/N_T$ ratios suggest that the standard error of the SIS1 estimator is generally insensitive to the choice of $M/N_T$ ratio; (b) in investigating the effects of the metamodeling inaccuracy for the global pattern and different locality levels of $\mu(X)$, we do not find any clear patterns for this specific example. We also devise numerical examples with a multivariate input vector and observe the similar patterns discussed above (detailed in Appendix A).

In summary, SIS1 and SIS2 always outperform BIS and CMC in various settings. We obtain remarkable improvements of computational efficiency when the original input density and SIS1 (or SIS2) density are different. Also, as the target failure probability gets smaller, the efficiencies of SIS1 and SIS2 increase. Overall, SIS1 yields smaller standard errors than SIS2 in most cases. However, when it is difficult to build a good-quality metamodel (e.g., due to complex response surface over the input space), SIS2 would provide robust estimations because it is less sensitive to the metamodel quality.

## 2.6 Implementation With Wind Turbine Simulators

We implement the proposed approach to evaluate the reliability of a wind turbine operated in dynamic wind conditions (*Byon et al.*, 2010), using the NREL simulators. Implementation details are provided in Appendix A.

### 2.6.1 Description of NREL simulations

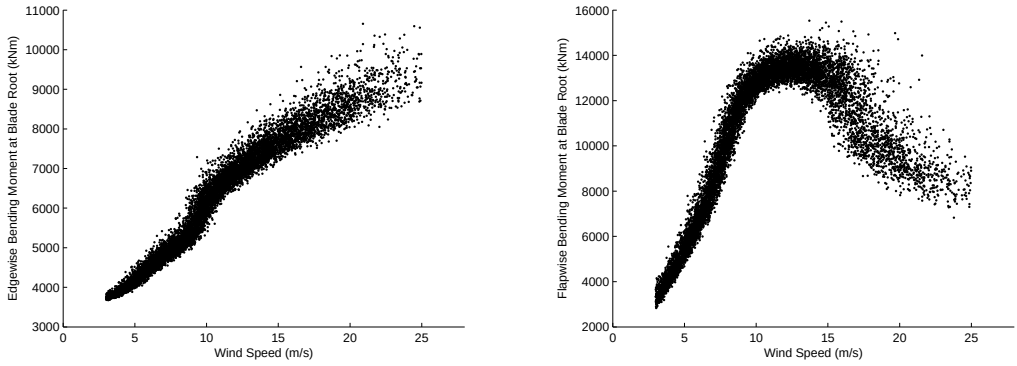
Following wind industry practice and the international standard, IEC 61400-1 (*International Electrotechnical Commission*, 2005), we use a 10-minute average wind speed as an input, $X$, to the NREL simulators. As the density of $X$, $f$, we use a Rayleigh density with a truncated support, following *Moriarty* (2008).

Given a 10-minute average wind speed, $X$, the NREL simulators, including TurbSim (*Jonkman*, 2009) and FAST (*Jonkman and Buhl Jr.*, 2005), simulate the turbine's 10-minute operations. We study two load response types, edgewise and flapwise bending moments at a blade root, as they are of great concern in ensuring a wind turbine's structural reliability. We calculate both load responses based on the equations in *Moriarty* (2008, p.564) using the in-plane and out-of-plane bending moments generated by FAST. Among the 10-minute load responses, we take the maximum response of a load type as an output variable, $Y$. Hereafter, a simulation replication denotes the 10-minute simulation which generates a 10-minute maximum load (hereafter, a load, or response), given a 10-minute average wind speed (hereafter, a wind speed).

Figure 2.1 shows the load outputs in a range of wind conditions. High wind speed tends to cause large edgewise moments, which are dominated by gravity loading. Flapwise moments depend on the pitch regulation (*Moriarty*, 2008; *Yampikulsakul et al.*, 2014) that controls the blade pitch angles to reduce the loading on the blades when the wind speed is higher than the rated speed (11.5 m/s in Figure 2.1(b)).

### 2.6.2 Approximation of POE with a metamodel

To implement SIS1, SIS2 and BIS, we need the conditional POE, $s(x)$, which is unknown in practice. We approximate it using a parametric regression model. *Lee et al.* (2013) model the load responses in wind turbine field data using a nonhomogeneous GEV distribution. We apply a similar procedure for approximating the conditional POE.




(a) Edgewise bending moment (b) Flapwise bending moment

Figure 2.1: Load outputs from the NREL simulators

To begin, we obtain a pilot sample of NREL simulations to build the metamodel. The pilot sample consists of 600 observations of $(X, Y)$ pairs, where $X$ is the wind speed uniformly sampled between 3 m/s and 25 m/s, and $Y$ is the corresponding load response from the NREL simulators. In the metamodel, we use a nonhomogeneous GEV distribution to approximate the conditional distribution of $Y|X = x$ and express the location and scale parameters as functions of wind speeds as in *Lee et al.* (2013). We also considered other parametric distributions including Weibull, Gamma, and lognormal distributions. However, GEV provides the best fit for our chosen load response types. The cumulative distribution function of GEV is expressed as follows, with the location parameter function, $\mu(x)$, the scale parameter function, $\sigma(x)$, and the shape parameter, $\xi$.

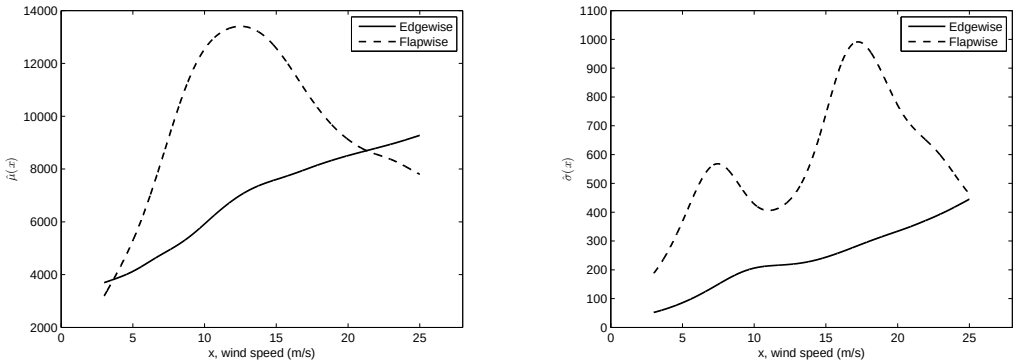
$$P(Y \leq y \mid X = x) = \begin{cases} \exp\left(-\left(1+\xi\left(\frac{y-\mu(x)}{\sigma(x)}\right)\right)^{-1/\xi}\right) & \text{for } \xi \neq 0 \\ \exp\left(-\exp\left(-\frac{y-\mu(x)}{\sigma(x)}\right)\right) & \text{for } \xi = 0\ . \end{cases}$$

We model the location and scale parameter functions with cubic smoothing spline functions. For the shape parameter, we use a constant, $\xi$, to avoid an overly complicated model as suggested in *Lee et al.* (2013). To estimate the spline function parameters and the shape parameter, we use the GAMLSS framework (*Rigby and*

*Stasinopoulos*, 2005). Specifically, we maximize the log-likelihood function penalized by the roughness of $\mu(x)$ and $\log \sigma(x)$ for fixed smoothing parameters, $\lambda_\mu$ and $\lambda_\sigma$:

$$\max \mathcal{L}_p = \mathcal{L} - \lambda_\mu \int_{-\infty}^{\infty} \mu''(x)^2 \, \mathrm{d}x - \lambda_\sigma \int_{-\infty}^{\infty} ((\log \sigma)''(x))^2 \, \mathrm{d}x,$$

where $\mathcal{L}$ is the log-likelihood function of the pilot data, $(X_i, Y_i), i = 1, 2, \cdots, 600$. The roughness penalties based on the second derivatives are commonly emplopyed in the literature (*Hastie and Tibshirani*, 1990; *Green and Silverman*, 1994). We find the smoothing parameters, $\lambda_\mu$ and $\lambda_\sigma$, that minimize the Bayesian information criterion (BIC) as suggested in *Rigby and Stasinopoulos* (2005). Figures 2.2(a) and (b) present the estimated location and scale parameter functions, $\hat{\mu}(x)$ and $\hat{\sigma}(x)$, respectively. The estimated shape parameters, $\hat{\xi}$, are -0.0359 and -0.0529 for the edgewise and flapwise moments, respectively.




(a) Estimated location parameter function, (b) Estimated scale parameter function, $\hat{\sigma}(x)$ $\hat{\mu}(x)$

Figure 2.2: Estimated parameter functions for edgewise and flapwise moments

Next, we conduct the Kolmogorov-Smirnov (KS) test to see the goodness-of-fit of the GEV distribution. We standardize the output, using the estimated location and scale functions shown in Figure 2.2, and perform the KS test on the standardized loads, $Z_i, i = 1, 2, \cdots, 600$, with the null hypothesis, $H_0 : Z \sim GEV(\mu = 0, \sigma = 1, \hat{\xi})$. The test results support the use of GEV distribution for the edgewise and flapwise

moments with the $p$-values of 0.716 and 0.818, respectively. In Appendix A, we include additional tests at important wind speeds, which also support the use of GEV distribution.

### 2.6.3 Sampling from IS densities

To avoid difficulties in drawing samples from the IS densities whose normalizing constants are unknown, we use the following acceptance-rejection algorithm (*Kroese et al.*, 2011).

**Acceptance-rejection algorithm**

Step 1: Sample $\mathbf{x}$ from the input distribution, $f$.

Step 2: Sample $u$ from the uniform distribution over the interval, $(0, f(\mathbf{x}))$.

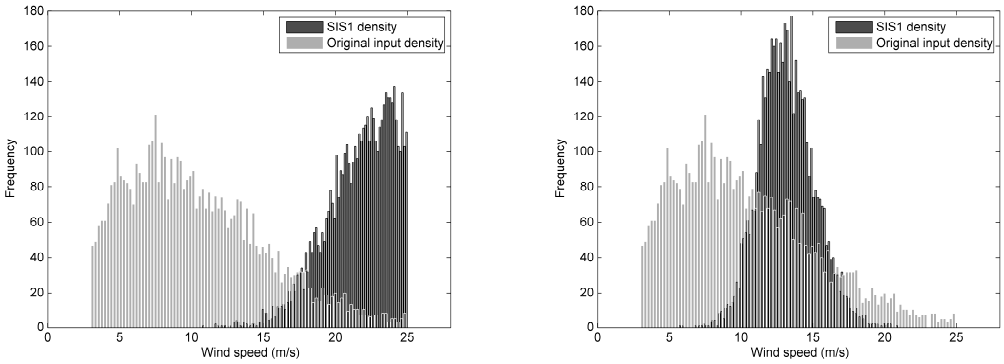
Step 3: If $u \leq C_q \cdot q(\mathbf{x})$, return $\mathbf{x}$; otherwise, repeat from Step 1.

Here, $C_q$ denotes the normalizing constant corresponding to the IS density, i.e., $C_{q1}$ for SIS1, $C_{q2}$ for SIS2, and $P(Y > l)$ for BIS. Note that $C_q \cdot q(\mathbf{x})$ only involves $f(\mathbf{x})$ and $s(\mathbf{x})$. Thus, without a knowledge of $C_q$, this algorithm returns $\mathbf{x}$, which follows the target IS density, $q$. This algorithm exactly samples from $q$ when the inequality condition, $f(\mathbf{x}) \geq C_q \cdot q(\mathbf{x})$, $\forall \mathbf{x} \in \mathcal{X}_f$ is satisfied. The IS densities, $q_{SIS1}$, $q_{SIS2}$, and $q_{BIS}$, satisfy this inequality condition. The acceptance rate of the algorithm is equal to $C_q$ (*Kroese et al.*, 2011).

The acceptance-rejection method has several advantages. First, this method keeps the unbiasedness of the estimator because of its independent and exact sampling nature. Second, we can always use the original input distribution, $f$, as an auxiliary distribution. However, we can also use other sampling methods such as Markov chain Monte Carlo (MCMC). MCMC method can be useful if the input, $\mathbf{X}$, is high dimensional (*Kroese et al.*, 2011). The choice of sampling method is flexible in implementing

SIS1 and SIS2. In practice, the computational cost of the sampling would be insignificant; e.g., sampling thousands of inputs from the IS densities is a matter of seconds, whereas thousands of the NREL simulation replications can take days.

Figure 2.3 shows the empirical SIS1 density using the sampled wind speeds from the acceptance-rejection algorithm. In Figure 2.3(a), compared to the original input density, the SIS1 density for the edgewise moments has a higher mass at high wind speeds where high loads likely occur and high load variability is observed (see Figure 2.1(a)). Similarly, the SIS1 density for the flapwise moments in Figure 2.3(b) centers around the rated speed, 11.5 m/s, where high loads and variability are observed (see Figure 2.1(b)). Using the same acceptance-rejection algorithm, we also draw wind speeds from the SIS2 and BIS densities.




(a) Edgewise moments with $l$ =9,300 $kNm$ (b) Flapwise moments with $l$ =14,300 $kNm$

Figure 2.3: Comparison of empirical densities: original input density, $f$, versus SIS1 density, $q_{SIS1}$

Even though we sample inputs from the IS densities without knowing the value of the normalizing constant, $C_q$, we still need to compute $C_q$ for estimating the failure probability because the likelihood ratio in the IS estimators, $f(\mathbf{X})/q(\mathbf{X})$, need to be evaluated to ensure the unbiasedness of the estimators. This issue has been studied in the literature (*Hesterberg*, 1995). In this study, we employ a numerical integration to compute $C_q$ since a state-of-the-art numerical integration leads to an accurate evaluation of $C_q$ (*Shampine*, 2008). Our numerical studies in Appendix A also show

that the numerical integration does not affect the POE estimation accuracy.

### 2.6.4 Sensitivity analysis with different $M$ in SIS1

Recall that in SIS1, we derived the optimal SIS1 density, $q_{SIS1}$, and the optimal allocation size, $N_i$, $i = 1, 2, \cdots, M$, for a given input sample size, $M$, and a total computational resource, $N_T$. To see the effect of the ratio of $M$ to $N_T$ on POE estimation, we consider the four ratios of $M$ to $N_T$, 10%, 30%, 50%, and 80%. Table 2.4 summarizes the sample average and standard error based on 50 POE estimates. We also obtain the 95% CI of the standard error by using the bootstrap percentile interval (*Efron and Tibshirani*, 1993). Overall, the standard errors are comparable among different ratios.

Similar results are also observed in the extensive numerical studies where we have tested 10%, 30%, 50%, 70%, and 90% of $M/N_T$ ratios for the univariate and multivariate examples (see Appendix A). All of these results indicate that the estimation accuracy is not sensitive to the size of $M$, given $N_T$. In the subsequent implementations, we use the ratio of 10% and 30% for the edgewise and flapwise bending moments, respectively.

Table 2.4: Failure probability estimation by SIS1 method with different ratios of $M$ to $N_T$

| $M/N_T$ | Edgewise ($l$ = 8,600 *kNm*, $N_T$ = 1,000) | | Flapwise ($l$ = 13,800 *kNm*, $N_T$ = 2,000) | |
|---|---|---|---|---|
| | Sample Average | Standard Error (95% bootstrap CI) | Sample Average | Standard Error (95% bootstrap CI) |
| 10% | 0.0486 | 0.0016 (0.0012, 0.0020) | 0.0523 | 0.0034 (0.0026, 0.0041) |
| 30% | 0.0486 | 0.0018 (0.0014, 0.0022) | 0.0514 | 0.0028 (0.0022, 0.0033) |
| 50% | 0.0487 | 0.0022 (0.0017, 0.0026) | 0.0516 | 0.0032 (0.0024, 0.0039) |
| 80% | 0.0483 | 0.0022 (0.0017, 0.0025) | 0.0527 | 0.0033 (0.0024, 0.0041) |

### 2.6.5 Implementation results

Tables 2.5 and 2.6 summarize the implementation results for the edgewise and flapwise bending moments, respectively, using 50 POE estimates for SIS1, SIS2 and BIS. For each response type, we use two different values of the resistance level, $l$. In general, the SIS1's standard errors appear to be slightly smaller than the SIS2's. In all cases, SIS1 and SIS2 outperform BIS, which confirms the theoretical advantage of their variance reductions.

We also assess the computational gains of the IS methods over CMC. Let $N_T^{(CMC)}$ denote the number of CMC simulation replications to achieve the same standard error of the corresponding method in each row of Tables 2.5 and 2.6. With $N_T^{(CMC)}$ replications, the standard error of the CMC estimator is $\sqrt{P\left(1-P\right)/N_T^{(CMC)}}$, where $P$ is the true failure probability, $P(Y > l)$. Since $P$ is unknown, we use the sample average of SIS1 for $P$ because SIS1 generates the smallest standard error in all cases. With the estimated $N_T^{(CMC)}$, we compute the relative ratio, $N_T/N_T^{(CMC)}$, as shown in Tables 2.5 and 2.6. For the edgewise moment, the SIS methods need about 5% to 9% of the CMC efforts. In other words, for $l$ = 8,600 $kNm$, CMC needs about 11,000 to 18,000 replications to obtain the same accuracy achieved by SIS1 and SIS2 with 1,000 replications. For $l$ = 9,300 $kNm$, CMC needs 51,000 to 61,000 replications compared to SIS1 and SIS2 with 3,000 replications.

Table 2.5: Estimation results of the failure probability for edgewise bending moments

| Method | $l$ = 8,600 $kNm$, $N_T$ = 1,000 | | | $l$ = 9,300 $kNm$, $N_T$ = 3,000 | | |
|---|---|---|---|---|---|---|
| | Sample Average | Standard Error (95% bootstrap CI) | Relative Ratio | Sample Average | Standard Error (95% bootstrap CI) | Relative Ratio |
| SIS1 | 0.0486 | 0.0016 (0.0012, 0.0020) | 5.5% | 0.00992 | 0.00040 (0.00032, 0.00047) | 4.9% |
| SIS2 | 0.0485 | 0.0020 (0.0016, 0.0024) | 8.7% | 0.01005 | 0.00044 (0.00036, 0.00051) | 5.9% |
| BIS | 0.0488 | 0.0029 (0.0020, 0.0037) | 18 % | 0.00995 | 0.00056 (0.00042, 0.00068) | 9.6% |

We explain the fact that the computational gains of the SIS methods for the

Table 2.6: Estimation results of the failure probability for flapwise bending moments

| Method | $l$ = 13,800 $kNm$, $N_T$ = 2,000 | | | $l$ = 14,300 $kNm$, $N_T$ = 9,000 | | |
|---|---|---|---|---|---|---|
| | Sample Average | Standard Error (95% bootstrap CI) | Relative Ratio | Sample Average | Standard Error (95% bootstrap CI) | Relative Ratio |
| SIS1 | 0.0514 | 0.0028 (0.0022, 0.0033) | 32% | 0.01070 | 0.00061 (0.00047, 0.00074) | 32% |
| SIS2 | 0.0527 | 0.0032 (0.0025, 0.0038) | 42% | 0.01037 | 0.00063 (0.00046, 0.00078) | 34% |
| BIS | 0.0528 | 0.0038 (0.0030, 0.0044) | 59% | 0.01054 | 0.00083 (0.00055, 0.00110) | 59% |

flapwise moment are not as substantial as for the edgewise moment using Figure 2.3; the SIS1 density for the flapwise moment is not as different from the original input density as is the SIS1 density for the edgewise moment. We observe similar results for the SIS2 density. As a result, the computational gains by biasing the input distribution using the SIS methods become less obvious for the flapwise moment than the edgewise moment. Recall that we observed the similar pattern in the numerical studies discussed in Section 2.5, where the computational gains of SIS1 and SIS2 are less remarkable when the optimal IS density is similar to the original input density (with $\delta = -1$ in (2.17)).

## 2.7 Summary

This chapter proposes an extended framework of IS for the reliability evaluation using a stochastic simulation model. The applicability of the existing IS methods is limited to simulations with deterministic simulation models where an output is uniquely determined for a given input.

By accounting for different sources of output variability in stochastic simulation models, we develop two methods for estimating a failure probability. For SIS1, which allows multiple replications at each sampled input, we derive the optimal IS density and allocation size that minimize the variance of the estimator. For SIS2, which uses one replication at each sampled input, we derive the optimal IS density. Since

SIS2 imposes an additional restriction on the allocation size, SIS1 is more flexible. However, SIS2 does not need to determine the input sample size and the optimal allocation size. The implementation results suggest that the performance of SIS1 is comparable to SIS2 in most cases and that both SIS methods can significantly improve the estimation accuracy over the two benchmark methods, BIS and CMC. We also observe that the computational gains of the SIS methods become larger when a smaller POE needs to be estimated and when the difference between the IS density and the original input density is larger.

# CHAPTER III

# Uncertainty Quantification of Importance Sampling Estimators for Stochastic Computer Experiments

## 3.1 Introduction

To improve the computational efficiency of reliability estimations using stochastic simulation models, two important questions need to be answered: (1) what is the optimal allocation of computational resources to minimize the estimation uncertainty and (2) how to quantify the estimation uncertainty. Chapter II addresses the first question and proposes two SIS methods to efficiently evaluate the system reliability. This chapter aims to answer the second question by proposing methods to measure the estimation uncertainty when SIS methods are used.

To this end, we establish the CLT for each of two SIS estimators under mild assumptions. Based on the CLTs, we quantify the uncertainties of SIS estimators by constructing CIs. We validate the proposed procedures using numerical studies, and demonstrate the utility of the methods via a case study on the wind turbine reliability evaluation.

## 3.2 Background

Note that we use slightly different notations to better present our methods in this chapter. First, the SIS1 estimator of the failure probability, $p_y \equiv \mathbb{P}(Y > y)$, is

$$\hat{P}_{1,n}(y) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{I}\left(Y_j^{(i)} > y\right)\right)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}, \tag{3.1}$$

where $m$ is the input sample size, denoting the number of times that the input, $\mathbf{X}$, is sampled independently from a new density, $q$; $N_i$ is the allocation size, denoting the number of simulation replications alloted to $\mathbf{X}_i$; $Y_j^{(i)}$ is the $j$th replication output at $\mathbf{X}_i$. In other words, SIS1 samples $m$ inputs, $\mathbf{X}_1, \ldots, \mathbf{X}_m$, from $q$, and runs the simulator $N_i$ times at each sampled $\mathbf{X}_i, i = 1, \ldots, m$. As a result, we observe the total $n = \sum_{i=1}^{m} N_i$ outputs of $Y_j^{(i)}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, N_i$.

The estimator, $\hat{P}_{1,n}(y)$, in (3.1) is unbiased and has the minimum variance when we use the optimal SIS1 density, $q_{1,y}(\mathbf{x})$, and the optimal allocation size, $N_i^*, i = 1, \ldots, m$, as follows:

$$q_{1,y}(\mathbf{x}) = \frac{1}{C_{q1}} f(\mathbf{x})\sqrt{\frac{1}{n}s_y(\mathbf{x})\left(1 - s_y(\mathbf{x})\right) + s_y(\mathbf{x})^2} \tag{3.2}$$

and

$$N_i^* = n\frac{h^*(\mathbf{X}_i)}{\sum_{j=1}^{m} h^*(\mathbf{X}_j)}, \quad i = 1, \ldots, m, \tag{3.3}$$

where

$$h^*(\mathbf{x}) = \sqrt{\frac{n\left(1 - s_y(\mathbf{x})\right)}{1 + (n-1)\, s_y(\mathbf{x})}}. \tag{3.4}$$

Here, $s_y(\mathbf{x})$ is $\mathbb{P}(Y > y \mid \mathbf{X} = \mathbf{x})$ and $C_{q1}$ in (3.2) is the normalizing constant. Because the conditional probability, $s_y(\mathbf{x})$, is unknown in practice, the optimal solutions in

(3.2) and (3.3) need to be approximated for implementation in practice.

In contrast to SIS1, SIS2 uses a single replication at each input (i.e., $N_i = 1, i = 1, \ldots, m$). As such, the SIS2 estimator of the failure probability is

$$\hat{P}_{2,n}(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(Y_i > y\right) L_i, \tag{3.5}$$

where the likelihood ratio, $L_i$, denotes $f(\mathbf{X}_i)/q(\mathbf{X}_i)$. The optimal SIS2 density that minimizes the variance of $\hat{P}_{2,n}(y)$ takes the following form:

$$q_{2,y}(\mathbf{x}) = \frac{1}{C_{q2}} f(\mathbf{x}) \sqrt{s_y(\mathbf{x})}, \tag{3.6}$$

where $C_{q2}$ is the normalizing constant. This optimal density also needs an approximation in practice, due to the unknown $s_y(\mathbf{x})$.

Although the above optimal solutions minimizing the variances of the estimators in (3.1) and in (3.5) have been derived for stochastic simulation models, the distributional properties of the SIS estimators are not yet understood well. In particular, quantifying the estimation uncertainty by building a valid CI would be substantially important in practice.

In this chapter, we establish the CLTs for both SIS1 and SIS2 estimators. We also propose consistent estimators for the asymptotic variances involved in the CLTs, which lead us to construct asymptotically valid CIs. In the literature, the CLT for DIS estimator is well studied (*Geweke*, 2005). However, the existing derivations are not applicable to the SIS estimators due to the intrinsic randomness within the stochastic simulation model. In this study, we address the intrinsic randomness in constructing the CLTs and CIs for the SIS estimators.

## 3.3 Asymptotic Properties of SIS Estimators

This section presents the asymptotic properties of the SIS estimators and describes how to construct the CIs based on these properties. All the proofs are available in Appendix B. We use the following three assumptions:

**Assumption III.1.** *If* $q(\mathbf{x}) = 0$*, then* $\mathbb{P}(Y > y \mid \mathbf{X} = \mathbf{x})\, f(\mathbf{x}) = 0$ *for any* $\mathbf{x}$*.*

**Assumption III.2.** $\mathbb{E}_q\left[\mathbb{I}(Y > y)L^2\right] < \infty$ *holds, where the expectation is taken with respect to* $q$*.*

**Assumption III.3.** *The ratio,* $m/n = c_0$*, is fixed for a constant,* $0 < c_0 \leq 1$*.*

The SIS1 and SIS2 methods with their optimal densities satisfy Assumptions III.1 and III.2. Assumption III.1 implies that we should use the SIS density, $q(\mathbf{x})$, that makes the SIS estimator, $\hat{P}_{k,n}(y)$ (in (3.1) for $k = 1$ or in (3.5) for $k = 2$), unbiased. This assumption is satisfied when we use the optimal SIS densities in (3.2) and (3.6) for SIS1 and SIS2, respectively (*Choe et al.*, 2015). Assumption III.2 implies that the SIS estimator should have a finite variance. This assumption is also satisfied with the optimal SIS densities as stated in the following proposition:

**Proposition III.4.** *The optimal SIS density,* $q_{k,y}$ *(in* (3.2) *for* $k = 1$ *or in* (3.6) *for* $k = 2$*), satisfies Assumption III.2.*

Assumptions III.1 and III.2 are used to establish the CLT for SIS. Analogously, to prove the CLT for DIS, similar or stronger assumptions are commonly made in the literature (*Koopman et al.*, 2009).

Assumption III.3 concerns SIS1, because SIS2 has $m/n = 1$. In practice, $m/n$ ratio for SIS1 is set at a fixed level (e.g., 30%) according to the empirical finding and implementation guideline suggested in *Choe et al.* (2015).

### 3.3.1 Central Limit Theorems for SIS1 and SIS2

Adding much complexity to DIS, the SIS1 estimator in (3.1) involves the allocation size, $N_i$, which takes account of the intrinsic randomness within the stochastic simulation model. As the first step towards proving CLT for SIS1, we need to characterize the asymptotic behavior of $N_i$.

Recall that the allocation size, $N_i$, used in practice is an approximation of the optimal allocation size, $N_i^*$, in (3.3), because $N_i^*$ involves $h^*(\mathbf{X})$, which is unknown due to the unknown conditional probability, $s_y(\mathbf{X}) = \mathbb{P}(Y > y \mid \mathbf{X})$. Let $h(\mathbf{X})$ denote the function that approximates $h^*(\mathbf{X})$. Also, we round $N_i$ to the nearest integer and, to ensure the unbiasedness of the estimator in (3.1), set $N_i$ as one if the rounding is zero. As such, the actual $N_i$ can be expressed as

$$N_i \equiv \max\left(1, \left\lfloor n\frac{h(\mathbf{X}_i)}{\sum_{j=1}^m h(\mathbf{X}_j)} + \frac{1}{2}\right\rfloor\right), \quad i = 1, \ldots, m, \tag{3.7}$$

where the floor function, $\lfloor x \rfloor$, yields the largest integer not greater than $x$. Thus, $\lfloor x + 1/2 \rfloor$ is equivalent to rounding $x$. The sum of $N_i$, $i = 1, \ldots, m$, in (3.7) may deviate slightly from the pre-specified total sample size, $n$. If we want to ensure $n = \sum_{i=1}^m N_i$ in the implementation, we can adjust either $n$ or some $N_i$'s. For simplicity, we ignore such minor adjustments in the following discussions.

The allocation size, $N_i$, in (3.7) depends not only on $\mathbf{X}_i$ but also on all $\mathbf{X}_j, j = 1, \ldots, m$. Accordingly, $N_i$ is not independent of $N_j$ for $j \neq i$. This dependency makes the derivation of CLT for SIS1 nontrivial. We first address this issue in Lemma III.5 by showing that under certain regularity conditions, the allocation sizes become mutually independent as the total sample size, $n$, increases.

**Lemma III.5. (Asymptotic independence between allocation sizes)** *Suppose that Assumption III.3 holds and that the function, $h(\cdot)$, in* (3.7) *is nonnega-*

*tive and satisfies the conditions*

$$\mathbb{E}_q[h(\mathbf{X})] < \infty \tag{3.8}$$

*and*

$$\mathbb{P}\left(\frac{h(\mathbf{X})}{c_0\mathbb{E}_q[h(\mathbf{X})]} + \frac{1}{2} \in \mathcal{N}\right) = 0, \tag{3.9}$$

*where* $\mathcal{N} \equiv \{2, 3, \ldots\}$. *Then, for any index* $i \in \{1, \ldots, m\}$,

$$N_i \xrightarrow{P} \tilde{N}_i \tag{3.10}$$

$$\equiv \max\left(1, \left\lfloor \frac{h(\mathbf{X}_i)}{c_0\mathbb{E}_q[h(\mathbf{X})]} + \frac{1}{2} \right\rfloor\right), \tag{3.11}$$

*as* $m \to \infty$. *Therefore,* $N_i, i = 1, \ldots, m$, *is asymptotically independent of one another.*

The regularity conditions in (3.8) and (3.9) generally hold in practical situations. First, the condition in (3.8) implies that the expected value of $h(\mathbf{X})$ is finite when $\mathbf{X}$ is sampled from the SIS1 density, $q$. This condition holds in practice by Proposition III.6, which implies that if we use $s'_y(\mathbf{x})$, a metamodel of $s_y(\mathbf{x})$, in both $h(\mathbf{x})$ and $q(\mathbf{x})$ to approximate $h^*(\mathbf{x})$ and $q_{1,y}(\mathbf{x})$, respectively, then $\mathbb{E}_q[h(\mathbf{X})]$ is finite.

**Proposition III.6.** *The condition,* $\mathbb{E}_q[h(\mathbf{X})] < \infty$, *in* (3.8) *holds if* $q(\mathbf{x})$ *and* $h(\mathbf{x})$ *are a density function and a non-negative function, respectively, such that a function,* $0 \le s'_y(\mathbf{x}) \le 1$, *replaces* $s_y(\mathbf{x})$ *in both* $q_{1,y}(\mathbf{x})$ *in* (3.2) *and* $h^*(\mathbf{x})$ *in* (3.4), *to yield* $q(\mathbf{x})$ *and* $h(\mathbf{x})$, *respectively.*

Next, the condition in (3.9) is to address discontinuous points due to the rounding of $N_i$, implying that the limit of non-rounded $N_i$, $h(\mathbf{X})/(c_0\mathbb{E}_q[h(\mathbf{X})]) + 1/2$, should not belong to a set of integers greater than 1. The condition in (3.9) holds when we impose the continuity on $h(\cdot)$ for continuous $\mathbf{X}$. Note that $h(\cdot)$ is a function

that approximates $h^*(\cdot)$ in (3.4). Therefore, $h(\cdot)$ can be regarded as a metamodel or emulator for $h^*(\cdot)$. In general simulation studies that develop metamodels (or emulators), it is common to model an unknown function as a continuous function (*Plumlee and Tuo*, 2014; *Zhang and Apley*, 2014, 2015). Similarly, in our case, we expect $h(\mathbf{x}_1)$ to be close to $h(\mathbf{x}_2)$ for $\mathbf{x}_1$ close to $\mathbf{x}_2$, because the conditional failure probability at $\mathbf{x}_1$, $s_y(\mathbf{x}_1)$, is generally expected to be close to $s_y(\mathbf{x}_2)$.

Building upon Lemma III.5 that characterizes the asymptotic independence of the allocation sizes, we derive the CLT for SIS1 in Theorem III.7.

**Theorem III.7. (CLT for SIS1 estimator)**

*Suppose Assumptions III.1–III.3 and the conditions in Lemma III.5 hold. Then,*

$$\sqrt{\frac{m}{\sigma_{1,y}^2}}\left(\hat{P}_{1,n}(y) - p_y\right) \xrightarrow{d} N(0,1) \tag{3.12}$$

*as* $m \to \infty$*, where*

$$\sigma_{1,y}^2 = \mathbb{E}_q\left[\frac{1}{\tilde{N}} s_y(\mathbf{X})\left(1 - s_y(\mathbf{X})\right) L^2\right] + \mathbb{E}_q\left[s_y(\mathbf{X})^2 L^2\right] - p_y^2 \tag{3.13}$$

*with*

$$\tilde{N} = \max\left(1, \left\lfloor \frac{h(\mathbf{X})}{c_0 \mathbb{E}_q[h(\mathbf{X})]} + \frac{1}{2} \right\rfloor\right).$$

Theorem III.7 describes the asymptotic normality of the SIS1 estimator, $\hat{P}_{1,n}(y)$, in (3.1). As $m$ increases, the SIS1 estimator becomes close to a normal random variable with the mean of $p_y$ and the variance of $\sigma_{1,y}^2/m$. We note that '$m \to \infty$' is equivalent to '$n \to \infty$', because $m$ and $n$ are of the same order by Assumption III.3.

Next, Theorem III.8 states the CLT for SIS2, implying that as $n$ increases, $\hat{P}_{2,n}(y)$ becomes close to a normal random variable with the mean of $p_y$ and the variance of $\sigma_{2,y}^2/n$.

**Theorem III.8. (CLT for SIS2 estimator)**

*Under Assumptions III.1–III.2,*

$$\sqrt{\frac{n}{\sigma_{2,y}^2}}\left(\hat{P}_{2,n}(y) - p_y\right) \xrightarrow{d} N(0,1) \tag{3.14}$$

*as* $n \to \infty$*, where*

$$\sigma_{2,y}^2 = \mathbb{E}_q\left[s_y(\mathbf{X})L^2\right] - p_y^2. \tag{3.15}$$

Both Theorems III.7 and III.8 provide the information on the distributional properties of SIS1 and SIS2 estimators in the asymptotic regime. Yet, the asymptotic variances are unknown, because $\sigma_{i,y}^2$ (in (3.13) for $i = 1$ or in (3.15) for $i = 2$) involves $s_y(\mathbf{X}) = \mathbb{P}(Y > y \mid \mathbf{X})$ and $p_y$. In the next section, we devise consistent estimators of the asymptotic variances and use them to construct the asymptotically valid confidence intervals for $p_y$.

### 3.3.2 Confidence Intervals for SIS1 and SIS2

We note that by the Slutsky's theorem (*Jiang*, 2010, Theorem 2.13), replacing $\sigma_{i,y}^2$ for $i = 1, 2$ in the CLTs with their consistent estimators does not change the limiting distributions. Therefore, the asymptotic normalities in Theorems III.7 and III.8 still hold when we substitute the asymptotic variances with their consistent estimators.

Theorems III.9 and III.10 present the consistent estimators, $\hat{\sigma}_{i,y}^2$, for $\sigma_{i,y}^2$ for $i = 1, 2$, and construct the CIs for $p_y$. We define $z_{\alpha/2} \equiv \Phi^{-1}(1-\alpha/2)$ for $\alpha \in (0,1)$, where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$.

**Theorem III.9. (CI for SIS1)** *Suppose Assumptions III.1–III.3 and the conditions in Lemma III.5 hold.*

*(a) Then,*

$$\hat{\sigma}^2_{1,y} \xrightarrow{P} \sigma^2_{1,y} \tag{3.16}$$

*as* $m \to \infty$*, where*

$$\hat{\sigma}^2_{1,y} = \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{I}\left(Y_j^{(i)} > y\right)L_i - \hat{P}_{1,n}(y)\right)^2. \tag{3.17}$$

*(b)*

$$\sqrt{\frac{m}{\hat{\sigma}^2_{1,y}}}\left(\hat{P}_{1,n}(y) - p_y\right) \xrightarrow{d} N(0,1) \tag{3.18}$$

*as* $m \to \infty$*. Therefore,* $\mathbb{P}\Big(p_y \in \Big(\hat{P}_{1,n}(y) \pm z_{\alpha/2}\hat{\sigma}_{1,y}/\sqrt{m}\Big)\Big) \to 1-\alpha$ *for* $\alpha \in (0,1)$ *as* $m \to \infty$*. That is,* $\Big(\hat{P}_{1,n}(y) \pm z_{\alpha/2}\hat{\sigma}_{1,y}/\sqrt{m}\Big)$ *is a* $100(1-\alpha)\%$ *asymptotic confidence interval for* $p_y$*.*

**Theorem III.10. (CI for SIS2)**

*Suppose Assumptions III.1 and III.2 hold.*

*(a) Then,*

$$\hat{\sigma}^2_{2,y} \xrightarrow{P} \sigma^2_{2,y} \tag{3.19}$$

*as* $n \to \infty$*, where*

$$\hat{\sigma}^2_{2,y} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbb{I}\left(Y_i > y\right)L_i - \hat{P}_{2,n}(y)\right)^2. \tag{3.20}$$

*(b)*

$$\sqrt{\frac{n}{\hat{\sigma}^2_{2,y}}}\left(\hat{P}_{2,n}(y) - p_y\right) \xrightarrow{d} N(0,1) \tag{3.21}$$

*as* $n \to \infty$. *Therefore,* $\mathbb{P}\Big(p_y \in \Big(\hat{P}_{2,n}(y) \pm z_{\alpha/2}\hat{\sigma}_{2,y}/\sqrt{n}\Big)\Big) \to 1-\alpha$ *for* $\alpha \in (0,1)$ *as* $n \to \infty$. *That is,* $\Big(\hat{P}_{2,n}(y) \pm z_{\alpha/2}\hat{\sigma}_{2,y}/\sqrt{n}\Big)$ *is a* $100(1-\alpha)\%$ *asymptotic confidence interval for* $p_y$.

### 3.3.3 Confidence Intervals With Different Thresholds

The optimal SIS solutions depend on the failure threshold, $y$, leading to the sampling and simulation results optimized for the particular $y$. Suppose we obtain the simulation outputs with $y$. We can still use the same simulation outputs to estimate the failure probability at a different threshold, $\tilde{y}$, for $\tilde{y} > y$ without conducting experiments again.

Suppose we sample $\mathbf{X}_i, i = 1, \ldots, m$, from $q_{k,y}$ (in (3.2) for $k = 1$ or in (3.6) for $k = 2$) and obtain the simulation outputs, $Y_j^{(i)}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, N_i$ (note that $n = m$ in SIS2). Then, we can replace $y$ with $\tilde{y}$ and use the SIS estimator, $\hat{P}_{k,n}(\tilde{y})$ (in (3.1) for $k = 1$ or in (3.5) for $k = 2$), to estimate the failure probability, $p_{\tilde{y}} = \mathbb{P}(Y > \tilde{y})$. The estimator, $\hat{P}_{k,n}(\tilde{y})$, is an unbiased estimator of $p_{\tilde{y}}$ for $\tilde{y} > y$ (*Choe and Byon*, 2015). Moreover, we can construct the pointwise CI for $p_{\tilde{y}}$ by substituting $\tilde{y}$ for $y$ in Theorem III.9 for $k = 1$ (or Theorem III.10 for $k = 2$) for $\tilde{y} > y$, as stated in Corollary III.11 below.

**Corollary III.11. (Pointwise CI for $\tilde{y} > y$)**

*(a) Suppose the conditions in Theorem III.9 hold. Then, for* $\tilde{y} > y$, *the CI for* $p_{\tilde{y}}$, $\Big(\hat{P}_{1,n}(\tilde{y}) \pm z_{\alpha/2}\hat{\sigma}_{1,\tilde{y}}/\sqrt{m}\Big)$ *is asymptotically valid, i.e.,*

$$\mathbb{P}\Big(p_{\tilde{y}} \in \Big(\hat{P}_{1,n}(\tilde{y}) \pm z_{\alpha/2}\hat{\sigma}_{1,\tilde{y}}/\sqrt{m}\Big)\Big) \to 1-\alpha$$

*for* $\alpha \in (0,1)$ *as* $m \to \infty$.

*(b) Suppose the conditions in Theorem III.10 hold. Then, for* $\tilde{y} > y$*, the CI for* $p_{\tilde{y}}$*,* $\left(\hat{P}_{2,n}(\tilde{y}) \pm z_{\alpha/2}\hat{\sigma}_{2,\tilde{y}}/\sqrt{n}\right)$ *is asymptotically valid, i.e.,*

$$\mathbb{P}\Big(p_{\tilde{y}} \in \left(\hat{P}_{2,n}(\tilde{y}) \pm z_{\alpha/2}\hat{\sigma}_{2,\tilde{y}}/\sqrt{n}\right)\Big) \to 1-\alpha$$

*for* $\alpha \in (0,1)$ *as* $n \to \infty$.

We believe that the results in Corollary III.11, which justifies the CIs for $\tilde{y} > y$, are practically desirable. At the system design stage, designers want to estimate the failure probability and quantify the estimation uncertainties in a range of design parameters, $\tilde{y}$, rather than at a single value of $y$. In particular, system designers are interested in a large resistance level, $\tilde{y}$, which corresponds to a small failure probability, $p_{\tilde{y}}$, to ensure a high level of system reliability. Corollary III.11 suggests that we can construct the CIs for $p_{\tilde{y}}$ using the results optimized for $p_y$, without running the simulation with each $\tilde{y}$.

### 3.3.4 Implementation Summary

We summarize how to implement the proposed procedure. Recall that SIS2's input sample size, $m$, is equal to the total sample size, $n$, because SIS2 sets $N_i = 1$ for $i = 1, \ldots, m$.

**Implementation procedure ($k = 1$ for SIS1 or $k = 2$ for SIS2):**

1. Given $y$, sample $\mathbf{X}_i, i = 1, \ldots, m$, from the SIS density, $q_{k,y}$ (in (3.2) for $k = 1$ or in (3.6) for $k = 2$).

2. For each $\mathbf{X}_i$, run the simulator $N_i$ (in (3.3) for $k = 1$ or $N_i = 1$ for $k = 2$) times to obtain $Y_j^{(i)}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, N_i$.

3. Estimate the failure probability for $\tilde{y}$ by $\hat{P}_{k,n}(\tilde{y})$ (in (3.1) for $k = 1$ or in (3.5)

for $k = 2$) for $\tilde{y} \geq y$.

4. Obtain $\hat{\sigma}_{k,\tilde{y}}$ (in (3.17) for $k = 1$ or in (3.20) for $k = 2$).

5. Construct the $100(1-\alpha)\%$ pointwise CI for $p_{\tilde{y}}$ using $\left(\hat{P}_{k,n}(\tilde{y}) \pm z_{\alpha/2}\hat{\sigma}_{k,\tilde{y}}/\sqrt{m}\right)$.

In Steps 1 and 2, as noted in Section 2, the SIS density and allocation size need approximations, since the conditional probability, $s_y(\mathbf{x}) = \mathbb{P}(Y > y \mid \mathbf{X} = \mathbf{x})$, is unknown. Recall that Chapter II provides a guideline on how to approximate $s_y(\mathbf{x})$ using a metamodel.

## 3.4 Numerical Studies

This section presents numerical examples to show that the empirical coverage levels of the proposed CIs agree with the target coverage probability, $1 - \alpha$, under various settings. We use two data generating models.

### 3.4.1 Example 1

*Cannamela et al.* (2008) originally develop a deterministic simulation example, which is later modified by *Choe et al.* (2015) as the stochastic simulation example. We use the same stochastic data generating model as follows:

$$X \sim N(0,1)\,, \qquad Y|X \sim N\big(\mu(X)\,, \sigma^2(X)\big)\,, \tag{3.22}$$

where the mean, $\mu(X)$, and the standard deviation, $\sigma(X)$, of the normal distribution are

$$\begin{aligned} \mu(X) &= 0.95\delta X^2\left(1 + 0.5\cos(5X) + 0.5\cos(10X)\right), \\ \sigma(X) &= 1 + 0.7\,|X| + 0.4\cos(X) + 0.3\cos(14X), \end{aligned} \tag{3.23}$$

respectively. The metamodel of the conditional distribution of $Y|X$ is set as the normal distribution with the following mean and standard deviation:

$$\mu_{meta}(X) = 0.95\delta X^2 \left(1 + 0.5\rho\cos(5X) + 0.5\rho\cos(10X)\right), \tag{3.24}$$
$$\sigma_{meta}(X) = 1 + 0.7\,|X| + 0.4\rho\cos(X) + 0.3\rho\cos(14X).$$

In this example, all model and experiment parameters are set as in Chapter II (see Section 2.5). The model parameter, $\delta$, in (3.23) and (3.24) determines the similarity of the SIS density, $q_{k,y}, k = 1, 2$, in (3.2) and (3.6) to the original input density, $f$. For $\delta = 1$ $(-1)$, the important regions are far from (close to) $X = 0$, which is the mode of $f$, the density of $N(0, 1)$. Consequently, the SIS densities that focus on the important regions differ significantly for different $\delta$'s. Another model parameter, $\rho$, controls the metamodel accuracy: the metamodel with $\rho = 0$ captures only the global pattern of important region, whereas the metamodel with $\rho = 1$ is equivalent to the true data generating model. In this example, we set $\rho$ as 0.5, which represents a moderate metamodel quality. We use the failure threshold that corresponds to the true failure probability, $p_y = 0.01$. For SIS1, the ratio of $m/n$ is set as 30%.

To compute the empirical coverage level, we repeatedly construct the $100(1-\alpha)\%$ CI, $\left(\hat{P}_{k,n}(y) \pm z_{\alpha/2}\hat{\sigma}_{k,y}/\sqrt{m}\right)$, 10,000 times and calculate the proportion of the CIs covering the true failure probability, $p_y$. We consider the target coverage probability, $1-\alpha$, of 0.90 and 0.95. Table 3.1 shows the experiment results. We summarize the key observations as follows:

- With the moderate size of $n$ of 1000 (note that $p_y = 0.01$), the corresponding empirical coverages are close to the target coverage probabilities, $1-\alpha$, for both SIS1 and SIS2.

- As $n$ increases, the empirical coverage levels for both SIS1 and SIS2 reach the target coverage probability, $1-\alpha$. This result agrees with the asymptotic results

stated in Theorems III.9 and III.10.

- Across all cases, SIS1 and SIS2 maintain the same empirical coverage level up to the second decimal place, showing that their CIs perform similarly.
- The parameters, $\alpha$ and $\delta$, do not appear to significantly affect the behavior of CI coverage.

Table 3.1: Empirical coverage level in Example 1

| | $\delta = 1$ | | | | $\delta = -1$ | | | |
|---|---|---|---|---|---|---|---|---|
| $1-\alpha$ | 0.90 | | 0.95 | | 0.90 | | 0.95 | |
| $n$ | SIS1 | SIS2 | SIS1 | SIS2 | SIS1 | SIS2 | SIS1 | SIS2 |
| 1000 | 0.88 | 0.88 | 0.94 | 0.94 | 0.88 | 0.88 | 0.93 | 0.93 |
| 10000 | 0.89 | 0.89 | 0.95 | 0.95 | 0.90 | 0.90 | 0.95 | 0.95 |
| 100000 | 0.90 | 0.90 | 0.95 | 0.95 | 0.90 | 0.90 | 0.95 | 0.95 |

NOTE: The empirical coverage level is the proportion of CIs (out of 10,000 experiments) that include the true failure probability, $p_y = 0.01$.

### 3.4.2 Example 2

*Ackley* (1987) proposes a deterministic simulation example which is later modified by *Huang et al.* (2006) and *Choe et al.* (2015) into a stochastic simulation example with the three-dimensional input vector, $\mathbf{X} = (X_1, X_2, X_3)$, following a multivariate normal distribution. We use the same data generating model:

$$\mathbf{X} \sim MVN(\mathbf{0}, \mathbf{I}_3)\,, \qquad Y|\mathbf{X} \sim N\big(\mu(\mathbf{X})\,, \sigma^2(\mathbf{X})\big)\,,$$

where $\mathbf{I}_3$ is the 3 by 3 identity matrix. The mean function, $\mu(\mathbf{X})$, and the standard deviation function, $\sigma(\mathbf{X})$, take the following forms that represent highly nonlinear

response surface and heterogeneous variability over a range of input conditions:

$$\mu(\mathbf{X}) = 20\delta\left(1 - \exp\left(-0.2\sqrt{\frac{1}{3}\|\mathbf{X}\|^2}\right)\right) + \delta\left(\exp(1) - \exp\left(\frac{1}{3}\sum_{i=1}^{3}\cos(\pi X_i)\right)\right),$$
$$\sigma(\mathbf{X}) = 1 + 0.7\sqrt{\frac{1}{3}\|\mathbf{X}\|^2} + 0.4\left(\frac{1}{3}\sum_{i=1}^{3}\cos(3\pi X_i)\right).$$

As the metamodel of the conditional distribution, $Y|\mathbf{X}$, we use

$$N\big(\mu_{meta}(\mathbf{X}), \sigma^2_{meta}(\mathbf{X})\big),$$

where

$$\mu_{meta}(\mathbf{X}) = 20\delta\left(1 - \exp\left(-0.2\sqrt{\frac{1}{3}\|\mathbf{X}\|^2}\right)\right) + \rho\delta\left(\exp(1) - \exp\left(\frac{1}{3}\sum_{i=1}^{3}\cos(\pi X_i)\right)\right),$$
$$\sigma_{meta}(\mathbf{X}) = 1 + 0.7\sqrt{\frac{1}{3}\|\mathbf{X}\|^2} + 0.4\rho\left(\frac{1}{3}\sum_{i=1}^{3}\cos(3\pi X_i)\right).$$

Here, the model parameters, $\delta$ and $\rho$, play essentially the same roles as in the first example in Section 4.1. Namely, $\delta = 1$ $(-1)$ means that the important input conditions are far from (close to) the origin, $\mathbf{X} = \mathbf{0}$, which is the mode of $f$, $MVN(\mathbf{0}, \mathbf{I}_3)$. Also, $\rho$ is the metamodel accuracy tuning parameter with the same interpretation as the first example's $\rho$. As in the first example, we set $\rho$ as 0.5 and use the failure threshold associated with $p_y = 0.01$. The ratio of $m/n$ is fixed at 30% for SIS1.

Table 3.2 shows the empirical coverage level of the $100(1-\alpha)\%$ CI when the target coverage probability, $1-\alpha$, is 0.90 or 0.95. The results are similar to the first example's results, echoing the characteristics of the CIs observed previously. In particular, considering the complex mean and variance structure in this example, the good agreements even with the moderate size of $n = 1,000$ (or 10,000) for estimating the failure probability of $p_y = 0.01$ support the usefulness of the proposed CI with limited computational resources in practice.

Table 3.2: Empirical coverage level in Example 2

| | $\delta = 1$ | | | | $\delta = -1$ | | | |
|---|---|---|---|---|---|---|---|---|
| $1-\alpha$ | 0.90 | | 0.95 | | 0.90 | | 0.95 | |
| $n$ | SIS1 | SIS2 | SIS1 | SIS2 | SIS1 | SIS2 | SIS1 | SIS2 |
| 1000 | 0.87 | 0.87 | 0.92 | 0.92 | 0.88 | 0.88 | 0.93 | 0.93 |
| 10000 | 0.89 | 0.89 | 0.94 | 0.94 | 0.90 | 0.90 | 0.95 | 0.95 |
| 100000 | 0.90 | 0.90 | 0.95 | 0.95 | 0.90 | 0.90 | 0.95 | 0.95 |

NOTE: The empirical coverage level is the proportion of CIs (out of 10,000 experiments) that include the true failure probability, $p_y = 0.01$.

## 3.5 Case Study: Implementation With Wind Turbine Simulators

We use the same simulation setting as in Chapter II (see Section 2.6). This case study aims to estimate the probability that the load of interest, $Y$, will exceed a threshold, $y$. In particular, we estimate a small probability associated with an extreme load level, which can be observed rarely with the probability less than, or equal to, 0.01. Thus, the brute-force approach like CMC raises serious concerns on the computational cost (*Moriarty*, 2008; *Manuel et al.*, 2013). As a remedy, we use SIS and provide the CIs for probability estimation.

We first test whether the empirical coverage level of CI is similar to the target coverage probability. Unlike the numerical studies in Section 4 where we repeat the experiment 10,000 times, we limit the repetition to 50 times in this case study because of the high computational cost. For each experiment, we use the same setup used in Chapter II: namely, for the edgewise bending moment, we use $y = 9300$ $kNm$, $n = 3,000$, and $m/n = 10\%$; for the flapwise bending moment, we use $y = 14,300$ $kNm$, $n = 9,000$, and $m/n = 30\%$. Both $y$ values are associated with $p_y$ close to 0.01. Because $p_y$ is unknown, we estimate it with the sample average of the 50 failure probability estimates. We compute the empirical coverage level by obtaining the proportion of CIs that cover the estimated $p_y$.

Table 3.3 shows the empirical coverage level for the different target coverage prob-

ability ($1-\alpha = 0.90$ or 0.95) and the different load type (edgewise or flapwise bending moment). The observed coverage level is generally similar to the target level, considering that the proportion is subject to the randomness. We note that the empirical coverage level does not exactly match the target coverage probability, although the difference is small. We believe that the slight mismatch is due to the small number of repetitions (50 in this case study) and the possible deviation of the sample average of 50 estimates from the true, unknown $p_y$.

Table 3.3: CI coverage from 50 experiments in the case study (empirical coverage level)

| $1-\alpha$ | 0.90 | | 0.95 | |
|---|---|---|---|---|
| | SIS1 | SIS2 | SIS1 | SIS2 |
| Edgewise | 0.96 (48/50) | 0.96 (48/50) | 1.00 (50/50) | 0.98 (49/50) |
| Flapwise | 0.96 (48/50) | 0.92 (46/50) | 0.96 (48/50) | 0.92 (46/50) |

NOTE: The first number in each parenthesis denotes the number of experiments whose CIs include the estimated $p_y$.

Next, to illustrate how the CIs can help a wind turbine design process, we estimate the failure probability of $10^{-2}$ or less because such a small failure probability is desired in the wind industry (*Lee et al.*, 2013). To do so, we pool all the results from the 50 repetitions of experiments. The pooled estimator of the failure probability, $p_y$, is $\hat{P}_{k,50n}(y)$ (in (3.1) with $m$ replaced by $50m$ for SIS1 ($k = 1$) or in (3.5) with $n$ replaced by $50n$ for SIS2 ($k = 2$)). We also construct the CIs using the results in Theorems III.9 and III.10 with $50n$ and $50m$ in place of $n$ and $m$, respectively. Moreover, we obtain the pointwise CIs of $p_{\tilde{y}}$ for $\tilde{y} > y$, based on Corollary III.11.

To illustrate, Figure 3.1 shows the SIS1 point estimates and pointwise CIs for the failure probabilities corresponding to $\tilde{y}$ greater than, or equal to, $y = 9{,}300$ $kNm$ for edgewise bending moments (we omit the SIS2's result as it is similar to SIS1's). In Figure 3.1, we note that the CIs get wider as $\tilde{y}$ increases, reflecting the increasing uncertainty in the distribution tail. This is because the experiments were optimized to estimate $p_y$ for $y = 9{,}300$ $kNm$. As the threshold, $\tilde{y}$, increases, a smaller number

of simulation outputs, which were obtained from the original experiments with $y =$ 9,300 $kNm$, are used to compute $\hat{P}_{k,50n}(\tilde{y})$ (in (3.1) for SIS1 ($k = 1$) or in (3.5) for SIS2 ($k = 2$)) and the corresponding CIs in Corollary III.11, because a large number of outputs result in $\mathbb{I}\left(Y_j^{(i)} > \tilde{y}\right) = 0$ in (3.1) or $\mathbb{I}\left(Y_i > y\right) = 0$ in (3.5). Accordingly, as $\tilde{y}$ becomes substantially greater than $y$, the estimation uncertainties get larger. Note that the sharp decline in the lower CI bound at the tail (around 11,600 $kNm$) in Figure 3.1 is mainly due to the fact that the failure probability in the $y$-axis is in the log scale.

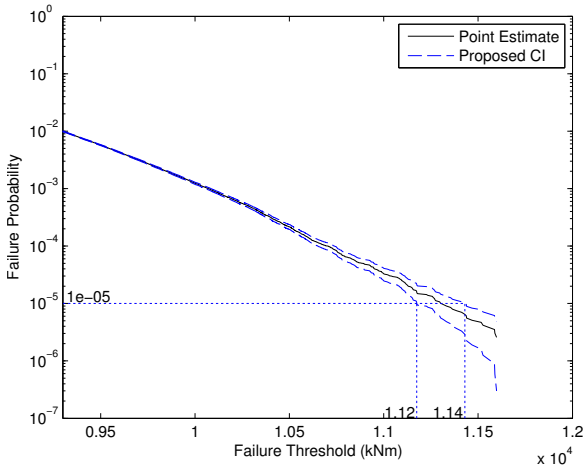



Figure 3.1: Failure probability estimates and 95% pointwise CIs from SIS1 for edgewise bending moments using the simulation outputs from 50 repetitions with $y =$ 9,300 $kNm$

## 3.6 Summary

SIS estimators can significantly save computational resources in estimating the probability associated with the output of stochastic simulation model. This chapter studies the asymptotic properties of the SIS estimators with a focus on measuring the estimation uncertainty. We prove the CLTs for the SIS estimators and construct the asymptotically valid CIs that use asymptotic variance estimators. Numerical studies show that the asymptotic CI's empirical coverage level indeed converges to the target

coverage probability. In our case study, we use the CI to quantify the uncertainty of the failure probability estimation for wind turbine reliability evaluation.

# CHAPTER IV

# EM-Based Cross-Entropy Method With an Asymptotically Unbiased Information Criterion

## 4.1 Introduction

As we discussed in the previous chapters, the theoretically optimal IS density is not implementable in practice, necessitating some approximations such as a metamodel-based approach (*Dubourg et al.*, 2013) or the CE method (*Rubinstein*, 1999). In Chapter II, we see that the performances of IS methods highly depend on the metamodel quality. When the metamodel does not approximate the unknown quantities well, the computational advantage of IS deteriorates. In this chapter, we study the CE method that does not require the metamodel construction. The proposed approach will be useful when it is difficult to build a good metamodel, especially when the response surface is complicated.

In the standard CE method, the candidate IS density is confined to a parametric family, often becoming too rigid to capture the complicated important region (*Botev et al.*, 2013). Nonparametric approaches can overcome such limitations, but encounter computational challenges (*Rubinstein*, 2005; *Botev et al.*, 2007).

This chapter aims to overcome the limitations in the existing CE methods and provides a new approach to find an appropriate IS density by using the Gaussian

mixture model (GMM) in a flexible and computationally efficient manner. One of the well-known issues of using the GMM in the statistical learning is the model selection problem (*Figueiredo and Jain*, 2002), because the number of mixture components (or model order), $k$, cannot be chosen by simply maximizing the likelihood. As a remedy, some theoretically valid criteria such as Akaike information criterion (AIC) (*Akaike*, 1974) and BIC (*Schwarz*, 1978) are adopted to balance between the model fitting and the model complexity. Noting an analogy between minimizing the deviation of a GMM from the optimal IS density and maximizing the likelihood of a GMM by the expectation–maximization (EM) algorithm, we derive a new information criterion similar to AIC. The resulting criterion shares the theoretical properties of AIC, and enables us to automatically identify the model order by balancing between the model fitting to the optimal IS density and the model complexity. The proposed criterion is applicable to both deterministic and stochastic simulation models.

## 4.2 Background

In this chapter, we use slightly different notations for better presentation. Specifically, the CMC estimator is

$$\hat{P}_{CMC} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(Y > l\right), \tag{4.1}$$

where $n$ is the number of total simulation replications. The IS estimator for deterministic simulation models is

$$\hat{P}_{DIS} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(Y_i > l\right) \frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}, \tag{4.2}$$

where $\mathbf{X}_i, i = 1, \ldots, n$, is sampled from $q$. $Y_i$ is the output corresponding to $\mathbf{X}_i$. We consider SIS1 (instead of SIS2) as a representative of SIS, because SIS1 estimator in (3.1) reduces to SIS2 estimator in (3.5) for $m = n$ and SIS1 density in (3.2) takes

a more complicated form to approximate than SIS2 density in (3.6). Thus, in this chapter, the SIS estimator denotes

$$\hat{P}_{SIS} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{I}\left(Y_j^{(i)} > l\right)\right)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}, \tag{4.3}$$

where $m$ is the number of input drawings such that $n = \sum_{i=1}^{m} N_i$. $Y_j^{(i)}$ is the $j$th output from $N_i$ replications at the input, $\mathbf{X}_i$.

### 4.2.1 Standard CE Method

The CE method (*Rubinstein*, 1999) is originally developed to find the density that best approximates the optimal density of DIS. We later show that CE is also applicable to SIS.

The standard CE method limits the search space for the optimal IS density, $q^*(\mathbf{x})$, to a pre-specified parametric family (e.g., Gaussian, Poisson, gamma, etc.), $\{q(\mathbf{x};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d\}$, and seeks the density, $q(\mathbf{x};\boldsymbol{\theta}^*)$, that is closest to the optimal density. The closeness is measured by the Kullback-Leibler divergence,

$$\mathbb{D}(q^*, q) = \int q^*(\mathbf{x}) \ln q^*(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int q^*(\mathbf{x}) \ln q(\mathbf{x};\boldsymbol{\theta})\,\mathrm{d}\mathbf{x}. \tag{4.4}$$

This quantity is always non-negative and takes zero if and only if $q^*(\mathbf{x}) = q(\mathbf{x};\boldsymbol{\theta})$ almost everywhere. Thus, minimizing $\mathbb{D}(q^*, q)$ over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ leads to $q(\mathbf{x};\boldsymbol{\theta}^*) = q^*(\mathbf{x})$ if $q^*$ belongs to the same parametric family.

Minimizing $\mathbb{D}(q^*, q)$ in (4.4) over $\boldsymbol{\theta}$ is equivalent to minimizing its second term, known as the cross-entropy

$$\mathbb{C}(q^*, q) = -\int q^*(\mathbf{x}) \log q\,(\mathbf{x};\boldsymbol{\theta})\,\mathrm{d}\mathbf{x}, \tag{4.5}$$

because the first term in (4.4) is constant. Noting that the IS optimal density can be

expressed as $q^*(\mathbf{x}) \propto h(\mathbf{x})f(\mathbf{x})$, where $h(\mathbf{x})$ is $\mathbb{I}\left(g(\mathbf{x}) > l\right)$ for DIS (later, we will also consider $h(\mathbf{x}) = \sqrt{s(\mathbf{x})\left(1 - s(\mathbf{x})\right)/n + s(\mathbf{x})^2}$ for SIS with $s(\mathbf{x}) = \mathbb{P}(Y > l \mid \mathbf{X} = \mathbf{x})$), the CE method aims to equivalently minimize

$$\mathcal{C}_{\boldsymbol{\theta}} = -\int h(\mathbf{x})f(\mathbf{x}) \log q\left(\mathbf{x}; \boldsymbol{\theta}\right) \mathrm{d}\mathbf{x} \tag{4.6}$$

over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

In practice, the CE method typically uses an iterative procedure. Let $\hat{\boldsymbol{\theta}}'$ denote the parameter estimate for the IS density, $q$, in the previous iteration. In the current iteration, the CE method finds $\hat{\boldsymbol{\theta}}$ that minimizes the following IS estimator of (4.6),

$$\bar{\mathcal{C}}_{\boldsymbol{\theta}} = -\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i) \log q(\mathbf{X}_i; \boldsymbol{\theta}) \tag{4.7}$$

where $w(\mathbf{X}_i)$ is the likelihood ratio, $f(\mathbf{X}_i)/q(\mathbf{X}_i; \hat{\boldsymbol{\theta}}')$, and $\mathbf{X}_i, i = 1, \ldots, n$, is sampled from $q(\mathbf{x}; \hat{\boldsymbol{\theta}}')$. We summarize the CE method as follows:

Step 1. Sample $\mathbf{X}_i, i = 1, \ldots, n$, from $q(\mathbf{x}; \hat{\boldsymbol{\theta}}')$. At the first iteration, $q(\mathbf{x}; \hat{\boldsymbol{\theta}}')$ can be flexible (e.g., $f$ is commonly used.)

Step 2. Find $\hat{\boldsymbol{\theta}} = \mathrm{argmin}_{\boldsymbol{\theta}}\, \bar{\mathcal{C}}_{\boldsymbol{\theta}}$, where $\bar{\mathcal{C}}_{\boldsymbol{\theta}}$ is in (4.7).

Step 3. Set $\hat{\boldsymbol{\theta}}' = \hat{\boldsymbol{\theta}}$ and start the next iteration from Step 1 until some stopping criterion is met.

This procedure iteratively refines $q(\mathbf{x}; \hat{\boldsymbol{\theta}})$. However, the refinement is limited, as the search space is less flexibly defined by a parametric family.

### 4.2.2 Variations of CE Method

Some studies (*Rubinstein*, 2005; *Botev et al.*, 2007) explore nonparametric approaches to allow greater flexibility on the candidate IS density than the standard

CE method. However, the flexibility comes with great costs: finding the optimal density (*Botev et al.*, 2007) or sampling from the optimized density (*Rubinstein*, 2005) is computationally challenging.

Bridging between the two extremes of the spectrum (a parametric density with $d \ll n$ or a nonparametric density with $d \asymp n$, where $d$ is the number of parameters in a candidate IS density and $n$ is the number of simulation replications), a few studies (*Botev et al.*, 2013; *Wang and Zhou*, 2015; *Kurtz and Song*, 2013) recently consider the mixture of parametric distributions, where $d$ can vary between 1 and $n$. This approach is particularly desirable for engineering applications because (a) it can be as flexible as we want; (b) it is easy and fast to sample from the candidate IS density; and (c) the optimized IS density provides an insight on the engineering system (e.g., means of mixture components often coincide with the so-called 'hot spots', where the system likely fails).

## 4.3 Methodology

This section uses the GMM to find the IS density under the CE framework, and derives a new asymptotically unbiased information criterion to automatically determine the model order, $k$, of the GMM.

### 4.3.1 Gaussian Mixture Model and EM algorithm

We express the candidate IS density by GMM:

$$q(\mathbf{x};\boldsymbol{\theta}) = \sum_{j=1}^{k} \alpha_j \, q_j\big(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\big)\,, \tag{4.8}$$

where the component weights, $\alpha_j, j = 1,\ldots,k$, sum to one. The $j$th Gaussian component density, $q_j$, is specified by the mean, $\boldsymbol{\mu}_j$, and the covariance $\boldsymbol{\Sigma}_j$. Thus, $\boldsymbol{\theta}$ denotes $(\alpha_1,\ldots,\alpha_k,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_k)$.

To minimize (4.7), the gradient of (4.7) is set to zero:

$$-\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i)\nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X}_i;\boldsymbol{\theta}\right) = 0. \tag{4.9}$$

This leads to the updating equations as derived in (*Kurtz and Song*, 2013):

$$\alpha_j = \frac{\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i)\gamma_{ij}}{\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i)}, \tag{4.10}$$

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i)\gamma_{ij}\mathbf{X}_i}{\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i)\gamma_{ij}}, \tag{4.11}$$

$$\boldsymbol{\Sigma}_j = \frac{\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i)\gamma_{ij}(\mathbf{X}_i-\boldsymbol{\mu}_j)(\mathbf{X}_i-\boldsymbol{\mu}_j)^T}{\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i)\gamma_{ij}}, \tag{4.12}$$

where

$$\gamma_{ij} = \frac{\alpha_j\, q_j\big(\mathbf{X}_i;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\big)}{\sum_{j'=1}^{k} \alpha_{j'}\, q_{j'}\big(\mathbf{X}_i;\boldsymbol{\mu}_{j'},\boldsymbol{\Sigma}_{j'}\big)}. \tag{4.13}$$

As the name suggests, the right-hand sides of the 'updating' equations (4.10), (4.11), (4.12) involve $\boldsymbol{\theta} = (\alpha_1,\ldots,\alpha_k,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_k)$ either explicitly or implicitly through $\gamma_{ij}$. As such, the updating equations are interlocking with each other and cannot be solved analytically. Thus, by starting with an initial value for $\boldsymbol{\theta}$ on the right-hand sides of the updating equations, we need to compute the left-hand sides and plug the results back to the right-hand sides iteratively until the convergence is reached.

This optimization procedure is called the EM algorithm that alternates between the expectation step (computing $\gamma_{ij}$) and the maximization step (updating $\boldsymbol{\theta}$). The study (*Kurtz and Song*, 2013) that derives the updating equations does not notice the connection with the EM algorithm. Moreover, the existing studies on the mixture model (*Botev et al.*, 2013; *Wang and Zhou*, 2015; *Kurtz and Song*, 2013) do not iterate the updating equations but solves them only once in each CE iteration. To actually 'minimize' (4.7), it is necessary to use the EM algorithm (i.e., iterating the updating

equations until convergence) within each CE iteration.

### 4.3.2 Cross-Entropy Information Criterion

Due to the difficulty choosing the number of mixture components, $k$, existing studies either assume that $k$ is given (*Botev et al.*, 2013; *Kurtz and Song*, 2013) or follow a rule of thumb based on "some understanding of the structure of the problem at hand" (*Wang and Zhou*, 2015). We derive an asymptotically unbiased criterion to choose $k$ automatically. We borrow the ideas of the information criteria widely used in statistical learning (*Figueiredo and Jain*, 2002), where the best model is chosen by minimizing a criterion generally expressed as

$$-\frac{1}{n}\sum_{i=1}^{n}\log q\Big(\mathbf{X}_i;\tilde{\boldsymbol{\theta}}\Big)+\mathcal{P}(d), \tag{4.14}$$

which balances between the model's goodness of fit and the model complexity: the first term is the average negative log likelihood of the model, which is minimized by the maximum likelihood estimator (MLE), $\tilde{\boldsymbol{\theta}}_{mle}$. The second term is a monotonically increasing function of $d$ to penalize the overly complex model (note that $d$ is the dimension of $\tilde{\boldsymbol{\theta}}$ and proportional to $k$.). For example, when $\mathcal{P}(d)=d/n$, the criterion in (4.14) becomes AIC (*Akaike*, 1974); when $\mathcal{P}(d)=d(\log n)/n$, the criterion becomes BIC (*Schwarz*, 1978).

However, we cannot directly use the existing criteria to find the best parameter of GMM approximating the optimal IS density, because our goal is not finding the best model explaining the given data (i.e., maximizing the likelihood). Instead, we need to minimize the CE in (4.5). We note that the estimator in (4.7) that estimates the CE (up to a multiplicative constant) is only different from the average negative log likelihood (the first term in (4.14)) by the weighting term, $h(\mathbf{X}_i)w(\mathbf{X}_i)$. Accordingly, the minimum cross-entropy estimator (MCE), $\hat{\boldsymbol{\theta}}$, that minimizes (4.7) shares the theoretical properties such as consistency and asymptotic normality of MLE, under

certain regularity conditions. More importantly, the similarity of MCE with MLE leads to a criterion for CE minimization, analogous to AIC. We call the new criterion cross-entropy information criterion (CIC). This criterion takes the following form

$$\mathrm{CIC} = \bar{\mathcal{C}}_{\hat{\boldsymbol{\theta}}} + K_{q^*}\frac{d}{n}, \tag{4.15}$$

where

$$\bar{\mathcal{C}}_{\hat{\boldsymbol{\theta}}} = -\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i)\log q\Big(\mathbf{X}_i;\hat{\boldsymbol{\theta}}\Big). \tag{4.16}$$

Here, $\bar{\mathcal{C}}_{\hat{\boldsymbol{\theta}}}$ is analogous to the first term in (4.14), i.e., the average negative log likelihood. The second term in (4.15) penalizes the model complexity by being linearly proportional to $d$, the dimension of $\hat{\boldsymbol{\theta}}$. Because (4.16) includes the weighting term, $h(\mathbf{X}_i)w(\mathbf{X}_i)$, the second term in (4.15) also includes

$$K_{q^*} = \mathbb{E}\left[h(\mathbf{X})w(\mathbf{X})\right] \tag{4.17}$$

so that the both terms in (4.15) can be balanced as in AIC.

Below we briefly explain the derivation of CIC. First, to establish the asymptotic unbiasedness of CIC in (4.15), we need two assumptions.

**Assumption IV.1.** *The optimal IS density is in the parametric family of* $q(\mathbf{x};\boldsymbol{\theta})$. *That is, there exists* $\boldsymbol{\theta}^*$ *such that* $q^*(\mathbf{x}) = q(\mathbf{x};\boldsymbol{\theta}^*)$.

**Assumption IV.2.** *Assume that* $2 \le \tau < \infty$, *where* $\tau$ *denotes the number of total CE iterations. As* $n \to \infty$, *simulation replications allocated to each CE iteration increase at the same rate.*

Under the stated assumptions and regularity conditions, the following theorem holds (see Appendix C).

**Theorem IV.3.**

$$\mathbb{E}\left[\bar{\mathcal{C}}_{\hat{\boldsymbol{\theta}}} - \mathcal{C}_{\hat{\boldsymbol{\theta}}}\right] = -K_{q^*}\frac{d}{n} + o\left(\frac{1}{n}\right), \tag{4.18}$$

*where the expectation is taken with respect to the data* $\mathbf{X}_1, \ldots, \mathbf{X}_n$*. The little-o term indicates that the approximation error goes to zero faster than the rate of* $1/n$*.*

Theorem IV.3 implies that the asymptotic bias of the estimator, $\bar{\mathcal{C}}_{\hat{\boldsymbol{\theta}}}$, in estimating $\mathcal{C}_{\hat{\boldsymbol{\theta}}}$ is $-K_{q^*}d/n$. Consequently, the CIC in (4.15) asymptotically corrects the bias and presents an asymptotically unbiased information criterion. As a result, among many possible choices of $k$, by finding the model order that minimizes the CIC in (4.15), we can find the best GMM that asymptotically minimizes the CE in (4.5). We also note that the bias correction term depending on $d$ prevents the overfitting, similar to AIC.

For illustration, Figure 4.1 shows a typical pattern of CIC observed in the DIS example in Section 4.4.1. As we use the GMM with unconstrained means and covariances, $d$ is $(k-1) + k(p + p(p+1)/2)$, where $p$ is the dimension of $\mathbf{X}$. Since $d$ is linearly proportional to $k$, we see that, as $k$ increases, CIC initially decreases and then levels off before increasing. As such, CIC guards against the overfitting. By minimizing CIC, we can find the best model that minimizes the CE in an asymptotically unbiased manner.

### 4.3.3 Approximations Necessary for Implementation

CIC in (4.15) involves $K_{q^*}$ in (4.17), which needs to be estimated in practice. For DIS, $\hat{P}_{DIS} = K_{q^*} + O_p(1/\sqrt{n})$ holds by the central limit theorem (*Keener*, 2010). Thus, the bias correction term derived in Theorem IV.3 remains valid when we use $\hat{P}_{DIS}$ in (4.2) as the estimator of $K_{q^*}$, $\hat{K}_{q^*}$. Similarly, for SIS, we use $\hat{P}_{SIS}$ in (4.3) as $\hat{K}_{q^*}$ .

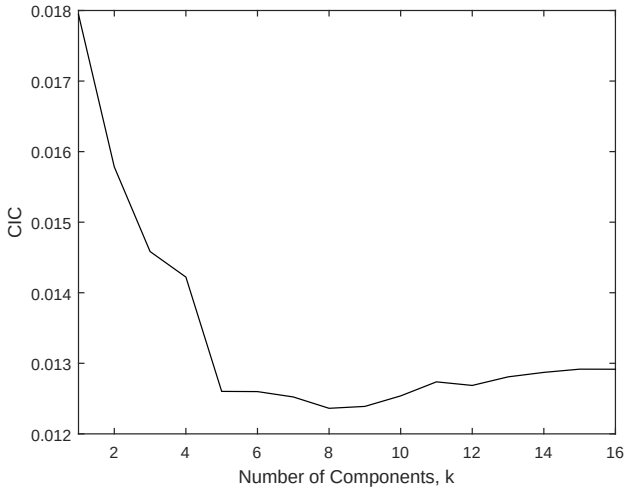



Figure 4.1: CIC observed in the DIS example in Section 4.4.1

To implement the proposed method, we need to know $h(\mathbf{x})$ in the EM algorithm equations (4.10)-(4.13) to find MCE, $\hat{\boldsymbol{\theta}}$ and in CIC in (4.15). For DIS, $h(\mathbf{x}) = \mathbb{I}(g(\mathbf{x}) > l)$ can be evaluated exactly at each $\mathbf{x}$ by running the simulation. However, for SIS, $h(\mathbf{x}) = \sqrt{s(\mathbf{x})(1-s(\mathbf{x}))/n + s(\mathbf{x})^2}$ needs to be estimated because $s(\mathbf{X}_i)$ is unknown. We estimate $s(\mathbf{X}_i)$ by

$$\hat{s}(\mathbf{X}_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbb{I}\left(Y_j^{(i)} > l\right) \tag{4.19}$$

and then estimate $h(\mathbf{x})$ by plugging in $\hat{s}(\mathbf{X}_i)$.

SIS also needs to allocate $N_i$ replications at each $\mathbf{X}_i$, as explained in Section 4.2. For a large $n \gg \max_{i=1}^{m}(1-s(\mathbf{X}_i))/s(\mathbf{X}_i)$, the optimal $N_i$ in (3.3) is approximately proportional to $\sqrt{w(\mathbf{X}_i) - \hat{P}_{SIS}}$ (see Appendix C). Thus, we decide $N_i$ based on this approximation. If $w(\mathbf{X}_i) - \hat{P}_{SIS} \leq 0$, we assign $N_i = 1$, to ensure the unbiasedness of $\hat{P}_{SIS}$ in (4.3).

### 4.3.4 Aggregated Failure Probability Estimation

Finally, to estimate the failure probability, we aggregate the samples obtained in all of the CE iterations. For DIS, instead of $\hat{P}_{DIS}$ in (4.2), we use

$$\hat{P}_{DIS'} = \frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{n^{(t)}}\sum_{i=1}^{n^{(t)}}\mathbb{I}\Big(Y_i^{(t)} > l\Big)\,\frac{f\Big(\mathbf{X}_i^{(t)}\Big)}{q\Big(\mathbf{X}_i^{(t)};\hat{\boldsymbol{\theta}}^{(t)}\Big)}, \tag{4.20}$$

where the superscript $^{(t)}$ denotes the $t$th CE iteration. Similarly, for SIS, instead of $\hat{P}_{SIS}$ in (4.3), we use

$$\hat{P}_{SIS'} = \frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{m^{(t)}}\sum_{i=1}^{m^{(t)}}\frac{1}{N_i^{(t)}}\sum_{j=1}^{N_i^{(t)}}\mathbb{I}\Big(Y_{ij}^{(t)} > l\Big)\,\frac{f\Big(\mathbf{X}_i^{(t)}\Big)}{q\Big(\mathbf{X}_i^{(t)};\hat{\boldsymbol{\theta}}^{(t)}\Big)}. \tag{4.21}$$

It should be noted that, by (4.20) and (4.21), we make further improvements over the standard CE method discussed in Section 4.2.1. The standard CE method does not use the intermediate CE sampling results for the failure probability estimation. Instead, the standard CE method uses $\hat{P}_{DIS}$ in (4.2) with the data obtained in the final iteration only.

### 4.3.5 Summary of the Proposed Method

For DIS, we use the following pseudo-code:

1. Set the iteration counter, $t = 1$. Sample $\mathbf{X}_i^{(t)}, i = 1, \ldots, n^{(t)}$ from an initial distribution (e.g., $f$).

2. At each $\mathbf{X}_i^{(t)}$, run the simulators and obtain the dataset $\mathcal{D}^{(t)} = \{(\mathbf{X}_i^{(t)}, Y_i^{(t)}) : i = 1, \ldots, n^{(t)}\}$.

3. If $t < \tau$, run the EM algorithm in Section 4.3.1 to find $\hat{\boldsymbol{\theta}}(k)$ for $k = k_{\min}, \ldots, k_{\max}$ and choose $k^* = \operatorname{argmin}_k \mathrm{CIC}(k)$, where $\mathrm{CIC}(k)$ in (4.15) is computed using $\hat{\boldsymbol{\theta}}(k)$, $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(t)}$ and $\hat{K}_{q^*} = \hat{P}_{DIS'}$ in (4.20). Otherwise, go to Step 5.

4. Increase $t$ by 1. Sample $\mathbf{X}_i^{(t)}, i = 1, \ldots, n^{(t)}$ from $q(\mathbf{x}; \hat{\boldsymbol{\theta}}(k^*))$ in (4.8). Go to Step 2.

5. Estimate the failure probability by $\hat{P}_{DIS'}$ in (4.20).

In Step 3, $\hat{K}_{q^*}$ is estimated using all the data available up to the current iteration, $t$. $k_{\min}$ and $k_{\max}$ can be flexibly chosen to find $k^*$ that minimizes CIC. For SIS, the pseudo-code is essentially the same except that the dataset $\mathcal{D}^{(t)}$ is constructed by running the simulator $N_i^{(t)}$ times at $\mathbf{X}_i^{(t)}, i = 1, \ldots, m^{(t)}$ and that we use $\hat{P}_{SIS'}$ in (4.21) instead of $\hat{P}_{DIS'}$. Hereafter, we call the proposed method EM-based cross-entropy (EMCE) method.

## 4.4 Numerical Examples

### 4.4.1 DIS Example

The closest work to ours is done by *Kurtz and Song* (2013), who use the GMM with a pre-specified value for $k$. Their method, called 'cross-entropy-based adaptive IS using Gaussian mixture (CE-AIS-GM)' is tested in *Kurtz and Song* (2013) using a classical example of the structural safety literature. In this example, the failure region is defined as $\{\mathbf{x} \in \mathbb{R}^2 : g(\mathbf{x}) \leq 0\}$, where

$$g(\mathbf{x}) = b - x_2 - \kappa (x_1 - e)^2 . \tag{4.22}$$

For comparison of CE-AIS-GM and EMCE, we vary the parameter $b = 1.5$, 2.0 and 2.5, to test three different failure thresholds. We fix the other two parameters, $\kappa = 0.1$ and $e = 0$ to maintain the shape of the failure region. We use the same sample size used in *Kurtz and Song* (2013), namely, the total of 8700 replications. As in *Kurtz and Song* (2013), CE-AIS-GM is set to use $k = 30$, whereas EMCE automatically chooses $k$.

Table 4.1 shows the estimation results based on 500 experiment repetitions. The sample mean of the failure probability estimates ('Mean') decreases as the threshold, $b$, increases. EMCE leads to at least twice smaller standard errors than CE-AIS-GM. This improvement of accuracy translates into computational saving: 'CMC Ratio' is the number of replications used in each row's method divided by the number of replications necessary for CMC in (4.1) to achieve the same standard error in the row. Although CE-AIS-GM saves significantly compared to CMC, EMCE saves even more by 4 to 6 times.

Table 4.1: Comparison between CE-AIS-GM and EMCE

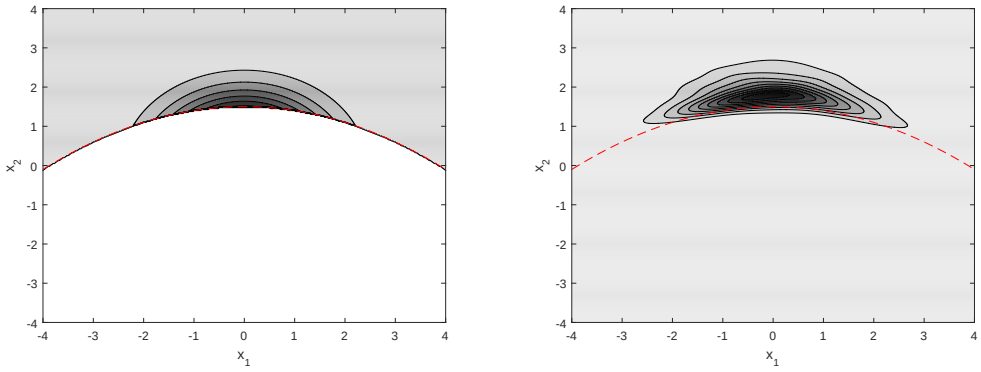
| $b$ | Method | Mean | Standard Error | CMC Ratio |
|---|---|---|---|---|
| 1.5 | CE-AIS-GM | 0.082902 | 0.001145 | 15.00% |
| | EMCE | 0.082911 | 0.000506 | 2.93% |
| 2.0 | CE-AIS-GM | 0.030174 | 0.000526 | 8.23% |
| | EMCE | 0.030173 | 0.000213 | 1.35% |
| 2.5 | CE-AIS-GM | 0.008908 | 0.000211 | 4.39% |
| | EMCE | 0.008910 | 0.000099 | 0.97% |

Figure 4.2 compares the theoretically optimal density in (2.3) and the EMCE density, for $b = 1.5$. We observe that the EMCE density with automatically chosen $k = 10$ is close to the theoretically optimal density, capturing the shape of important region.

### 4.4.2 SIS Example

For SIS, we test EMCE with the numerical example in Section 2.5 of Chapter II. Its data generating structure is as follows:

$$X \sim \mathcal{N}(0,1)\,, \qquad Y|X \sim \mathcal{N}\big(\mu(X)\,, \sigma^2(X)\big)\,,$$




(a) Optimal density (b) EMCE density ($k = 10$)

Figure 4.2: Comparison between the theoretically optimal density in (2.3) and the EMCE density, for the DIS example with $b = 1.5$. The red dashed line is the failure boundary, $g(\mathbf{x}) = 0$.

where the mean and the standard deviation are

$$\mu(X) = 0.95X^2\left(1 + 0.5\cos(5X) + 0.5\cos(10X)\right),$$
$$\sigma(X) = 1 + 0.7\,|X| + 0.4\cos(X) + 0.3\cos(14X).$$

To approximate the optimal density in (2.12) and the allocation in (2.13), Chapter II approximates the conditional probability, $s(X) = \mathbb{P}(Y > l \mid X)$, by using a metamodel. The metamodel is set as the normal distribution with the following mean and standard deviation:

$$\hat{\mu}(X) = 0.95X^2, \quad \hat{\sigma}(X) = 1 + 0.7\,|X|\,.$$

The total number of replications is set as 1000 for each method and the experiment is repeated 500 times to obtain the results in Table 4.2. Table 4.2 also shows the result from the optimal SIS that uses the true $s(X)$. It appears that EMCE is better than the metamodel-based approach, which captures the overall pattern of the true model, and close to the optimal SIS.

Table 4.2: Comparison between Metamodel-based SIS, EMCE, and the optimal SIS

| Method | Mean | Standard Error | CMC Ratio |
|---|---|---|---|
| Metamodel | 0.01011 | 0.00122 | 15.09% |
| EMCE | 0.00972 | 0.00073 | 8.65% |
| True Model | 0.00996 | 0.00052 | 2.74% |

## 4.5 Case Study

In the case study, we evaluate the reliability of a wind turbine using the same simulation setting described in Section 2.6 of Chapter II. In the previous chapters, we apply the metamodel-based SIS to this problem. This section compares the performance of EMCE with the metamodel-based SIS. For both methods, we use the same number of total replications, 1000 (2000) for the edgewise (flapwise) bending moment.

Table 4.3 compares the results based on 50 repetitions. EMCE has slightly smaller (larger) standard error than the metamodel-based approach for the edgewise (flapwise) bending moment. Accordingly, both methods save the similar level of computational resource compared to CMC, as indicated by 'CMC Ratio'.

Table 4.3: Comparison between the metamodel-based SIS and the EMCE for the case study

| Response | Method | Mean | Standard Error | CMC Ratio |
|---|---|---|---|---|
| Edgewise | Metamodel | 0.0486 | 0.0018 | 7.0% |
| | EMCE | 0.0486 | 0.0015 | 4.9% |
| Flapwise | Metamodel | 0.0514 | 0.0028 | 32% |
| | EMCE | 0.0535 | 0.0030 | 37% |

In the metamodel-based SIS, recall that the metamodel is carefully built by fitting a nonhomogeneous generalized extreme value distribution to the pilot data in Chapter II. As such, we can see that the performance of EMCE is comparable to that of metamodel-based SIS with a high quality metamodel. However, as seen in Section 4.4.2, when the metamodel quality is not good enough, EMCE provides a better computational efficiency. Since EMCE is an automated method, it can be a

promising method when building a metamodel is difficult.

## 4.6 Summary

We propose a method called EMCE that uses the EM algorithm to improve the updating scheme of the CE method. Furthermore, we derive an asymptotically unbiased model selection criterion, called CIC, to automatically find the model order that minimizes the cross-entropy between the optimal IS density and the candidate IS density. The numerical examples and case study demonstrate the superior performance of EMCE over the standard CE method and show the advantage of EMCE over the metamodel-based IS.

# CHAPTER V

# Conclusion

This dissertation develops three approaches to tackle computational challenges associated with reliability evaluation using stochastic simulation models. The computational challenges arise mainly from the fact that: (a) a simulation model, which accurately represents a stochastic system with millions of random variables, tends to be computationally expensive, (b) it is necessary to repeat running the simulation model many times to observe rare events which are critical for understanding system reliability, and (c) a conventional estimator of the rare event probability based on CMC is subject to large uncertainty, requiring sufficient enough simulation replications to observe several rare events in order to ensure a reasonable accuracy of the estimator.

Chapter II proposes SIS as the main solution approach for saving the computational resources when stochastic simulation models are used to estimate the probability of a failure event which occurs rarely. The goal of this chapter is to devise methods to optimally use stochastic simulation models under computational budget constraints. The proposed methods, SIS1 and SIS2, have the optimal properties of minimizing the variances of two different failure probability estimators. The first method, SIS1, prescribes how to optimally sample simulation inputs and allocate simulation resources at each sampled input, given the total number of simulation repli-

cations and the number of inputs to sample. The second method, SIS2, optimizes sampling efforts when only a single replication is allocated at each sampled input. Our numerical studies and case studies show that the performances of both methods are similar, while significantly outperforming the conventional method, CMC, and another benchmark method, BIS, which is similar to the existing IS designed for deterministic simulation models.

Chapter III develops computationally efficient approaches to quantify the uncertainty of SIS-based failure probability estimation. Such uncertainty quantification is important for reliability evaluation because any estimator based on stochastic simulations is subject to randomness and a highly uncertain estimator can be misleading in evaluating the system reliability. In Chapter II, to measure the uncertainty (variability or variance) of a failure probability estimator, we repeat obtaining failure probability estimates and compute the sample standard deviation. Such repetitions multiply the computational burden when obtaining a single estimate is already computationally expensive. Chapter III establishes CLTs for SIS1 and SIS2 and constructs asymptotic CIs for the failure probability estimation without repeating the estimation. Numerical studies validate that the resulting CIs indeed quantify the estimation uncertainty accurately. Case studies demonstrate the usefulness of having the CIs for the reliability evaluation of a wind turbine.

Chapter IV proposes a novel information criterion, CIC, to enhance the CE method that adaptively guides simulation process in efficiently estimating the failure probability of a system. The standard CE method has been widely used in practice for reliability evaluation with deterministic simulation models. The standard approach uses a parametric distribution, usually in the exponential family, as the IS distribution that focuses sampling efforts on important simulation inputs to improve the estimation accuracy. Because of the rigidity of common parametric distributions, recent studies propose using the mixture of the parametric distributions to have the

flexible shape of IS distribution. To the best of our knowledge, none of the studies, however, provides a rigorous approach to determine the number of component distributions in the mixture. We devise CIC that has a desirable asymptotic property to enable us to decide a good number of components in the mixture density, based on the information at hand. We use the EM algorithm, which minimizes an estimate of CE for the distribution parameter estimation, for the mixture-based CE method, and show that this method is applicable not only to deterministic simulation models but also to stochastic simulation models. Our numerical studies and case studies demonstrate that the proposed approach performs comparably or better than the benchmark methods we consider.

In the future, it would be interesting to investigate the methods that estimate a very small probability in the binomial distribution to improve the estimation of the conditional POE in stochastic simulation models. A new SIS method can be also developed to optimize a simulation experiment for evaluating the reliability associated with multiple responses. The resulting estimator will need an accompanying approach to quantify the estimation uncertainty, extending the work in Chapter III. Important extensions of CIC proposed in Chapter IV includes adopting the Bayesian paradigm and devising a CE-based criterion that is analogous to BIC (*Schwarz*, 1978) or an advanced criterion like in *Figueiredo and Jain* (2002), which may improve the stability and performance of EM algorithm.

The proposed approaches in this dissertation are applied to the reliability evaluation of a wind turbine in the case studies. We, however, expect that the methodologies are widely applicable to various domains. For structural safety evaluation in the civil engineering, IS has been used extensively to improve the reliability evaluation accuracy in civil infrastructure systems (*Dubourg et al.*, 2013; *Kurtz and Song*, 2013). Because uncertainty is a very important dimension to consider in many safety-critical systems, the results presented in this dissertation will benefit those who use stochastic

simulation models to evaluate the system safety. Finance is another area where rare events are of significant interests (*Wang and Zhou*, 2015). The proposed approaches in this dissertation that consider rare events under large uncertainties will benefit researchers and practitioners in finance to improve the estimation accuracy, quantify the associated uncertainty, and adaptively guide simulation process for efficient use of computational resources.

# APPENDICES

# APPENDIX A

# Appendix for Chapter II

In this Appendix, Section A.1 includes the derivations for the optimal solutions of SIS1 and SIS2 with the multivariate input vector. Sections A.2 and A.3 present the numerical examples with the univariate input variable and the multivariate input vector, respectively, which are used to investigate the impacts of various factors on the performances of the proposed methods. Section A.4 discusses the implementation details with the wind turbine simulators.

## A.1 Derivations for Optimal SIS

This section details the derivations of the optimal allocation size, $N_i, i = 1, \cdots, M$, and the optimal IS density, $q_{SIS1}$, for SIS1 and the optimal IS density, $q_{SIS2}$, for SIS2, presented in Section 2.3. In the sequel, we consider the multivariate input vector, $\mathbf{X} \in \mathbb{R}^p$. Note that the univariate input variable is a special case with $p = 1$.

### A.1.1 Optimal importance sampling density and allocations in SIS1

First, we consider the SIS1 estimator,

$$\begin{aligned}\hat{P}_{SIS1} &= \frac{1}{M}\sum_{i=1}^{M}\hat{P}\left(Y > l \mid \mathbf{X}_i\right)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)} \\ &= \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{I}\left(Y_j^{(i)} > l\right)\right)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}.\end{aligned}$$

We decompose the variance of this estimator into two components, the expectation of the conditional variance and the variance of the conditional expectation, as

$$\begin{aligned}Var\left[\hat{P}_{SIS1}\right] &= Var\left[\frac{1}{M}\sum_{i=1}^{M}\hat{P}\left(Y > l \mid \mathbf{X}_i\right)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}\right] \\ &= \frac{1}{M^2}E_q\left[Var\left[\sum_{i=1}^{M}\hat{P}\left(Y > l \mid \mathbf{X}_i\right)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}\;\middle|\;\mathbf{X}_1,\cdots,\mathbf{X}_M\right]\right] \\ &\quad + \frac{1}{M^2}Var_q\left[E\left[\sum_{i=1}^{M}\hat{P}\left(Y > l \mid \mathbf{X}_i\right)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}\;\middle|\;\mathbf{X}_1,\cdots,\mathbf{X}_M\right]\right], \quad \text{(A.1)}\end{aligned}$$

where the subscript $q$ appended to $E$ or $Var$ indicates that the expectation or the variance is taken with respect to $q$. For simplicity, let $s(\mathbf{X})$ denote the conditional POE, $P\left(Y > l \mid \mathbf{X}\right)$. Using the fact that $\mathbf{X}_i \overset{i.i.d}{\sim} q$ for $i = 1, 2, \cdots, M$, we simplify $Var\left[\hat{P}_{SIS1}\right]$ in (A.1) to

$$\begin{aligned}Var\left[\hat{P}_{SIS1}\right] &= \frac{1}{M^2}E_q\left[Var\left[\sum_{i=1}^{M}\left(\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{I}\left(Y_j^{(i)} > l\right)\right)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}\;\middle|\;\mathbf{X}_1,\cdots,\mathbf{X}_M\right]\right] \\ &\quad + \frac{1}{M^2}Var_q\left[\sum_{i=1}^{M}s(\mathbf{X}_i)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}\right] \\ &= \frac{1}{M^2}E_q\left[\sum_{i=1}^{M}\left(\frac{1}{N_i^2}\sum_{j=1}^{N_i}s(\mathbf{X}_i)\left(1-s(\mathbf{X}_i)\right)\right)\frac{f(\mathbf{X}_i)^2}{q(\mathbf{X}_i)^2}\right] + \frac{1}{M}Var_q\left[s(\mathbf{X})\frac{f(\mathbf{X})}{q(\mathbf{X})}\right] \\ &= \frac{1}{M^2}E_q\left[\sum_{i=1}^{M}\frac{1}{N_i}s(\mathbf{X}_i)\left(1-s(\mathbf{X}_i)\right)\frac{f(\mathbf{X}_i)^2}{q(\mathbf{X}_i)^2}\right] + \frac{1}{M}Var_q\left[s(\mathbf{X})\frac{f(\mathbf{X})}{q(\mathbf{X})}\right].\end{aligned}$$

We express the allocation size, $N_i$, at $\mathbf{X}_i$ as a proportion of the total simulation budget, $N_T$,

$$N_i = N_T \cdot \frac{c(\mathbf{X}_i)}{\sum_{j=1}^{M} c(\mathbf{X}_j)}, \qquad i = 1, 2, \cdots, M, \tag{A.2}$$

where $c(\mathbf{X})$ is a non-negative function. Lemma II.1 presents the optimal assignment of simulation replications, $N_i$, to each $\mathbf{X}_i$ for any given $q$.

**Lemma II.1** *Given $q$, the variance in (2.9) is minimized if and only if*

$$N_i = \frac{\sqrt{s(\mathbf{X}_i)\left(1 - s(\mathbf{X}_i)\right)} f(\mathbf{X}_i) / q(\mathbf{X}_i)}{\sum_{j=1}^{M} \sqrt{s(\mathbf{X}_j)\left(1 - s(\mathbf{X}_j)\right)} f(\mathbf{X}_j) / q(\mathbf{X}_j)} \cdot N_T \quad \text{for } i = 1, 2, \cdots, M.$$

***Proof.*** *We want to find $N_i$, $i = 1, 2, \cdots, M$, that minimizes the variance in (2.9) for any given density function, $q(\mathbf{X})$. Note that the second term in (2.9) is constant, provided that the function $q(\mathbf{X})$ is given, and the other functions, $f(\mathbf{X})$ and $s(\mathbf{X})$, are fixed. Thus, we find $N_i$ that minimizes the first term in (2.9),*

$$\begin{aligned}
&\frac{1}{M^2} E_q \left[ \sum_{i=1}^{M} \frac{1}{N_i} s(\mathbf{X}_i)\left(1 - s(\mathbf{X}_i)\right) \frac{f(\mathbf{X}_i)^2}{q(\mathbf{X}_i)^2} \right] \\
&= \frac{1}{M^2} \sum_{i=1}^{M} E_q \left[ \frac{1}{N_i} s(\mathbf{X}_i)\left(1 - s(\mathbf{X}_i)\right) \frac{f(\mathbf{X}_i)^2}{q(\mathbf{X}_i)^2} \right] \\
&= \frac{1}{M} E_q \left[ \frac{1}{N_1} s(\mathbf{X}_1)\left(1 - s(\mathbf{X}_1)\right) \frac{f(\mathbf{X}_1)^2}{q(\mathbf{X}_1)^2} \right] && \text{(A.3)} \\
&= \frac{1}{M} \frac{1}{N_T} E_q \left[ \frac{\sum_{j=1}^{M} c(\mathbf{X}_j)}{c(\mathbf{X}_1)} s(\mathbf{X}_1)\left(1 - s(\mathbf{X}_1)\right) \frac{f(\mathbf{X}_1)^2}{q(\mathbf{X}_1)^2} \right] && \text{(A.4)} \\
&= \frac{1}{M} \frac{1}{N_T} \left( \sum_{j=1}^{M} E_q \left[ \frac{c(\mathbf{X}_j)}{c(\mathbf{X}_1)} s(\mathbf{X}_1)\left(1 - s(\mathbf{X}_1)\right) \frac{f(\mathbf{X}_1)^2}{q(\mathbf{X}_1)^2} \right] \right) \\
&= \frac{1}{M} \frac{1}{N_T} \left( E_q \left[ s(\mathbf{X}_1)\left(1 - s(\mathbf{X}_1)\right) \frac{f(\mathbf{X}_1)^2}{q(\mathbf{X}_1)^2} \right] + \sum_{j=2}^{M} E_q \left[ \frac{c(\mathbf{X}_j)}{c(\mathbf{X}_1)} s(\mathbf{X}_1)\left(1 - s(\mathbf{X}_1)\right) \frac{f(\mathbf{X}_1)^2}{q(\mathbf{X}_1)^2} \right] \right) \\
&= \frac{1}{M} \frac{1}{N_T} \left( E_q \left[ s(\mathbf{X})\left(1 - s(\mathbf{X})\right) \frac{f(\mathbf{X})^2}{q(\mathbf{X})^2} \right] + (M-1) \cdot E_q \left[ c(\mathbf{X}) \right] \cdot E_q \left[ \frac{1}{c(\mathbf{X})} s(\mathbf{X})\left(1 - s(\mathbf{X})\right) \frac{f(\mathbf{X})^2}{q(\mathbf{X})^2} \right] \right) && \text{(A.5)} \\
&\geq \frac{1}{M} \frac{1}{N_T} \left( E_q \left[ s(\mathbf{X})\left(1 - s(\mathbf{X})\right) \frac{f(\mathbf{X})^2}{q(\mathbf{X})^2} \right] + (M-1) \cdot \left( E_q \left[ \sqrt{s(\mathbf{X})\left(1 - s(\mathbf{X})\right)} \frac{f(\mathbf{X})}{q(\mathbf{X})} \right] \right)^2 \right) && \text{(A.6)}
\end{aligned}$$

*The equalities in* (A.*3*) *and* (A.*5*) *are due to the fact that* $\mathbf{X}_i$, $i = 1, 2, \cdots, M$, *is independent and identically distributed. We use the definition in* (A.*2*) *for* (A.*4*). *The inequality in* (A.6) *follows by applying the Cauchy-Schwarz inequality to the second term in* (A.*5*). *The equality in* (A.6) *holds if and only if*

$$c(\mathbf{X}) = k\sqrt{s(\mathbf{X})\left(1 - s(\mathbf{X})\right)} f(\mathbf{X}) / q(\mathbf{X}),$$

*where* $k$ *is a positive constant. Therefore, by the definition in* (A.2), *the optimal allocation size in (2.10) follows.* □

Plugging $N_i$'s in (2.10) into the estimator variance in (2.9) leads to

$$\begin{aligned} Var\left[\hat{P}_{SIS1}\right] &= \frac{1}{M}\frac{1}{N_T}\left(E_q\left[s(\mathbf{X})\left(1 - s(\mathbf{X})\right)\frac{f(\mathbf{X})^2}{q(\mathbf{X})^2}\right] + (M-1)\left(E_f\left[\sqrt{s(\mathbf{X})\left(1 - s(\mathbf{X})\right)}\right]\right)^2\right) \\ &\quad + \frac{1}{M} Var_q\left[s(\mathbf{X})\frac{f(\mathbf{X})}{q(\mathbf{X})}\right] \qquad \text{(A.7)} \\ &= \frac{1}{M}\frac{1}{N_T}\left(E_f\left[s(\mathbf{X})\left(1 - s(\mathbf{X})\right)\frac{f(\mathbf{X})}{q(\mathbf{X})}\right] + (M-1)\left(E_f\left[\sqrt{s(\mathbf{X})\left(1 - s(\mathbf{X})\right)}\right]\right)^2\right) \\ &\quad + \frac{1}{M}\left(E_q\left[s(\mathbf{X})^2\frac{f(\mathbf{X})^2}{q(\mathbf{X})^2}\right] - \left(E_q\left[s(\mathbf{X})\frac{f(\mathbf{X})}{q(\mathbf{X})}\right]\right)^2\right) \\ &= \frac{1}{M}\frac{1}{N_T}\left(E_f\left[s(\mathbf{X})\left(1 - s(\mathbf{X})\right)\frac{f(\mathbf{X})}{q(\mathbf{X})}\right] + (M-1)\left(E_f\left[\sqrt{s(\mathbf{X})\left(1 - s(\mathbf{X})\right)}\right]\right)^2\right) \\ &\quad + \frac{1}{M}\left(E_f\left[s(\mathbf{X})^2\frac{f(\mathbf{X})}{q(\mathbf{X})}\right] - P\left(Y > l\right)^2\right), \qquad \text{(A.8)} \end{aligned}$$

where we obtain the equation in (A.7) using the expression in (A.6). Please note that $E_q\left[\sqrt{s(\mathbf{X})\left(1 - s(\mathbf{X})\right)}\frac{f(\mathbf{X})}{q(\mathbf{X})}\right] = E_f\left[\sqrt{s(\mathbf{X})\left(1 - s(\mathbf{X})\right)}\right]$.

Recall that $s(\mathbf{X})$ denotes $P\left(Y > l \mid \mathbf{X}\right)$. Thus, only the following terms in (A.8) contain $q$,

$$\begin{aligned} &\frac{1}{M}\frac{1}{N_T}E_f\left[s(\mathbf{X})\left(1 - s(\mathbf{X})\right)\frac{f(\mathbf{X})}{q(\mathbf{X})}\right] + \frac{1}{M}E_f\left[s(\mathbf{X})^2\frac{f(\mathbf{X})}{q(\mathbf{X})}\right] \\ &= \frac{1}{M}\int_{\mathcal{X}_f}\left(\frac{1}{N_T}s(\mathbf{x})\cdot\left(1 - s(\mathbf{x})\right) + s(\mathbf{x})^2\right)\frac{f^2(\mathbf{x})}{q(\mathbf{x})}\,\mathrm{d}\mathbf{x}, \qquad \text{(A.9)} \end{aligned}$$

where $\mathcal{X}_f = \{\mathbf{x} \in \mathbb{R}^p : f\left(\mathbf{x}\right) > 0\}$ is the support of $f$. Finding $q$ that minimizes (A.9)

is a functional minimization problem. To specify the boundary conditions, we define the joint cumulative distribution function (CDF) of $\mathbf{X} \in \mathbb{R}^p$ with the IS density, $q$, as

$$Q\left(x_1, x_2, \cdots, x_p\right) \equiv \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_p} q\left(\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_p\right) \mathrm{d}\tilde{x}_1 \mathrm{d}\tilde{x}_2 \cdots \mathrm{d}\tilde{x}_p.$$

Then, we impose the boundary conditions,

$$Q\left(-\infty, -\infty, \cdots, -\infty\right) = 0,$$
$$Q\left(\infty, \infty, \cdots, \infty\right) = 1.$$

Therefore, we minimize the functional in (A.9) over the set of functions,

$$\{q : Q\left(-\infty, -\infty, \cdots, -\infty\right) = 0;\ Q\left(\infty, \infty, \cdots, \infty\right) = 1;\ q\left(\mathbf{x}\right) \geq 0, \forall \mathbf{x} \in \mathbb{R}^p\}.$$

In the following, we use principles of the calculus of variations. The integrand in (A.9) is the Lagrangian function, $\mathcal{L}(x_1, x_2, \cdots, x_p, q)$. The optimal $q$ should satisfy the Euler-Lagrange equation (*Courant and Hilbert*, 1989),

$$\begin{aligned} 0 &= (-1)^p \frac{\partial^p}{\partial x_1 \partial x_2 \cdots \partial x_p} \left( \frac{\partial \mathcal{L}}{\partial q}(x_1, x_2, \cdots, x_p, q) \right) \\ &= (-1)^p \frac{\partial^p}{\partial x_1 \partial x_2 \cdots \partial x_p} \left( -\frac{\mathcal{L}(x_1, x_2, \cdots, x_p, q)}{q(x_1, x_2, \cdots, x_p)} \right). \end{aligned}$$

This Euler-Lagrange equation is satisfied if the function $q$ satisfies

$$C_{q1}^2 = \left( \frac{1}{N_T} s(\mathbf{x}) \left(1 - s(\mathbf{x})\right) + s(\mathbf{x})^2 \right) \frac{f^2(\mathbf{x})}{q^2\left(\mathbf{x}\right)},$$

where $C_{q1}$ is a positive constant. Rearranging the above equation gives

$$q(\mathbf{x}) = \frac{1}{C_{q1}} f\left(\mathbf{x}\right) \sqrt{\frac{1}{N_T} s(\mathbf{x}) \left(1 - s(\mathbf{x})\right) + s(\mathbf{x})^2}. \tag{A.10}$$

This function $q$ also satisfies the boundary conditions on $Q$ by setting $C_{q1}$ to satisfy the normalizing constraint of the joint IS density, $q$, as follows:

$$\begin{aligned} C_{q1} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x})\sqrt{\frac{1}{N_T}s(\mathbf{x})\cdot(1-s(\mathbf{x}))+s(\mathbf{x})^2}\,\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_p \\ &\equiv \int_{\mathcal{X}_f} f(\mathbf{x})\sqrt{\frac{1}{N_T}s(\mathbf{x})\cdot(1-s(\mathbf{x}))+s(\mathbf{x})^2}\,\mathrm{d}\mathbf{x}. \end{aligned} \tag{A.11}$$

To guarantee that the resulting $q$ is a minimizer of the functional in (A.9), we verify that the following second variation (*Courant and Hilbert*, 1989) is positive definite,

$$J[Q;R] = \int_{\mathcal{X}_q} R^2\frac{\partial^2\mathcal{L}}{\partial Q^2} + 2Rr\frac{\partial^2\mathcal{L}}{\partial Q\partial q} + r^2\frac{\partial^2\mathcal{L}}{\partial q^2}\,\mathrm{d}\mathbf{x}, \tag{A.12}$$

where $\mathcal{X}_q = \{\mathbf{x}\in\mathbb{R}^p : q(\mathbf{x}) > 0\}$ is the support of $q$. The function, $R(x_1, x_2, \cdots, x_p)$, in (A.12) represents a variation that should satisfy the boundary conditions,

$$\begin{aligned} R(-\infty,-\infty,\cdots,-\infty) &= 0, \\ R(\infty,\infty,\cdots,\infty) &= 0, \end{aligned}$$

so that the varied function, $\tilde{Q}(x_1, x_2, \cdots, x_p) \equiv Q(x_1, x_2, \cdots, x_p) + R(x_1, x_2, \cdots, x_p)$, satisfies the prescribed boundary conditions,

$$\begin{aligned} \tilde{Q}(-\infty,-\infty,\cdots,-\infty) &= 0, \\ \tilde{Q}(\infty,\infty,\cdots,\infty) &= 1. \end{aligned}$$

The function, $r(x_1, x_2, \cdots, x_p)$, in (A.12) is

$$r(x_1, x_2, \cdots, x_p) \equiv \frac{\partial^p R}{\partial x_1\partial x_2\cdots\partial x_p}(x_1, x_2, \cdots, x_p).$$

We note that

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial Q^2}(x_1, x_2, \cdots, x_p, q) &= 0, \\
\frac{\partial^2 \mathcal{L}}{\partial Q \partial q}(x_1, x_2, \cdots, x_p, q) &= 0, \\
\frac{\partial^2 \mathcal{L}}{\partial q^2}(x_1, x_2, \cdots, x_p, q) &= 2\left(\frac{1}{N_T} s(\mathbf{x})\left(1 - s(\mathbf{x})\right) + s(\mathbf{x})^2\right) \frac{f^2(\mathbf{x})}{q^3(\mathbf{x})} \\
&> 0 \text{ for all } \mathbf{x} \in \mathcal{X}_q = \{\tilde{\mathbf{x}} \in \mathbb{R}^p : q(\tilde{\mathbf{x}}) > 0\}.
\end{aligned}$$

Therefore, the second variation in (A.12) is reduced to

$$J[Q; R] = \int_{\mathcal{X}_q} r^2 \frac{\partial^2 \mathcal{L}}{\partial q^2} \, \mathrm{d}\mathbf{x},$$

where $\frac{\partial^2 \mathcal{L}}{\partial q^2}$ is positive. Therefore, $J[Q; R]$ vanishes if and only if $r(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}_q$. The latter condition implies that $R(\mathbf{x})$ is a constant function of 0, since $R(\mathbf{x})$ is 0 at $(x_1, x_2, \cdots, x_p) = (-\infty, -\infty, \cdots, -\infty)$ and $(x_1, x_2, \cdots, x_p) = (\infty, \infty, \cdots, \infty)$. Therefore, for all allowable nonzero variations, $R(\mathbf{x})$, the second variation is positive definite (i.e., $J[Q; R] > 0$). This verifies that the IS density, $q$, in (A.10) with the normalizing constant in (A.11) is the minimizing function of the variance in (A.8). We also plug this $q$ into (2.10) to obtain the optimal allocation size, which leads to Theorem II.2.

**Theorem II.2** *(a) The variance of the estimator in* (2.6) *is minimized if the following IS density and the allocation size are used.*

$$\begin{aligned}
q_{SIS1}(\mathbf{x}) &= \frac{1}{C_{q1}} f(\mathbf{x}) \sqrt{\frac{1}{N_T} s(\mathbf{x})\left(1 - s(\mathbf{x})\right) + s(\mathbf{x})^2}, \\
N_i &= N_T \frac{\sqrt{\frac{N_T(1 - s(\mathbf{x}_i))}{1 + (N_T - 1)s(\mathbf{x}_i)}}}{\sum_{j=1}^{M} \sqrt{\frac{N_T(1 - s(\mathbf{x}_j))}{1 + (N_T - 1)s(\mathbf{x}_j)}}}, \quad i = 1, 2, \cdots, M,
\end{aligned}$$

*where* $C_{q1}$ *is* $\int_{\mathcal{X}_f} f(\mathbf{x}) \sqrt{\frac{1}{N_T} s(\mathbf{x}) \cdot (1 - s(\mathbf{x})) + s(\mathbf{x})^2}\, \mathrm{d}\mathbf{x}$ *and* $s(\mathbf{x})$ *is* $P(Y > l | \mathbf{X} = \mathbf{x})$.

*(b) With* $q_{SIS1}$ *and* $N_i, i = 1, 2, \cdots, M$, *the estimator in* (2.6) *is unbiased.*

***Proof.*** *(a) We already derived the optimal* $q_{SIS1}$ *in* (2.12) *from the above discussion. Plugging the optimal* $q_{SIS1}$ *into the formula of* $N_i$ *in* (2.10) *gives*

$$\begin{aligned}
N_i &\propto \sqrt{s(\mathbf{x}_i)\,(1 - s(\mathbf{x}_i))} \frac{f(\mathbf{x}_i)}{q_{SIS1}(\mathbf{x}_i)} \\
&= \sqrt{s(\mathbf{x}_i)\,(1 - s(\mathbf{x}_i))} f(\mathbf{x}_i) \left( \frac{1}{C_{q1}} f(\mathbf{x}_i) \sqrt{\frac{1}{N_T} s(\mathbf{x}_i)\,(1 - s(\mathbf{x}_i)) + s(\mathbf{x}_i)^2} \right)^{-1} \\
&\propto \sqrt{\frac{s(\mathbf{x}_i)\,(1 - s(\mathbf{x}_i))}{\frac{1}{N_T} s(\mathbf{x}_i)\,(1 - s(\mathbf{x}_i)) + s(\mathbf{x}_i)^2}} \\
&= \sqrt{\frac{N_T (1 - s(\mathbf{x}_i))}{1 - s(\mathbf{x}_i) + N_T s(\mathbf{x}_i)}} \\
&= \sqrt{\frac{N_T (1 - s(\mathbf{x}_i))}{1 + (N_T - 1)\, s(\mathbf{x}_i)}}.
\end{aligned}$$

*By imposing the normalizing constraint of* $N_T = \sum_{i=1}^{M} N_i$, *the expression of the optimal allocation size in* (2.13) *follows.*

*(b) The estimator in* (2.6) *is unbiased if* $q_{SIS1}(\mathbf{x}_i) = 0$ *implies*

$$\begin{aligned}
\hat{P}\,(Y > l \mid \mathbf{X} = \mathbf{x}_i)\, f(\mathbf{x}_i) &= \left( \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbb{I}\left( Y_j^{(i)} > l \right) \right) f(\mathbf{x}_i) \\
&= 0
\end{aligned}$$

*for any* $\mathbf{x}_i$. *Note that* $q_{SIS1}(\mathbf{x}) = 0$ *holds only if* $f(\mathbf{x}) = 0$ *or* $s(\mathbf{x}) = 0$. *If* $s(\mathbf{x}) = 0$, *then* $\hat{P}\,(Y > l | \mathbf{X} = \mathbf{x}) = 0$. *Therefore, if* $q_{SIS1}(\mathbf{x}) = 0$, *then* $\hat{P}\,(Y > l | \mathbf{X} = \mathbf{x})\, f(\mathbf{x}) = 0$, *which concludes the proof.* □

### A.1.2 Optimal importance sampling density in SIS2

Now we consider the SIS2 estimator with a multivariate input vector, $\mathbf{X} \in \mathbb{R}^p$,

$$\hat{P}_{SIS2} = \frac{1}{N_T} \sum_{i=1}^{N_T} \mathbb{I}\left(Y_i > l\right) \frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)},$$

where $Y_i$ is an output at $\mathbf{X}_i$, $i = 1, 2, \cdots, N_T$. Theorem II.3 presents the optimal IS density, $q$, for the estimator in (2.7). Similar to the derivation of $q_{SIS1}$ in (2.12), we first decompose the estimator variance and apply the principles of the calculus of variation.

**Theorem II.3** *(a) The variance of the estimator in (2.7) is minimized with the density,*

$$q_{SIS2}(\mathbf{x}) = \frac{1}{C_{q2}} \sqrt{s(\mathbf{x})} f(\mathbf{x}) \,,$$

*where* $C_{q2}$ *is* $\int_{\mathcal{X}_f} \sqrt{s(\mathbf{x})} f(\mathbf{x}) \, \mathrm{d}\mathbf{x}$ *and* $s(\mathbf{x})$ *is* $P(Y > l | \mathbf{X} = \mathbf{x})$.

*(b) With* $q_{SIS2}$*, the estimator in (2.7) is unbiased.*

***Proof.*** *(a)*

$$
\begin{aligned}
Var\left[\hat{P}_{SIS2}\right] &= Var\left[\frac{1}{N_T}\sum_{i=1}^{N_T}\mathbb{I}\left(Y_i > l\right)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}\right] \\
&= \frac{1}{N_T^2}E_q\left[Var\left[\sum_{i=1}^{N_T}\mathbb{I}\left(Y_i > l\right)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}\;\middle|\;\mathbf{X}_1,\cdots,\mathbf{X}_{N_T}\right]\right] \\
&\quad + \frac{1}{N_T^2}Var_q\left[E\left[\sum_{i=1}^{N_T}\mathbb{I}\left(Y_i > l\right)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}\;\middle|\;\mathbf{X}_1,\cdots,\mathbf{X}_{N_T}\right]\right] \\
&= \frac{1}{N_T^2}E_q\left[\sum_{i=1}^{N_T}s(\mathbf{X}_i)\cdot\left(1-s(\mathbf{X}_i)\right)\frac{f(\mathbf{X}_i)^2}{q(\mathbf{X}_i)^2}\right] \\
&\quad + \frac{1}{N_T^2}Var_q\left[\sum_{i=1}^{N_T}s(\mathbf{X}_i)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}\right] \\
&= \frac{1}{N_T}E_q\left[s(\mathbf{X})\cdot\left(1-s(\mathbf{X})\right)\frac{f(\mathbf{X})^2}{q(\mathbf{X})^2}\right] \\
&\quad + \frac{1}{N_T}\left(E_q\left[s(\mathbf{X})^2\frac{f(\mathbf{X})^2}{q(\mathbf{X})^2}\right] - \left(E_q\left[s(\mathbf{X})\frac{f(\mathbf{X})}{q(\mathbf{X})}\right]\right)^2\right) \\
&= \frac{1}{N_T}E_q\left[s(\mathbf{X})\frac{f(\mathbf{X})^2}{q(\mathbf{X})^2}\right] - \frac{1}{N_T}\left(E_q\left[s(\mathbf{X})\frac{f(\mathbf{X})}{q(\mathbf{X})}\right]\right)^2 \\
&= \frac{1}{N_T}E_f\left[s(\mathbf{X})\frac{f(\mathbf{X})}{q(\mathbf{X})}\right] - \frac{1}{N_T}P\left(Y > l\right)^2. \qquad \text{(A.13)}
\end{aligned}
$$

*To find the optimal IS density which minimizes the functional in (*A.*13), we apply the similar procedure discussed for SIS1. Since only the first term of* (A.13) *involves* $q$*, we consider the following Lagrangian function,*

$$\mathcal{L}(\mathbf{x}, q) = s(\mathbf{x})\frac{f^2(\mathbf{x})}{q(\mathbf{x})}.$$

*Note that the Lagrangian function for SIS2 replaces*

$$\left(\frac{1}{N_T}s(\mathbf{x})\cdot\left(1-s(\mathbf{x})\right) + s(\mathbf{x})^2\right)$$

*in the Lagrangian function for SIS1 (i.e., the integrand in (*A.*9)) with* $s(\mathbf{x})$*. Therefore,*

*the Euler-Lagrange equation and the second variation for SIS2 are analogous to those for SIS1. They lead to the minimizing function in* (2.14) *for SIS2, which is also analogous to the minimizing function in* (2.12) *for SIS1.*

*(b) The estimator in* (2.7) *is unbiased if* $q_{SIS2}(\mathbf{x}) = 0$ *implies* $\mathbb{I}(Y > l) f(\mathbf{x}) = 0$ *for any* $\mathbf{x}$*. Note that* $Y$ *is an output corresponding to* $\mathbf{x}$*.* $q_{SIS2}(\mathbf{x}) = 0$ *holds only if* $f(\mathbf{x}) = 0$ *or* $s(\mathbf{x}) = 0$*. Also, if* $s(\mathbf{x}) = 0$*, then* $\mathbb{I}(Y > l) = 0$*. Therefore, it follows that* $\mathbb{I}(Y > l) f(\mathbf{x}) = 0$ *if* $q_{SIS2}(\mathbf{x}) = 0$*.* $\square$

## A.2 Univariate Example

To design a univariate stochastic example, we take a deterministic simulation example in *Cannamela et al.* (2008) and modify it to have stochastic elements. Specifically, we have the following data generating structure:

$$X \sim N(0, 1),$$
$$Y|X \sim N\big(\mu(X), \sigma^2(X)\big),$$

where the mean, $\mu(X)$, and the standard deviation, $\sigma(X)$, are

$$\mu(X) = 0.95\delta X^2 \left(1 + 0.5\cos(10\kappa X) + 0.5\cos(20\kappa X)\right),$$
$$\sigma(X) = 1 + 0.7\,|X| + 0.4\cos(X) + 0.3\cos(14X),$$

respectively. The metamodel of the conditional distribution, $Y|X$, is

$$N\big(\hat{\mu}(X), \hat{\sigma}^2(X)\big),$$

where

$$\hat{\mu}(X) = 0.95\beta\delta X^2 \left(1 + 0.5\rho\cos(10\kappa X) + 0.5\rho\cos(20\kappa X)\right),$$
$$\hat{\sigma}(X) = 1 + 0.7\,|X| + 0.4\rho\cos(X) + 0.3\rho\cos(14X).$$

We vary the following parameters to test different aspects of our proposed methods compared to alternative methods.

- $P_T$, the magnitude of target failure probability: By varying $P_T = P(Y > l)$, where $l$ is a threshold for the system failure, we want to see how the proposed methods perform at different levels of $P_T$. Based on 1 million CMC simulation replications, we decide $l$ that corresponds to the target failure probability, $P_T$. We consider the three levels of $P_T$, namely, 0.10, 0.05, and 0.01.

- $\delta$, the difference between the original input density, $f$, and the optimal IS density, $q_{SIS1}$ (or $q_{SIS2}$): We want to investigate how the computational gains of SIS1 and SIS2 change when the optimal IS density is more different from the original input density, $f$. Note that the original input density, $f$, is a standard normal density with a mode at 0. When $\delta = 1$, $q_{SIS1}$ and $q_{SIS2}$ will focus their sampling efforts on the input regions far from 0, since the response variable, $Y$, tends to be large in such regions due to the term, $0.95X^2$, in $\mu(X)$. Conversely, when $\delta = -1$, $q_{SIS1}$ and $q_{SIS2}$ will focus their sampling efforts on the regions close to 0.

- $\rho$, the metamodeling accuracy for the oscillating pattern: We vary $\rho$ in $\hat{\mu}(X)$ and $\hat{\sigma}(X)$ to control the quality of the metamodel in capturing the oscillating pattern of the true model with $\mu(X)$ and $\sigma(X)$. We consider $\rho$ of 0, 0.5, and 1. When $\rho = 1$, the metamodel mimics the oscillating pattern perfectly in both the mean and standard deviation, whereas $\rho = 0$ means that the metamodel

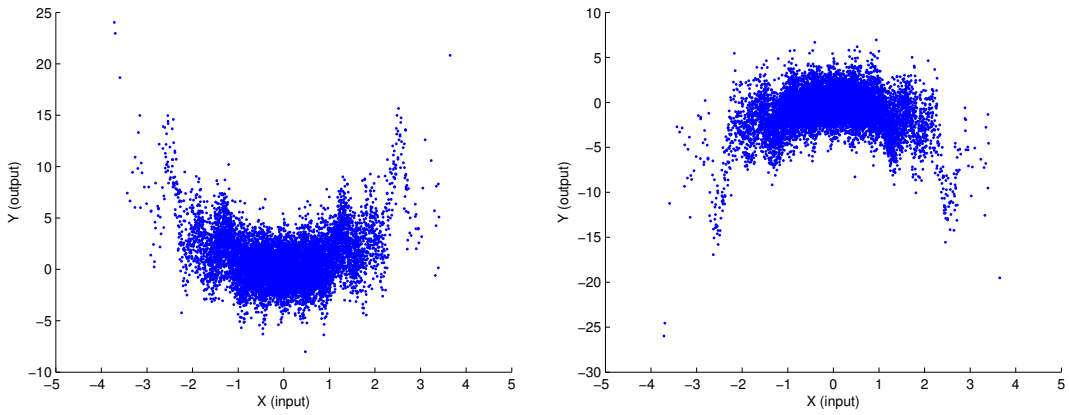
fails to capture the oscillating pattern.

- $\beta$, the metamodeling accuracy for the global pattern: We consider a variation of $\beta$ in $\hat{\mu}(X)$ with five levels, $\beta = 0.90$, 0.95, 1, 1.05, and 1.10. Note that when $\beta = 1$ (and $\rho = 1$), the metamodel perfectly mimics the true model.

- $M/N_T$, the ratio of the input sample size to the total number of simulation replications: We consider various choices of $M/N_T$ including 10%, 30%, 50%, 70%, and 90% to see how sensitive the performance of SIS1 is to the choice of $M/N_T$.

- $\kappa$, the locality (or roughness, nonlinearity) of the location function, $\mu(X)$: We consider the three levels of $\kappa = 0$, 0.5, and 1. When $\kappa$ is far from zero, the cosine terms in $\mu(X)$ add locality, roughness, or nonlinearity to the shape of $\mu(X)$. On the other hand, when $\kappa = 0$, the location function, $\mu(X)$, simply becomes a quadratic function of $X$.

We use the following setup as a baseline and vary each parameter to see its effect on the performances of the proposed methods: $P_T = 0.01$, $\delta = 1$, $M/N_T = 30\%$, $\rho = 1$, $\beta = 1$, and $\kappa = 0.5$. Figure A.1 shows the scatter plots at the baseline setup with $\delta = 1$ and $-1$.




(a) $\delta = 1$ (b) $\delta = -1$

Figure A.1: Scatter plots of the baseline univariate example with different $\delta$

We set $N_T$, the total simulation replications, as 1,000. To obtain the sample average and standard error of each method's POE estimation, we repeat the experiment 500 times.

### A.2.1 Effects of $P_T$ and $\delta$

Table A.1 summarizes the effects of $P_T$ and $\delta$. Except these two, we keep the other parameters at their baseline values. We use the perfect metamodel (i.e., $\rho = 1$, $\beta = 1$) so that we can examine the main effect of $P_T$ and $\delta$ without any interaction effects with the metamodel quality.

We compute the relative ratio, $N_T/N_T^{(CMC)}$, as follows. Let $N_T^{(CMC)}$ denote the number of CMC simulation replications to achieve the same standard error of each method in the table. With $N_T^{(CMC)}$ replications, the standard error of the CMC failure probability estimator is $\sqrt{P_T\left(1-P_T\right)/N_T^{(CMC)}}$. Table A.1 shows that the relative ratios of SIS1 and SIS2 are comparable to each other and clearly better than BIS, and that they generally decrease as $P_T$ gets smaller. That is, the efficiencies of the SIS methods against CMC improve as $P_T$ gets close to zero. For example, when $\delta = 1$ and $P_T$ are 0.10, 0.05, and 0.01, SIS1 requires 51%, 32%, and 2.5% of the CMC simulation efforts to achieve the same estimation accuracy, respectively (in other words, CMC needs about twice, three times, and forty times more simulation efforts than SIS1, respectively.) These remarkable computational savings are also observed in our case study with the wind turbine simulators (see Table 2.5). Specifically, SIS1 and SIS2 respectively lead to 4.9% and 5.9% of the relative ratios for edgewise bending moments with $l$ = 9,300 $kNm$. Note that the corresponding sample averages, namely 0.00992 and 0.01005, are close to the failure probability of $P_T = 0.01$.

Table A.1 also shows that the computational gains of SIS1 and SIS2 are much more significant when $\delta = 1$ (i.e., when $f$ and $q_{SIS1}$ (or $q_{SIS2}$) are quite different) than when $\delta = -1$. This finding is intuitive and also consistent with the observation

in the wind turbine simulation that the computational gains of SIS1 and SIS2 for the edgewise bending moments are much more remarkable than those for the flapwise bending moments. Interestingly, when $\delta = -1$, BIS has no advantage over CMC, whereas the proposed methods still lead to lower standard errors than CMC.

Table A.1: POE estimation results with different $\delta$ and $P_T$

| Method | | $\delta = 1$ | | | $\delta = -1$ | | |
|---|---|---|---|---|---|---|---|
| | | $P_T$ | | | $P_T$ | | |
| | | 0.10 | 0.05 | 0.01 | 0.10 | 0.05 | 0.01 |
| SIS1 | Sample Average | 0.1004 | 0.0502 | 0.0100 | 0.1001 | 0.0500 | 0.0100 |
| | Standard Error | 0.0068 | 0.0039 | 0.0005 | 0.0090 | 0.0062 | 0.0026 |
| | Relative Ratio | 51% | 32% | 2.5% | 90% | 81% | 68% |
| SIS2 | Sample Average | 0.0999 | 0.0501 | 0.0100 | 0.1001 | 0.0500 | 0.0099 |
| | Standard Error | 0.0069 | 0.0042 | 0.0006 | 0.0086 | 0.0064 | 0.0028 |
| | Relative Ratio | 53% | 37% | 3.6% | 82% | 86% | 79% |
| BIS | Sample Average | 0.1002 | 0.0505 | 0.0101 | 0.1009 | 0.0503 | 0.0101 |
| | Standard Error | 0.0089 | 0.0068 | 0.0014 | 0.0095 | 0.0067 | 0.0031 |
| | Relative Ratio | 88% | 97% | 20% | 100% | 95% | 97% |
| CMC | Sample Average | 0.1005 | 0.0506 | 0.0100 | 0.1005 | 0.0498 | 0.0100 |
| | Standard Error | 0.0092 | 0.0070 | 0.0030 | 0.0096 | 0.0071 | 0.0031 |

### A.2.2 Effects of metamodel accuracy

Now, we consider how computational efficiency varies when the metamodel accuracy changes. First, we study the effect of $\rho$, the metamodeling accuracy for the oscillating pattern. We keep all other parameters at their baseline values. The results in Table A.2 suggest that the standard errors for SIS1, SIS2, and BIS increase as $\rho$ decreases (i.e., the metamodel quality deteriorates). However, the standard errors for both SIS1 and SIS2 increase more slowly than for BIS as $\rho$ decreases. Also, SIS1 and SIS2 produce lower standard errors than BIS by 50-85% and CMC by 40-85%. Interestingly, the increase of the SIS2's standard error is minimal, indicating that SIS2 is the least sensitive to the metamodel quality. The performance of BIS differs significantly depending on the metamodel quality, and BIS generates an even higher standard error than CMC when $\rho = 0$.

Table A.2: POE estimation results with different $\rho$

| Method | | $\rho$ 1.00 | 0.50 | 0 |
|---|---|---|---|---|
| SIS1 | Sample Average | 0.0100 | 0.0100 | 0.0101 |
| | Standard Error | 0.0005 | 0.0008 | 0.0017 |
| SIS2 | Sample Average | 0.0100 | 0.0101 | 0.0100 |
| | Standard Error | 0.0006 | 0.0007 | 0.0010 |
| BIS | Sample Average | 0.0101 | 0.0100 | 0.0102 |
| | Standard Error | 0.0014 | 0.0018 | 0.0063 |
| CMC | Sample Average | 0.0099 | 0.0099 | 0.0099 |
| | Standard Error | 0.0030 | 0.0030 | 0.0030 |

Second, we consider the effect of $\beta$, the metamodeling accuracy for the global pattern. We keep all other parameters at their baseline values. The results in Table A.3 do not show any clear patterns to explain the impact of the metamodel accuracy of the global pattern on the performances of SIS1 and SIS2. However, in all cases, SIS1 and SIS2 outperform BIS and CMC, reducing the standard errors by 45-70% and 80-85%, respectively.

Table A.3: POE estimation results with different $\beta$

| Method | | $\beta$ 0.90 | 0.95 | 1.00 | 1.05 | 1.10 |
|---|---|---|---|---|---|---|
| SIS1 | Sample Average | 0.0101 | 0.0101 | 0.0100 | 0.0101 | 0.0101 |
| | Standard Error | 0.0005 | 0.0005 | 0.0005 | 0.0005 | 0.0005 |
| SIS2 | Sample Average | 0.0101 | 0.0100 | 0.0100 | 0.0100 | 0.0101 |
| | Standard Error | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0006 |
| BIS | Sample Average | 0.0101 | 0.0100 | 0.0101 | 0.0101 | 0.0101 |
| | Standard Error | 0.0013 | 0.0016 | 0.0014 | 0.0013 | 0.0011 |
| CMC | Sample Average | 0.0100 | 0.0100 | 0.0099 | 0.0100 | 0.0099 |
| | Standard Error | 0.0031 | 0.0031 | 0.0030 | 0.0030 | 0.0030 |

Third, we investigate the effect of the metamodel quality on the computational gains of the proposed methods as the failure probability gets smaller, when the metamodel is poor. Specifically, we consider the cases of $(\rho = 0.5, \beta = 1)$, $(\rho = 0, \beta = 0.6)$, and $(\rho = 0, \beta = 1.2)$. We keep all other parameters at their baseline values. Table A.4 shows that the computational efficiencies of SIS1 and SIS2 are substantially better

than BIS in all cases. Similar to the pattern in Table A.2, SIS2 tends to perform better than SIS1 when the metamodel is inaccurate, and when $P_T$ changes from 0.10 to 0.01, the efficiencies of SIS1 and SIS2 improve remarkably. However, we note that there are some cases (e.g., SIS1 with $\rho = 0, \beta = 1.2$ and SIS2 with $\rho = 0, \beta = 0.6$) where the efficiency slightly diminishes when $P_T$ changes from 0.10 to 0.05. This result indicates that if the metamodel is inaccurate, the efficiencies of SIS1 and SIS2 do not necessarily improve when smaller $P_T$ is estimated. Even so, SIS1 and SIS2 perform much better than BIS.

Table A.4: POE estimation results with different $\rho$ and $\beta$

| Method | | $\rho = 0.5, \beta = 1$ | | | $\rho = 0, \beta = 0.6$ | | | $\rho = 0, \beta = 1.2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $P_T$ | | | $P_T$ | | | $P_T$ | | |
| | | 0.10 | 0.05 | 0.01 | 0.10 | 0.05 | 0.01 | 0.10 | 0.05 | 0.01 |
| SIS1 | Ave. | 0.0998 | 0.0503 | 0.0100 | 0.0998 | 0.0503 | 0.0100 | 0.1001 | 0.0503 | 0.0102 |
| | S.E. | 0.0080 | 0.0046 | 0.0008 | 0.0104 | 0.0066 | 0.0016 | 0.0120 | 0.0090 | 0.0024 |
| | Ratio | 71% | 44% | 6.4% | 120% | 91% | 26% | 160% | 170% | 58% |
| SIS2 | Ave. | 0.0999 | 0.0503 | 0.0101 | 0.0999 | 0.0506 | 0.0100 | 0.0993 | 0.0503 | 0.0101 |
| | S.E. | 0.0068 | 0.0045 | 0.0007 | 0.0082 | 0.0064 | 0.0009 | 0.0078 | 0.0054 | 0.0010 |
| | Ratio | 51% | 42% | 4.9% | 75% | 86% | 8.1% | 67% | 61% | 10% |
| BIS | Ave. | 0.1007 | 0.0502 | 0.0100 | 0.1014 | 0.0493 | 0.0105 | 0.1028 | 0.0511 | 0.0105 |
| | S.E. | 0.0134 | 0.0078 | 0.0018 | 0.0355 | 0.0086 | 0.0082 | 0.0665 | 0.0184 | 0.0104 |
| | Ratio | 199% | 128% | 32% | 1398% | 155% | 673% | 4905% | 710% | 1082% |
| CMC | Ave. | 0.1004 | 0.0506 | 0.0099 | 0.1005 | 0.0504 | 0.0100 | 0.1001 | 0.0504 | 0.0099 |
| | S.E. | 0.0091 | 0.0071 | 0.0030 | 0.0093 | 0.0071 | 0.0030 | 0.0093 | 0.0070 | 0.0030 |

*Notes:* 'Ave.' denotes the sample average, 'S.E.' denotes the standard error, and 'Ratio' denotes the relative ratio of $N_T/N_T^{CMC}$.

### A.2.3 Effects of the ratio, $M/N_T$

Here, we want to see how sensitive SIS1 is to the choice of $M/N_T$. We keep all other parameters at their baseline values. The results in Table A.5 suggest that the standard error of the SIS1 estimator is generally insensitive to the choice of $M/N_T$. This result is consistent with the result of the wind turbine simulations. Note that the standard error in Table A.5 is presented up to 5 digits (not 4 digits).

Table A.5: Effect of different $M/N_T$ ratios in the univariate example

| $M/N_T$ | Sample Average | Standard Error |
|---|---|---|
| 10% | 0.0100 | 0.00055 |
| 30% | 0.0100 | 0.00050 |
| 50% | 0.0101 | 0.00059 |
| 70% | 0.0101 | 0.00063 |
| 90% | 0.0100 | 0.00076 |

### A.2.4 Effects of locality, $\kappa$

We consider the effect of $\kappa$, the locality (or roughness, nonlinearity) of the location function, $\mu(X)$. We keep all other parameters at their baseline values. The results in Table A.6 show that the standard errors for SIS1 and SIS2 slightly increase as $\kappa$ increases. However, regardless of $\kappa$, SIS1 and SIS2 outperform BIS and CMC, lowering the standard errors by 30-65% and 75-90%, respectively.

Table A.6: POE estimation results with different $\kappa$

| Method | | $\kappa$ = 0 | $\kappa$ = 0.50 | $\kappa$ = 1.00 |
|---|---|---|---|---|
| SIS1 | Sample Average | 0.0100 | 0.0100 | 0.0100 |
| | Standard Error | 0.0004 | 0.0005 | 0.0007 |
| SIS2 | Sample Average | 0.0100 | 0.0100 | 0.0101 |
| | Standard Error | 0.0005 | 0.0006 | 0.0007 |
| BIS | Sample Average | 0.0100 | 0.0101 | 0.0100 |
| | Standard Error | 0.0008 | 0.0014 | 0.0010 |
| CMC | Sample Average | 0.0100 | 0.0099 | 0.0099 |
| | Standard Error | 0.0031 | 0.0030 | 0.0031 |

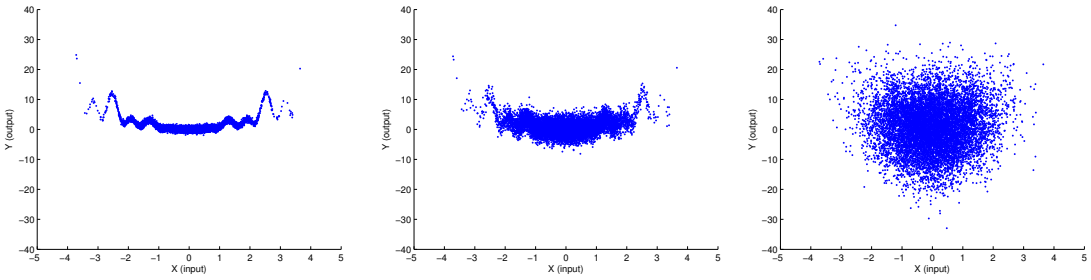
### A.2.5 Effects of variation of $\epsilon$

Theoretically, SIS1 and SIS2 are reduced to DIS when the simulator is deterministic. Recall that the standard error for DIS with $q_{DIS}$ is zero. Thus, in a stochastic computer model, if the uncontrollable randomness represented by $\epsilon$ has a smaller level of variation, then the standard errors for SIS1 and SIS2 will get closer to zero. We conduct a numerical study to illustrate the impact of the variance of $\epsilon$. We consider

the same data generating structure as before except that the variance of $\epsilon$ does not depend on the input, $X$, but is constant:

$$\sigma^2(X) = \tau^2.$$

Equivalently, we consider $Y = \mu(X) + \epsilon$, where $\epsilon$ follows a normal distribution with mean zero and standard deviation, $\tau$. We use the optimal IS densities for SIS1 and SIS2 with the perfect knowledge of $s(X)$. We consider $\tau$ of 0.5, 1, 2, 4, and 8. In Figure A.2, we can see the scatter plots of $Y$ versus $X$ for $\tau$ of 0.5, 2, and 8, by which the variation of $Y$ given $X$ is controlled. We set all other parameters at their baseline values: $P_T = 0.01$, $\delta = 1$, $M/N_T = 30\%$, and $\kappa = 0.5$.




(a) $\tau = 0.5$ (b) $\tau = 2$ (c) $\tau = 8$

Figure A.2: Scatter plots of the baseline case with different $\tau$

Table A.7 shows that as $\tau$ gets close to zero (please see from right to left), so do the standard errors of SIS1 and SIS2. The results indicate that the proposed methods practically reduce to DIS, since the optimal DIS density makes the standard error zero for the deterministic simulation (i.e., the case with $\tau = 0$).

Also, Figure A.3 illustrates that the optimal SIS1 and SIS2 densities are almost the same as the BIS density when the variation of $\epsilon$ is very small (in the figure, we use $\tau = 0.5$). Since the BIS density theoretically reduces to the DIS density for deterministic simulation and closely mimics the DIS density when $\tau$ is negligibly small, we can see that the proposed methods practically reduce to DIS when the

Table A.7: POE estimation results with different $\tau$

| Method | | $\tau$ 0.50 | 1.00 | 2.00 | 4.00 | 8.00 |
|---|---|---|---|---|---|---|
| SIS1 | Sample Average | 0.0102 | 0.0101 | 0.0101 | 0.0102 | 0.0100 |
| | Standard Error | 0.0001 | 0.0001 | 0.0005 | 0.0021 | 0.0028 |
| SIS2 | Sample Average | 0.0102 | 0.0101 | 0.0101 | 0.0104 | 0.0100 |
| | Standard Error | 0.0001 | 0.0002 | 0.0006 | 0.0023 | 0.0028 |
| BIS | Sample Average | 0.0102 | 0.0101 | 0.0100 | 0.0103 | 0.0101 |
| | Standard Error | 0.0002 | 0.0003 | 0.0010 | 0.0033 | 0.0033 |
| CMC | Sample Average | 0.0100 | 0.0100 | 0.0099 | 0.0101 | 0.0101 |
| | Standard Error | 0.0030 | 0.0031 | 0.0030 | 0.0032 | 0.0031 |

*Notes:* SIS1's standard errors for $\tau = 0.50$ and $\tau = 1.00$ are 0.00007 and 0.00013, respectively, in one more digit.

variation of the uncontrollable randomness is very small.

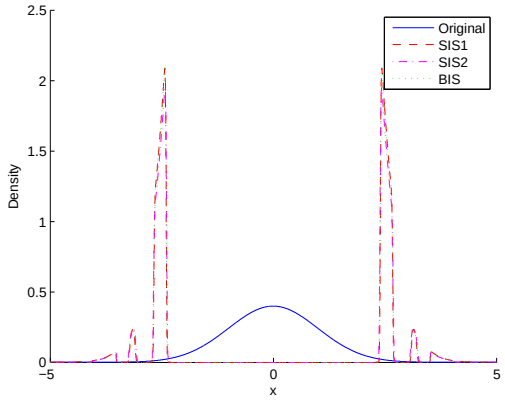



Figure A.3: Density plots for SIS1, SIS2, and BIS optimal densities when $\tau = 0.50$ along with the original input density

### A.2.6 Precision of numerical integration

When we use the numerical integration to compute the normalizing constant of an IS density, we make sure that the numerical precision is accurate enough so that the POE estimation accuracy is unaffected. We present POE estimation results up to 5 digits after the decimal point. Given that we bound the numerical error by -7 orders of magnitude or smaller, the numerical integration does not contribute to the error of POE estimation. To check the precision, we also conduct numerical studies

with the same data generating structure used in Section A.2.5. In Table A.8, the sample averages and standard errors are based on 500 POE estimates. The POEs in the last column are estimated by CMC with 100 million replications. We note that the estimated POE values from SIS1 and SIS2 are the same as the values from CMC.

Table A.8: POE estimation results for SIS1 and SIS2, compared to the POE estimated by CMC with 100 million replications, for different $\tau$

| $\tau$ | Sample Average (Standard Error) | | CMC |
|---|---|---|---|
| | SIS1 | SIS2 | |
| 0.50 | 0.0102 (0.0001) | 0.0102 (0.0001) | 0.0102 |
| 1.00 | 0.0101 (0.0001) | 0.0101 (0.0002) | 0.0101 |
| 2.00 | 0.0101 (0.0005) | 0.0101 (0.0006) | 0.0101 |

## A.3 Multivariate Example

We also design a multivariate stochastic example. We take an example in *Huang et al.* (2006), which adds a normal stochastic noise to a deterministic example originally in *Ackley* (1987). We slightly revise the example in *Huang et al.* (2006) by adding more complexity to the stochastic elements, and use the following data generating structure where the input vector, $\mathbf{X} = (X_1, X_2, X_3)$, follows a multivariate normal distribution:

$$\mathbf{X} \sim MVN(0, I_3)\,,$$
$$Y|\mathbf{X} \sim N\big(\mu(\mathbf{X})\,, \sigma^2(\mathbf{X})\big)\,,$$

where the mean, $\mu(\mathbf{X})$, and the standard deviation, $\sigma(\mathbf{X})$, are

$$\begin{aligned}\mu(\mathbf{X}) &= 20\delta\left(1-\exp\left(-0.2\sqrt{\frac{1}{3}\|\mathbf{X}\|^2}\right)\right)\\ &\quad+\delta\left(\exp(1)-\exp\left(\frac{1}{3}\sum_{i=1}^{3}\cos(2\pi\kappa X_i)\right)\right),\\ \sigma(\mathbf{X}) &= 1+0.7\sqrt{\frac{1}{3}\|\mathbf{X}\|^2}+0.4\left(\frac{1}{3}\sum_{i=1}^{3}\cos(3\pi X_i)\right),\end{aligned}$$

respectively. The output, $Y$, with the above $\mu(\mathbf{X})$ and $\sigma(\mathbf{X})$ presents a very complicated pattern over the input domain. The metamodel for the conditional distribution, $Y|\mathbf{X}$, is $N(\hat{\mu}(\mathbf{X}), \hat{\sigma}^2(\mathbf{X}))$, where

$$\begin{aligned}\hat{\mu}(\mathbf{X}) &= 20\beta\delta\left(1-\exp\left(-0.2\sqrt{\frac{1}{3}\|\mathbf{X}\|^2}\right)\right)+\rho\delta\left(\exp(1)-\exp\left(\frac{1}{3}\sum_{i=1}^{3}\cos(2\pi\kappa X_i)\right)\right),\\ \hat{\sigma}(\mathbf{X}) &= 1+0.7\sqrt{\frac{1}{3}\|\mathbf{X}\|^2}+0.4\rho\left(\frac{1}{3}\sum_{i=1}^{3}\cos(3\pi X_i)\right).\end{aligned}$$

The parameters in the above equations take similar roles in the univariate example. We use the same baseline setup we used in the univariate example, namely, $P_T = 0.01$, $\delta = 1$, $M/N_T = 30\%$, $\rho = 1$, $\beta = 1$, and $\kappa = 0.5$. We explain each parameter as follows.

- $P_T$, the magnitude of target failure probability: Based on 10 million CMC simulation replications, we decide $l$ that corresponds to the target failure probability, $P_T = P(Y > l)$. We consider the three levels of $P_T$, 0.10, 0.05, and 0.01.

- $\delta$, the difference between the original input density, $f$, and the optimal IS density, $q_{SIS1}$(or $q_{SIS2}$): We consider $\delta$ of 1 or $-1$. The densities, $f$ and $q_{SIS1}$ (or $q_{SIS2}$), are more different from each other when $\delta = 1$ than when $\delta = -1$. Note that the original input density, $f$, has the highest likelihood at the origin. When $\delta = 1$, $q_{SIS1}$ and $q_{SIS2}$ will focus their sampling efforts on the regions far from the origin, since the response variable, $Y$, tends to be large in such

regions due to the term, $20\left(1-\exp\left(-0.2\sqrt{\frac{1}{3}\|\mathbf{X}\|^2}\right)\right)$, in $\mu(\mathbf{X})$. Conversely, when $\delta = -1$, $q_{SIS1}$ and $q_{SIS2}$ will focus their sampling efforts on the regions close to the origin.

- $\rho$, the metamodeling accuracy for the oscillating pattern: We consider $\rho$ of 0, 0.5, and 1. When $\rho = 1$, the metamodel mimics the oscillating pattern perfectly, whereas $\rho = 0$ implies that the metamodel captures no oscillating term.

- $\beta$, the metamodeling accuracy for the global pattern: We consider $\beta = 0.95$, 1, and 1.05. Note that when $\beta = 1$ (and $\rho = 1$), the metamodel perfectly mimics the true model.

- $M/N_T$, the ratio of the input sample size to the total number of simulation replications: We consider $M/N_T$ of 10%, 30%, 50%, 70%, and 90%.

- $\kappa$, the locality (or roughness, nonlinearity) of the location function, $\mu(\mathbf{X})$: We consider the four levels of $\kappa$, 0, 0.5, 1, and 2. When $\kappa$ is far from zero, the cosine terms in $\mu(\mathbf{X})$ add locality, roughness, or nonlinearity to the shape of $\mu(\mathbf{X})$. On the other hand, when $\kappa = 0$, the location function, $\mu(\mathbf{X})$, simply becomes a monotonically increasing function of $\|\mathbf{X}\|$.

We set $N_T$, the total simulation replications, as 1,000. To obtain the sample average and standard error of each method's POE estimation, we repeat the experiment 2,000 times.

### A.3.1 Effects of $P_T$ and $\delta$

Table A.9 summarizes the effects of $P_T$ and $\delta$. We keep all other parameters at their baseline values. Similar to the univariate example, the experiment results suggest that the computational gains of SIS1 and SIS2 against CMC increase as $P_T$ gets smaller. We also see that the computational gains of SIS1 and SIS2 are more

significant when $\delta = 1$ (i.e., $f$ and $q_{SIS1}$ (or $q_{SIS2}$) are quite different) than when $\delta = -1$. In all cases, SIS1 and SIS2 perform better than BIS and CMC.

We note that when $\delta = 1$, the relative ratios of SIS1 and SIS2 decrease more slowly than the univariate input example results in Table A.1. Specifically, for $P_T = 0.01$, SIS1 and SIS2 yield 2.5% and 3.6% of the relative ratios in Table A.1; but, both methods give 29% of the relative ratio in Table A.9. We attribute such differences in the two example results to the differences in the data generating structures. The data generating structure of the univariate example in Section A.2 and that of the multivariate example in Section A.3 are different not only in the input dimension but also in the mean function, $\mu(\mathbf{x})$, and the standard deviation function, $\sigma(\mathbf{X})$. We detail this point in Section A.3.5.

Table A.9: POE estimation results with different $\delta$ and $P_T$ in the multivariate example

| Method | | $\delta = 1$ | | | $\delta = -1$ | | |
|---|---|---|---|---|---|---|---|
| | | $P_T$ | | | $P_T$ | | |
| | | 0.10 | 0.05 | 0.01 | 0.10 | 0.05 | 0.01 |
| SIS1 | Sample Average | 0.1002 | 0.0501 | 0.0100 | 0.1000 | 0.0500 | 0.0100 |
| | Standard Error | 0.0070 | 0.0046 | 0.0017 | 0.0072 | 0.0051 | 0.0020 |
| | Relative Ratio | 54% | 45% | 29% | 58% | 55% | 40% |
| SIS2 | Sample Average | 0.1002 | 0.0499 | 0.0100 | 0.1001 | 0.0499 | 0.0100 |
| | Standard Error | 0.0070 | 0.0048 | 0.0017 | 0.0078 | 0.0050 | 0.0020 |
| | Relative Ratio | 54% | 49% | 29% | 68% | 53% | 40% |
| BIS | Sample Average | 0.1000 | 0.0500 | 0.0100 | 0.1001 | 0.0500 | 0.0102 |
| | Standard Error | 0.0082 | 0.0062 | 0.0026 | 0.0096 | 0.0069 | 0.0036 |
| | Relative Ratio | 75% | 81% | 68% | 102% | 100% | 131% |
| CMC | Sample Average | 0.0997 | 0.0500 | 0.0101 | 0.0998 | 0.0499 | 0.0101 |
| | Standard Error | 0.0094 | 0.0069 | 0.0031 | 0.0093 | 0.0069 | 0.0031 |

### A.3.2 Effects of metamodel accuracy

We consider the effect of $\rho$, the metamodeling accuracy for the oscillating pattern. We keep all other parameters at their baseline values. Similar to the univariate example, Table A.10 shows that the standard errors of the SIS1 and SIS2 estimators

increase as $\rho$ decreases. Also, the standard error for SIS2 increases more slowly than that for SIS1, which shows that SIS2 is less sensitive than SIS1 to the metamodel quality. It appears that the performance of BIS is the most sensitive to the metamodel quality. In all cases, SIS1 and SIS2 lead to smaller standard errors than BIS and CMC by 20–60% and 20–50%, respectively.

Table A.10: POE estimation results with different $\rho$ in the multivariate example

| Method | | $\rho$ 1.00 | 0.50 | 0 |
|---|---|---|---|---|
| SIS1 | Sample Average | 0.0100 | 0.0101 | 0.0100 |
| | Standard Error | 0.0017 | 0.0019 | 0.0024 |
| SIS2 | Sample Average | 0.0100 | 0.0100 | 0.0099 |
| | Standard Error | 0.0016 | 0.0018 | 0.0020 |
| BIS | Sample Average | 0.0100 | 0.0100 | 0.0098 |
| | Standard Error | 0.0022 | 0.0040 | 0.0047 |
| CMC | Sample Average | 0.0101 | 0.0102 | 0.0101 |
| | Standard Error | 0.0031 | 0.0031 | 0.0031 |

*Notes.* At $\rho = 1$, standard errors for SIS1 and SIS2 are 0.00167 and 0.00163, respectively, in one more digit.

We consider the effect of $\beta$, the metamodeling accuracy for the global pattern. We keep all other parameters at their baseline values. Table A.11 shows that the standard errors of the SIS1 and SIS2 estimators do not vary significantly, so the performances of SIS1 and SIS2 are insensitive to the metamodeling accuracy for the global pattern in this example. In all cases, SIS1 and SIS2 outperform BIS and CMC, providing lower standard errors than BIS and CMC by 25–40% and 45-50%, respectively. .

### A.3.3 Effects of the ratio, $M/N_T$

We want to see how sensitive SIS1 is to the choice of $M/N_T$. We keep all other parameters at their baseline values. The results in Table A.12 suggest that the standard error of the SIS1 estimator is generally insensitive to the choice of $M/N_T$ as we observed in the univariate example and the wind turbine simulations. Note that the standard error is presented up to 5 digits (not 4 digits).

Table A.11: POE estimation results with different $\beta$ in the multivariate example

| Method | | $\beta$ | | |
|---|---|---|---|---|
| | | 0.95 | 1.00 | 1.05 |
| SIS1 | Sample Average | 0.0099 | 0.0099 | 0.0100 |
| | Standard Error | 0.0016 | 0.0017 | 0.0017 |
| SIS2 | Sample Average | 0.0100 | 0.0100 | 0.0100 |
| | Standard Error | 0.0017 | 0.0017 | 0.0017 |
| BIS | Sample Average | 0.0101 | 0.0100 | 0.0100 |
| | Standard Error | 0.0025 | 0.0023 | 0.0023 |
| CMC | Sample Average | 0.0101 | 0.0101 | 0.0101 |
| | Standard Error | 0.0031 | 0.0031 | 0.0031 |

Table A.12: Effect of different $M/N_T$ ratios in the multivariate example

| $M/N_T$ | Sample Average | Standard Error |
|---|---|---|
| 10% | 0.0100 | 0.00168 |
| 30% | 0.0100 | 0.00167 |
| 50% | 0.0100 | 0.00168 |
| 70% | 0.0100 | 0.00173 |
| 90% | 0.0100 | 0.00185 |

### A.3.4 Effects of locality, $\kappa$

We consider the effect of $\kappa$, the locality (or roughness, nonlinearity) of the location function, $\mu(\mathbf{X})$. We keep all other parameters at their baseline values. The results in Table A.13 suggest that $\kappa$ has little effect on the standard errors of the SIS1 and SIS2 estimators in this specific example. For all $\kappa$ values, SIS1 and SIS2 lead to smaller standard errors than BIS and CMC by 20–50% and 45–50%, respectively.

Table A.13: POE estimation results with different $\kappa$ in the multivariate example

| Method | | $\kappa$ | | | |
|---|---|---|---|---|---|
| | | 0 | 0.5 | 1 | 2 |
| SIS1 | Sample Average | 0.0099 | 0.0100 | 0.0101 | 0.0099 |
| | Standard Error | 0.0017 | 0.0017 | 0.0016 | 0.0016 |
| SIS2 | Sample Average | 0.0101 | 0.0100 | 0.0100 | 0.0100 |
| | Standard Error | 0.0017 | 0.0017 | 0.0016 | 0.0016 |
| BIS | Sample Average | 0.0101 | 0.0100 | 0.0101 | 0.0100 |
| | Standard Error | 0.0031 | 0.0022 | 0.0033 | 0.0031 |
| CMC | Sample Average | 0.0102 | 0.0101 | 0.0102 | 0.0103 |
| | Standard Error | 0.0031 | 0.0031 | 0.0031 | 0.0032 |

### A.3.5 Analysis with the univariate input

In Section A.3.1, as $P_T$ gets smaller, we observe that the relative ratios of SIS1 and SIS2 with $\delta = 1$ decrease more slowly than those in the univariate example (see Tables A.1 and A.9 with $\delta = 1$). These different patterns in the two numerical examples are mainly due to the different data generating structures not only in the input dimension but also in the mean function, $\mu(\mathbf{x})$, and the standard deviation function, $\sigma(\mathbf{X})$. For the univariate example in Section A.2, we take a deterministic simulation example in Cannamela *et al.* (*Cannamela et al.*, 2008) and modify it by adding stochastic elements to it, whereas for the multivariate example in Section A.3, we add a normal stochastic noise to a deterministic multivariate example originally in Ackley (*Ackley*, 1987). In the sequel, we call these univariate and multivariate examples as Cannamela1D and Ackley3D, respectively, based on their respective sources (*Cannamela et al.*, 2008; *Ackley*, 1987).

To clarify the different patterns in Cannamela1D and Ackley3D, we devise a new univariate example which is one-dimensional version of Ackley3D, and we call this new example as Ackley1D. Specifically, we consider the following data generating structure:

$$
\begin{aligned}
X &\sim N(0,1)\,, \\
Y|X &\sim N\big(\mu(X)\,, \sigma^2(X)\big)\,,
\end{aligned}
$$

where the mean, $\mu(X)$, and the standard deviation, $\sigma(X)$, are

$$
\begin{aligned}
\mu(X) &= 20\delta\left(1 - \exp\left(-0.2|X|\right)\right) + \delta\left(\exp\left(1\right) - \exp\left(\cos(2\pi\kappa X)\right)\right), \\
\sigma(X) &= 1 + 0.7|X| + 0.4\cos(3\pi X),
\end{aligned}
$$

respectively.

The metamodel for the conditional distribution, $Y|X$, is $N(\hat{\mu}(X), \hat{\sigma}^2(X))$, where

$$\hat{\mu}(X) = 20\beta\delta\left(1 - \exp\left(-0.2|X|\right)\right) + \rho\delta\left(\exp\left(1\right) - \exp\left(\cos(2\pi\kappa X)\right)\right),$$
$$\hat{\sigma}(X) = 1 + 0.7|X| + 0.4\rho\cos(3\pi X).$$

For the experiments of Ackley1D, we use the same baseline setup used in Cannamela1D and Ackley3D, namely, $P_T = 0.01$, $\delta = 1$, $M/N_T = 30\%$, $\rho = 1$, $\beta = 1$, and $\kappa = 0.5$. Note that $\rho = 1$ and $\beta = 1$ imply that the metamodel is perfect so that the optimal IS densities and allocations can be used.

Table A.14 below compares the results of Ackley1D and Ackley3D. We note that the relative ratios of SIS1 and SIS2 for $P_T = 0.01$ in Ackley1D, namely, 15% and 17%, are smaller than those in Ackley3D, namely, 29% and 29%. Yet, the performances in Ackley1D are not as remarkable as those in Cannamela1D in Table A.1, namely, 2.5% and 3.6%. Such performance differences in Cannamela1D and Ackley1D can be explained mainly by the difference in their underlying data generating structures: See Figure A.4 below, where we plot the optimal SIS1 density along with the original input density for both examples. Apparently, the optimal SIS1 density for Cannamela1D is deviating much more from the original input density than that for Ackley1D is. We observe the similar pattern for SIS2. This explains the better performances of SIS1 and SIS2 for Cannamela1D.

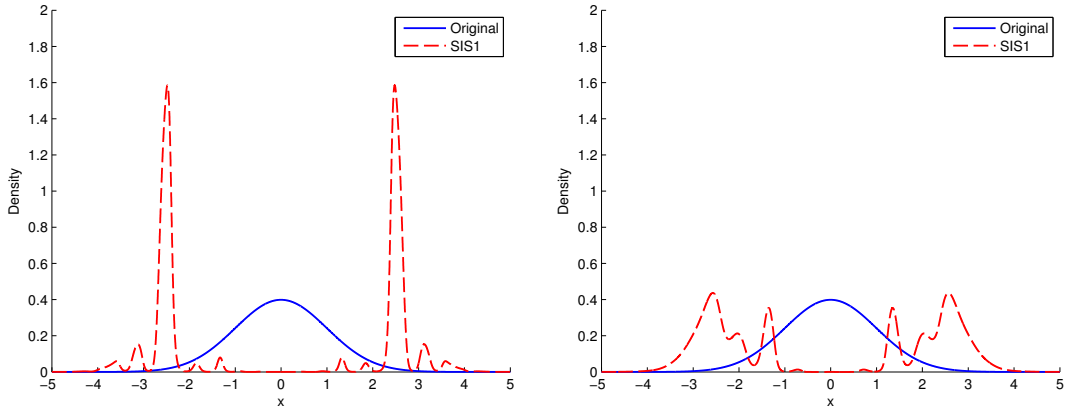
Obviously, the computational gains of SIS1 and SIS2 over CMC largely depend on the general trend represented by the location parameter function, $\mu(X)$. In addition, the scale parameter function, $\sigma(X)$, also makes a difference in the performances of SIS1 and SIS2 for Cannamela1D and Ackley1D. We plot 20,000 input-output pairs, $(X, Y)$'s, generated from the baseline setups for Cannamela1D and Ackley1D in Figures A.5(a) and (b), respectively. We draw the solid horizontal line in each plot to indicate the resistance level, $l$, corresponding to $P_T = 0.01$. We observe that the lo-

cation parameter functions, $\mu(X)$, in both examples tend to have large values at the regions where $f(X)$ is small. However, the scale parameter functions, $\sigma(X)$, lead to a major difference around the region, $(-2,-1)\cup(1,2)$, where $\mu(X)$ itself is not yet close to $l$ but many responses of Ackley1D in Figure A.5(b) exceed $l$ unlike Cannamela1D in Figure A.5(a). Accordingly, we observe the relevant peaks at $(-2,-1)\cup(1,2)$ in Figure A.4(b), which disperse sampling efforts in a larger input area and make $q_{SIS1}$ (and $q_{SIS2}$) more overlapped with $f$ for Ackley1D.

Table A.14: POE estimation results with different input dimension and target failure probability, $P_T$, for the numerical examples based on *Ackley* (1987)

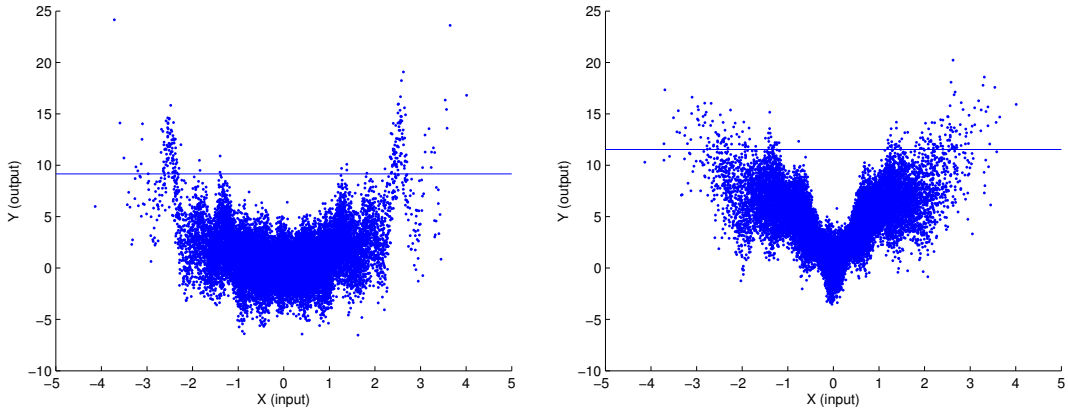
| Method | | Ackley1D | | | Ackley3D | | |
|---|---|---|---|---|---|---|---|
| | | $P_T$ | | | $P_T$ | | |
| | | 0.10 | 0.05 | 0.01 | 0.10 | 0.05 | 0.01 |
| SIS1 | Sample Average | 0.1001 | 0.0501 | 0.0100 | 0.1002 | 0.0501 | 0.0100 |
| | Standard Error | 0.0059 | 0.0038 | 0.0012 | 0.0070 | 0.0046 | 0.0017 |
| | Relative Ratio | 39% | 30% | 15% | 54% | 45% | 29% |
| SIS2 | Sample Average | 0.0998 | 0.0501 | 0.0100 | 0.1002 | 0.0499 | 0.0100 |
| | Standard Error | 0.0060 | 0.0040 | 0.0013 | 0.0070 | 0.0048 | 0.0017 |
| | Relative Ratio | 40% | 34% | 17% | 54% | 49% | 29% |
| BIS | Sample Average | 0.1000 | 0.0499 | 0.0100 | 0.1000 | 0.0500 | 0.0100 |
| | Standard Error | 0.0072 | 0.0052 | 0.0027 | 0.0082 | 0.0062 | 0.0026 |
| | Relative Ratio | 58% | 57% | 74% | 75% | 81% | 68% |
| CMC | Sample Average | 0.1001 | 0.0501 | 0.0100 | 0.0997 | 0.0500 | 0.0101 |
| | Standard Error | 0.0098 | 0.0071 | 0.0031 | 0.0094 | 0.0069 | 0.0031 |

In summary, the performances of the proposed methods will depend on the characteristics of the simulation model. Note that the variances of the proposed estimators depend only on the functions, $s(\mathbf{x})$ and $f(\mathbf{x})$, according to Theorems II.2 and II.3 (note that the SIS1 and SIS2 densities are also expressed in $s(\mathbf{x})$ and $f(\mathbf{x})$) and both functions are determined by the true data generating structure. Lastly, we remark that the higher relative ratios of SIS1 and SIS2 for Ackley3D compared to those for Ackley1D should not be generalized as that the input dimension negatively affects the performances of SIS1 and SIS2. In the case of Ackley3D, due to the highly oscillating response over the *three* dimensional input space, the sampling efforts are




(a) Cannamela1D (b) Ackley1D

Figure A.4: Comparison of the optimal SIS1 density and the original input density for the two examples




(a) Cannamela1D (b) Ackley1D

Figure A.5: Scatter plots of the data generated from the baseline data generating structures: the solid horizontal line is the quantile, $l$, corresponding to $P_T = 0.01$

more distributed in the larger input space and the resulting $q_{SIS1}$ (and $q_{SIS2}$) is more overlapped with $f$, compared to the case of Ackley1D. However, even with high dimensional input vectors, significant computational reduction can be achieved when the joint density of the input vector, $f$, and the optimal SIS1 (and SIS2) density, $q_{SIS1}$ (and $q_{SIS2}$), are different.

### A.3.6 Summary

Overall, we observe similar patterns both in the univariate example and the multivariate example. These patterns are also consistent with the wind turbine simulation results. For a wide range of parameter settings, the performances of SIS1 and SIS2 are superior to BIS and CMC.

## A.4 Implementation Details with Wind Turbine Simulators

In this section, we present the implementation details with wind turbine simulators.

### A.4.1 NREL simulators and the original input distribution

The NREL simulators used in this study include TurbSim (*Jonkman*, 2009) and FAST (*Jonkman and Buhl Jr.*, 2005). Given a wind condition (e.g., 10-minute average wind speed), TurbSim produces a three-dimensional stochastic wind profile. FAST, taking the generated wind profile as an input, simulates load responses (or loads) at turbine subsystems such as blades and shafts. Noting that there are many types of load responses, we limit our study to consider edgewise and flapwise bending moments at a blade root as output variables, where edgewise (flapwise) bending moments imply structural loads parallel (perpendicular) to the rotor span at a blade root. These two load types are of great concern in ensuring a wind turbine's structural reliability (*Moriarty*, 2008).

As in *Moriarty* (2008), we use the same turbine specification for an onshore version of an NREL 5-MW baseline wind turbine (*Jonkman et al.*, 2009). The target turbine operates within a specified wind speed range between the cut-in speed, $x_{in} = 3$ meter per second ($m/s$), and the cut-out speed, $x_{out} = 25\ m/s$. Following wind industry practice and the international standard, IEC 61400-1 (*International Electrotechnical*

*Commission*, 2005), we use a 10-minute average wind speed as an input, $X$, to the simulators. We use a Rayleigh density for $X$ with a truncated support of $[x_{in}, x_{out}]$ as in *Moriarty* (2008):

$$f(x) = \frac{f_R(x)}{F_R(x_{out}) - F_R(x_{in})},$$

where $F_R(x) = 1 - e^{-x^2/2\tau^2}$ denotes the cumulative distribution function of Rayleigh distribution with a scale parameter, $\tau = \sqrt{2/\pi} \cdot 10$ (unit: m/s). Also, $f_R$ denotes the Rayleigh density function with the same scale parameter.

### A.4.2 Acceptance rates of the acceptance-rejection algorithm

We use the acceptance-rejection algorithm in the implementation. The algorithm's acceptance rate is equal to the normalizing constant of each IS density because we use the original input density, $f$, as an instrumental (or auxiliary) density for the algorithm (*Kroese et al.*, 2011). Note that the normalizing constants of the optimal IS densities are $C_{q1}$ for SIS1, $C_{q2}$ for SIS2, and $P(Y > l)$ for BIS.

The acceptance rates differ, depending on POE, $P(Y > l)$. In our implementation, when POE is around 0.05 (i.e., edgewise moments with $l$ = 8,600 $kNm$ or flapwise moments with $l$ = 13,800 $kNm$), the acceptance rates are 5–21%. When POE is around 0.01 (i.e., edgewise moments with $l$ = 9,300 $kNm$ or flapwise moments with $l$ = 14,300 $kNm$), the acceptance rates are 1–14%. In practice, the computational cost of the acceptance-rejection algorithm would be insignificant. For example, sampling thousands of inputs from the IS densities is a matter of seconds, whereas thousands of the NREL simulation replications can take days.

### A.4.3 Goodness-of-fit test for the model

In constructing the metamodel, we assume no prior information on important area, so sampling $X$ from the uniform distribution would be generally suitable. We use the GEV distribution for approximating the conditional POE given $X$ regardless of the choice of distribution for $X$, and the GEV distribution is employed over the entire input space with varying location and scale parameters, $\mu(X)$ and $\sigma(X)$. In our implementation, we use the metamodel based on the GEV distribution to approximate the theoretically optimal IS density, $q_{SIS1}$ (or $q_{SIS2}$). That is, the GEV distribution is used as a means to find the good IS density. Then, we run the real simulators (not the metamodel) to gather $Y$ for each $X$ sampled from $q_{SIS1}$ (or $q_{SIS2}$).

Obviously, the metamodel quality affects the performance of the proposed approach. Therefore, in our study, we used the GEV goodness-of-fit to check if the GEV provides a good approximation of the conditional distribution over the entire input space, as shown in Chapter II. In this section, we additionally check if the GEV is suitable in the area where $X$ is likely sampled. Noting that high edgewise (flapwise) bending moments are most likely observed when wind speeds are between 17 and 25 (11 and 19), we take 50 observations each at $17, 19, \cdots, 25$ $(11, 13, \cdots, 19)$ $m/s$ and conduct Kolmogorov-Smirnov (KS) tests to assess the goodness-of-fit of the GEV distribution at each wind speed. The results in Table A.15 below support the use of GEV distribution for edgewise and flapwise bending moments, as the $p$-values are greater than a reasonable significance level, say, 5%.

Table A.15: KS tests for GEV at imporant wind speeds

| Edgewise bending moments | | Flapwise bending moments | |
|---|---|---|---|
| $x$ $(m/s)$ | $p$-value | $x$ $(m/s)$ | $p$-value |
| 17 | 0.34 | 11 | 0.31 |
| 19 | 0.60 | 13 | 0.52 |
| 21 | 0.89 | 15 | 0.35 |
| 23 | 0.19 | 17 | 0.57 |
| 25 | 0.64 | 19 | 0.36 |

### A.4.4 CMC simulations

We want to ensure that the estimations of $N_T^{(CMC)}$ are accurate, which are used to compute the relative ratios in Tables 2.5 and 2.6. Thus, we run CMC simulations with $N_T^{(CMC)}$ corresponding to SIS1 and SIS2 for the flapwise moment with $l = 13,800$ $kNm$ and compute the standard errors based on 50 repetitions. The corresponding $N_T^{(CMC)}$ for SIS1 and SIS2 are 6,219 and 4,762, respectively. In addition, we run simulations with $N_T^{(CMC)} = 5,000$ and $N_T^{(CMC)} = 6,000$. With $N_T^{(CMC)}$ of 6,000 and 6,219, we obtain the CMC's standard error of 0.0028, which is the same with the SIS1's standard error with $N_T = 2,000$. With $N_T^{(CMC)} = 4,762$ and $N_T^{(CMC)} = 5,000$, we obtain the CMC's standard errors of 0.0036 and 0.0033, respectively, which are close to the SIS2's standard error of 0.0032. We omit the CMC implementation for other cases due to the intensive computational requirement.

# APPENDIX B

# Appendix for Chapter III

This Appendix contains the proofs of Propositions III.4–III.6, Lemma III.5, Theorems III.7–III.10, and Corollary III.11.

## B.1 Proof of Proposition III.4

We first prove that the optimal SIS1 density, $q_{1,y}$, satisfies Assumption III.2 and then that the optimal SIS2 density, $q_{2,y}$, does too. Because both optimal SIS densities satisfy Assumption III.1 (see Chapter II), we will use the property, $\mathbb{E}_q(L) = \mathbb{E}_f(1)$, in the subsequent derivation.

- Proof for $q_{1,y}$ satisfying Assumption III.2: By plugging the optimal SIS1 density,

$q_{1,y}$, in (3.2) into $\mathbb{E}_q\left[\mathbb{I}(Y > y)L^2\right]$ leads to

$$\begin{aligned}
\mathbb{E}_q\left[\mathbb{I}(Y > y)L^2\right] &= \mathbb{E}_q\left[\mathbb{E}\left[\mathbb{I}(Y > y) \mid \mathbf{X}\right] L^2\right] \\
&= \mathbb{E}_q\left[s_y(\mathbf{X})\, L^2\right] \\
&= \mathbb{E}_f\left[s_y(\mathbf{X})\, \frac{f(\mathbf{X})}{q_{1,y}(\mathbf{X})}\right] \\
&= \mathbb{E}_f\left[s_y(\mathbf{X})\, \frac{f(\mathbf{X})}{\frac{1}{C_{q1}} f(\mathbf{X}) \sqrt{\frac{1}{n} s_y(\mathbf{X})\left(1 - s_y(\mathbf{X})\right) + s_y(\mathbf{X})^2}}\right] \\
&= C_{q1}\mathbb{E}_f\left[\frac{s_y(\mathbf{X})}{\sqrt{\frac{1}{n} s_y(\mathbf{X})\left(1 - s_y(\mathbf{X})\right) + s_y(\mathbf{X})^2}}\right] \\
&\leq C_{q1}\mathbb{E}_f\left[\frac{s_y(\mathbf{X})}{\sqrt{s_y(\mathbf{X})^2}}\right] \qquad \text{(B.1)} \\
&= C_{q1}.
\end{aligned}$$

The inequality in (B.1) holds because $\frac{1}{n} s_y(\mathbf{X})\left(1 - s_y(\mathbf{X})\right) \geq 0$. Here,

$$\begin{aligned}
C_{q1} &= \int_{\mathcal{X}_f} f(\mathbf{x}) \sqrt{\frac{1}{n} s_y(\mathbf{x}) \cdot \left(1 - s_y(\mathbf{x})\right) + s_y(\mathbf{x})^2}\, \mathrm{d}\mathbf{x} \\
&\leq \int_{\mathcal{X}_f} f(\mathbf{x}) \sqrt{(1+1)}\, \mathrm{d}\mathbf{x} \qquad \text{(B.2)} \\
&= \sqrt{2} \\
&< \infty,
\end{aligned}$$

where $\mathcal{X}_f$ is the support of $f$. The inequality in (B.2) holds because the both summands within the square root are bounded above by 1. Therefore, $\mathbb{E}_q\left[\mathbb{I}(Y > y)L^2\right] < \infty$ holds for the optimal SIS1 density.

- Proof for $q_{2,y}$ satisfying Assumption III.2: Now consider the optimal SIS2 den-

sity, $q_{2,y}$, in (3.6). With this density,

$$\begin{aligned}
\mathbb{E}_q\left[s_y(\mathbf{X})\,L^2\right] &= \mathbb{E}_f\left[s_y(\mathbf{X})\,\frac{f(\mathbf{X})}{\frac{1}{C_{q2}}\sqrt{s_y(\mathbf{X})}f(\mathbf{X})}\right] \\
&= C_{q2}\mathbb{E}_f\left[\sqrt{s_y(\mathbf{X})}\right] \\
&= C_{q2}^2,
\end{aligned}$$

where the last equality holds because $C_{q2}$ is $\mathbb{E}_f\left[\sqrt{s_y(\mathbf{X})}\right]$ by definition. Here,

$$\begin{aligned}
C_{q2} &= \int_{\mathcal{X}_f} \sqrt{s_y(\mathbf{x})}f(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&\leq \int_{\mathcal{X}_f} f(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= 1 \\
&< \infty.
\end{aligned}$$

Therefore, $\mathbb{E}_q\left[\mathbb{I}(Y > y)L^2\right] < \infty$ holds for the optimal SIS2 density. $\square$

## B.2 Proof of Lemma III.5

To prove $N_i \xrightarrow{P} \tilde{N}_i$ in (3.10), we first define

$$\eta_i \equiv n\frac{h(\mathbf{X}_i)}{\sum_{j=1}^m h(\mathbf{X}_j)} + \frac{1}{2}, \quad i = 1, \ldots, m, \tag{B.3}$$

$$\tilde{\eta}_i \equiv \frac{h(\mathbf{X}_i)}{c_0\mathbb{E}_q[h(\mathbf{X})]} + \frac{1}{2}, \quad i = 1, \ldots, m,$$

and

$$r(x) \equiv \max\left(1, \lfloor x \rfloor\right), \tag{B.4}$$

so that $N_i$ in (3.7) and $\tilde{N}_i$ in (3.11) can be expressed as

$$\begin{aligned} N_i &= \max\left(1, \left\lfloor n\frac{h(\mathbf{X}_i)}{\sum_{j=1}^m h(\mathbf{X}_j)} + \frac{1}{2} \right\rfloor\right), \quad i = 1, \ldots, m \\ &= r(\eta_i), \quad i = 1, \ldots, m, \end{aligned}$$

and

$$\begin{aligned} \tilde{N}_i &= \max\left(1, \left\lfloor \frac{h(\mathbf{X}_i)}{c_0 \mathbb{E}_q[h(\mathbf{X})]} + \frac{1}{2} \right\rfloor\right), \quad i = 1, \ldots, m \\ &= r(\tilde{\eta}_i), \quad i = 1, \ldots, m, \end{aligned}$$

respectively.

Next, we prove $\eta_i \xrightarrow{P} \tilde{\eta}_i$ and then $r(\eta_i) \xrightarrow{P} r(\tilde{\eta}_i)$, which in turn implies that $N_i, i = 1, \ldots, m$, is asymptotically independent of one another.

- Proof of $\eta_i \xrightarrow{P} \tilde{\eta}_i$: Note that $\eta_i$ in (B.3) can be expressed as

$$\begin{aligned} \eta_i &= n\frac{h(\mathbf{X}_i)}{\sum_{j=1}^m h(\mathbf{X}_j)} + \frac{1}{2}, \quad i = 1, \ldots, m \\ &= \frac{1}{c_0} \frac{h(\mathbf{X}_i)}{\frac{1}{m}\sum_{j=1}^m h(\mathbf{X}_j)} + \frac{1}{2}, \quad i = 1, \ldots, m, \end{aligned}$$

  where in the denominator of the first term, we note

$$\frac{1}{m}\sum_{j=1}^m h(\mathbf{X}_j) \xrightarrow{P} \mathbb{E}_q[h(\mathbf{X})]$$

  as $m \to \infty$ by the weak law of large numbers (*Jiang*, 2010, Theorem 6.1)

because $h(\mathbf{X}_j), j = 1, \ldots, m$, are i.i.d. random variables with a finite mean of $\mathbb{E}_q[h(\mathbf{X})]$ by the condition in (3.8). Thus, by the continuous mapping theorem (*Van der Vaart*, 1998, Theorem 2.3), it follows that

$$\eta_i \xrightarrow{P} \tilde{\eta}_i. \tag{B.5}$$

- Proof of $r(\eta_i) \xrightarrow{P} r(\tilde{\eta}_i)$: By definition, we prove the following convergence for any $\epsilon > 0$,

$$\mathbb{P}\left(|r(\eta_i) - r(\tilde{\eta}_i)| > \epsilon\right) \to 0 \tag{B.6}$$

as $m \to \infty$.

For any fixed $\delta > 0$, the left-hand side of (B.6) can be expressed as

$$\begin{aligned}\mathbb{P}\left(|r(\eta_i) - r(\tilde{\eta}_i)| > \epsilon\right) &= \mathbb{P}\left(|r(\eta_i) - r(\tilde{\eta}_i)| > \epsilon, |\eta_i - \tilde{\eta}_i| > \delta\right) \\ &\quad + \mathbb{P}\left(|r(\eta_i) - r(\tilde{\eta}_i)| > \epsilon, |\eta_i - \tilde{\eta}_i| \leq \delta\right) \\ &\equiv \alpha_1 + \alpha_2,\end{aligned}$$

where

$$\alpha_1 = \mathbb{P}\left(|r(\eta_i) - r(\tilde{\eta}_i)| > \epsilon, |\eta_i - \tilde{\eta}_i| > \delta\right) \leq \mathbb{P}\left(|\eta_i - \tilde{\eta}_i| > \delta\right) \to 0 \tag{B.7}$$

as $m \to \infty$, because of (B.5). On the other hand, to prove $\alpha_2 \to 0$, we define the set

$$G_\delta \equiv \{x \in \mathbb{R} - \mathcal{N} \mid \exists y : |r(y) - r(x)| > \epsilon, |y - x| \leq \delta\}$$

for each $\delta > 0$. Because $r(x)$ in (B.4) is continuous at $x \in \mathbb{R} - \mathcal{N}$, it follows

that

$$\lim_{\delta \to 0} G_\delta = \emptyset,$$

which implies that $\mathbb{P}(\tilde{\eta}_i \in G_\delta) \to 0$ as $\delta \to 0$. Thus,

$$\mathbb{P}(|r(\eta_i) - r(\tilde{\eta}_i)| > \epsilon, |\eta_i - \tilde{\eta}_i| \leq \delta, \tilde{\eta}_i \notin \mathcal{N}) \leq \mathbb{P}(\tilde{\eta}_i \in G_\delta) \tag{B.8}$$

$$\to 0 \tag{B.9}$$

as $\delta \to 0$. Therefore,

$$\begin{aligned}
\alpha_2 &= \mathbb{P}\left(|r(\eta_i) - r(\tilde{\eta}_i)| > \epsilon, |\eta_i - \tilde{\eta}_i| \leq \delta\right) \\
&= \mathbb{P}(|r(\eta_i) - r(\tilde{\eta}_i)| > \epsilon, |\eta_i - \tilde{\eta}_i| \leq \delta, \tilde{\eta}_i \notin \mathcal{N}) \\
&\quad + \mathbb{P}(|r(\eta_i) - r(\tilde{\eta}_i)| > \epsilon, |\eta_i - \tilde{\eta}_i| \leq \delta, \tilde{\eta}_i \in \mathcal{N}) \\
&\leq \mathbb{P}(\tilde{\eta}_i \in G_\delta) + \mathbb{P}(|r(\eta_i) - r(\tilde{\eta}_i)| > \epsilon, |\eta_i - \tilde{\eta}_i| \leq \delta, \tilde{\eta}_i \in \mathcal{N}) && \text{(B.10)} \\
&\leq \mathbb{P}(\tilde{\eta}_i \in G_\delta) + \mathbb{P}(\tilde{\eta}_i \in \mathcal{N}) \\
&= \mathbb{P}(\tilde{\eta}_i \in G_\delta) && \text{(B.11)} \\
&\to 0, && \text{(B.12)}
\end{aligned}$$

as $\delta \to 0$. The inequality in (B.10) is due to (B.8). The equation in (B.11) is due to the condition in (3.9). The convergence in (B.12) is due to (B.9).

In summary, (B.7) and (B.12) together imply (B.6), completing the proof of $r(\eta_i) \xrightarrow{P} r(\tilde{\eta}_i)$ in (3.10). Because $\tilde{N}_i$ in (3.11) depends only on $\mathbf{X}_i$ (not $\mathbf{X}_j, j \neq i$), it follows that $\tilde{N}_i, i = 1, \ldots, m$, is independent of one another. $\square$

## B.3 Proof of Proposition III.6

Under the given conditions, we want to show $\mathbb{E}_q[h(\mathbf{X})] < \infty$. We bound $\mathbb{E}_q[h(\mathbf{X})]$ from above by a constant:

$$\begin{aligned}
\mathbb{E}_q[h(\mathbf{X})] &= \mathbb{E}_f\left[h(\mathbf{X})\frac{1}{C'_{q1}}\sqrt{\frac{1}{n}s'_y(\mathbf{X})\left(1-s'_y(\mathbf{X})\right)+s'_y(\mathbf{X})^2}\right] \\
&\leq \frac{1}{C'_{q1}}\mathbb{E}_f\left[h(\mathbf{X})\sqrt{s'_y(\mathbf{X})}\right] \qquad \text{(B.13)} \\
&= \frac{1}{C'_{q1}}\mathbb{E}_f\left[\sqrt{\frac{n\left(1-s'_y(\mathbf{X})\right)}{1+(n-1)\,s'_y(\mathbf{X})}}\sqrt{s'_y(\mathbf{X})}\right] \\
&= \frac{1}{C'_{q1}}\mathbb{E}_f\left[\sqrt{\frac{1-s'_y(\mathbf{X})}{1/n+(1-1/n)\,s'_y(\mathbf{X})}}\sqrt{s'_y(\mathbf{X})}\right] \\
&\leq \frac{1}{C'_{q1}}\mathbb{E}_f\left[\sqrt{2}\sqrt{\frac{1-s'_y(\mathbf{X})}{s'_y(\mathbf{X})}}\sqrt{s'_y(\mathbf{X})}\right] \qquad \text{(B.14)} \\
&= \frac{\sqrt{2}}{C'_{q1}}\mathbb{E}_f\left[\sqrt{1-s'_y(\mathbf{X})}\right] \\
&\leq \frac{\sqrt{2}}{C'_{q1}} \qquad \text{(B.15)} \\
&< \infty
\end{aligned}$$

where $C'_{q1}$ is the normalizing constant of $q$ when $s_y(\mathbf{x})$ in $q_{1,y}(\mathbf{x})$ in (3.2) is replaced by $s'_y(\mathbf{x})$. Because $q$ is a density function, $C'_{q1}$ is a positive constant. The inequalities in (B.13) and (B.14) hold because $n \geq 1$. The inequality in (B.15) holds because $s'_y(\mathbf{x}) \geq 0$. $\square$

## B.4 Proof of Theorem III.7

To prove the CLT in (3.12),

$$\sqrt{\frac{m}{\sigma^2_{1,y}}}\left(\hat{P}_{1,n}(y)-p_y\right) \xrightarrow{d} N(0,1),$$

we introduce the following estimator:

$$\tilde{P}_{1,n}(y) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{\tilde{N}_i}\sum_{j=1}^{\tilde{N}_i}\mathbb{I}\left(Y_j^{(i)} > y\right)\right)\frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}. \tag{B.16}$$

Then, we express the left-hand side of (3.12) as

$$\begin{aligned}\sqrt{\frac{m}{\sigma_{1,y}^2}}\left(\hat{P}_{1,n}(y) - p_y\right) &= \sqrt{\frac{m}{\sigma_{1,y}^2}}\left(\hat{P}_{1,n}(y) - \tilde{P}_{1,n}(y) + \tilde{P}_{1,n}(y) - p_y\right) \\ &= \sqrt{\frac{m}{\sigma_{1,y}^2}}\left(\hat{P}_{1,n}(y) - \tilde{P}_{1,n}(y)\right) + \sqrt{\frac{m}{\sigma_{1,y}^2}}\left(\tilde{P}_{1,n}(y) - p_y\right) \end{aligned} \tag{B.17}$$

**Our proof for** (3.12) **consists of three main steps**:

1. Proof for the first term in (B.17) converging to zero in probability:

$$\sqrt{\frac{m}{\sigma_{1,y}^2}}\left(\hat{P}_{1,n}(y) - \tilde{P}_{1,n}(y)\right) \xrightarrow{P} 0. \tag{B.18}$$

2. Proof for the second term in (B.17) converging to $N(0,1)$ in distribution:

$$\sqrt{\frac{m}{\sigma_{1,y}^2}}\left(\tilde{P}_{1,n}(y) - p_y\right) \xrightarrow{d} N(0,1). \tag{B.19}$$

3. Application of the Slutsky's theorem (*Jiang*, 2010, Theorem 2.13) to (B.17).

**To prove the first main step's result** in (B.18), we show

$$\mathbb{P}\left(\left|\sqrt{m}\left(\hat{P}_{1,n}(y) - \tilde{P}_{1,n}(y)\right)\right| > \epsilon\right) \to 0 \tag{B.20}$$

for any $\epsilon > 0$ as $m \to \infty$. Both estimators, $\hat{P}_{1,n}(y)$ and $\tilde{P}_{1,n}(y)$, are unbiased estima-

tors of $p_y$ by Assumption III.1, making

$$\begin{aligned}\mathbb{E}_q\left[\hat{P}_{1,n}(y) - \tilde{P}_{1,n}(y)\right] &= \mathbb{E}_q\left[\hat{P}_{1,n}(y)\right] - \mathbb{E}_q\left[\tilde{P}_{1,n}(y)\right] \\ &= p_y - p_y \\ &= 0.\end{aligned}$$

By Chebyshev's inequality (*Jiang*, 2010, Equation (5.77)), the left-hand side of (B.20) is bounded from above as follows:

$$\mathbb{P}\left(\left|\sqrt{m}\left(\hat{P}_{1,n}(y) - \tilde{P}_{1,n}(y)\right)\right| > \epsilon\right) \leq \frac{m}{\epsilon^2} Var_q\left[\hat{P}_{1,n}(y) - \tilde{P}_{1,n}(y)\right]. \tag{B.21}$$

Now we show that the right-hand side of (B.21) converges to zero as $m \to \infty$. We obtain

$$\begin{aligned}&\frac{m}{\epsilon^2} Var_q\left[\hat{P}_{1,n}(y) - \tilde{P}_{1,n}(y)\right] \\ &= \frac{m}{\epsilon^2}\left(\mathbb{E}_q\left[Var\left[\hat{P}_{1,n}(y) - \tilde{P}_{1,n}(y) \mid \mathbf{X}_1, \ldots, \mathbf{X}_m\right]\right]\right. \\ &\quad \left. + Var_q\left[\mathbb{E}\left[\hat{P}_{1,n}(y) - \tilde{P}_{1,n}(y) \mid \mathbf{X}_1, \ldots, \mathbf{X}_m\right]\right]\right)\end{aligned} \tag{B.22}$$

by variance decomposition (*Kroese et al.*, 2011). The second term of (B.22) vanishes because

$$\begin{aligned}&\mathbb{E}\left[\hat{P}_{1,n}(y) - \tilde{P}_{1,n}(y) \mid \mathbf{X}_1, \ldots, \mathbf{X}_m\right] \\ &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{I}\left(Y_j^{(i)} > y\right) - \frac{1}{\tilde{N}_i}\sum_{k=1}^{\tilde{N}_i}\mathbb{I}\left(Y_k^{(i)} > y\right)\right) L_i \mid \mathbf{X}_1, \ldots, \mathbf{X}_m\right] \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(s_y(\mathbf{X}_i) - s_y(\mathbf{X}_i)\right) L_i \\ &= 0.\end{aligned}$$

In the first term of (B.22), we obtain

$$
\begin{aligned}
&Var\left[\hat{P}_{1,n}(y) - \tilde{P}_{1,n}(y) \mid \mathbf{X}_1, \ldots, \mathbf{X}_m\right] \\
&= Var\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{I}\left(Y_j^{(i)} > y\right) - \frac{1}{\tilde{N}_i}\sum_{k=1}^{\tilde{N}_i}\mathbb{I}\left(Y_k^{(i)} > y\right)\right) L_i \mid \mathbf{X}_1, \ldots, \mathbf{X}_m\right] \\
&= \frac{1}{m^2}\sum_{i=1}^{m} Var\left[\left(\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{I}\left(Y_j^{(i)} > y\right) - \frac{1}{\tilde{N}_i}\sum_{k=1}^{\tilde{N}_i}\mathbb{I}\left(Y_k^{(i)} > y\right)\right) \mid \mathbf{X}_1, \ldots, \mathbf{X}_m\right] L_i^2 \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\mathbb{E}\left[\left(\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{I}\left(Y_j^{(i)} > y\right) - \frac{1}{\tilde{N}_i}\sum_{k=1}^{\tilde{N}_i}\mathbb{I}\left(Y_k^{(i)} > y\right)\right)^2 \mid \mathbf{X}_1, \ldots, \mathbf{X}_m\right] L_i^2.
\end{aligned}
$$

Here, the conditional expectation in the last equation can be simplified as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{I}\left(Y_j^{(i)} > y\right) - \frac{1}{\tilde{N}_i}\sum_{k=1}^{\tilde{N}_i}\mathbb{I}\left(Y_k^{(i)} > y\right)\right)^2 \mid \mathbf{X}_1, \ldots, \mathbf{X}_m\right] \\
&= \mathbb{E}\left[\frac{1}{N_i^2}\sum_{j=1}^{N_i}\mathbb{I}\left(Y_j^{(i)} > y\right) + \frac{2}{N_i^2}\sum_{k=1}^{N_i}\sum_{l>k}^{N_i}\mathbb{I}\left(Y_k^{(i)} > y\right)\mathbb{I}\left(Y_l^{(i)} > y\right)\right. \\
&\quad + \frac{1}{\tilde{N}_i^2}\sum_{j=1}^{\tilde{N}_i}\mathbb{I}\left(Y_j^{(i)} > y\right) + \frac{2}{\tilde{N}_i^2}\sum_{k=1}^{\tilde{N}_i}\sum_{l>k}^{\tilde{N}_i}\mathbb{I}\left(Y_k^{(i)} > y\right)\mathbb{I}\left(Y_l^{(i)} > y\right) \\
&\quad \left. - \frac{2}{N_i\tilde{N}_i}\sum_{k=1}^{N_i}\sum_{l=1}^{\tilde{N}_i}\mathbb{I}\left(Y_k^{(i)} > y\right)\mathbb{I}\left(Y_l^{(i)} > y\right) \mid \mathbf{X}_1, \ldots, \mathbf{X}_m\right] \\
&= \frac{1}{N_i}\left(s_y(\mathbf{X}_i) + (N_i - 1)\, s_y^2(\mathbf{X}_i)\right) + \frac{1}{\tilde{N}_i}\left(s_y(\mathbf{X}_i) + \left(\tilde{N}_i - 1\right) s_y^2(\mathbf{X}_i)\right) \\
&\quad - \frac{2}{N_i\tilde{N}_i}\left(\min(N_i, \tilde{N}_i) s_y(\mathbf{X}_i) + \left(N_i\tilde{N}_i - \min(N_i, \tilde{N}_i)\right) s_y^2(\mathbf{X}_i)\right) \\
&= s_y(\mathbf{X}_i)\left(1 - s_y(\mathbf{X}_i)\right)\frac{N_i + \tilde{N}_i - 2\min(N_i, \tilde{N}_i)}{N_i\tilde{N}_i} \\
&= s_y(\mathbf{X}_i)\left(1 - s_y(\mathbf{X}_i)\right)\frac{\left|N_i - \tilde{N}_i\right|}{N_i\tilde{N}_i} \\
&= s_y(\mathbf{X}_i)\left(1 - s_y(\mathbf{X}_i)\right)\left|\frac{1}{N_i} - \frac{1}{\tilde{N}_i}\right| \qquad \text{(B.23)}
\end{aligned}
$$

Therefore, the equation in (B.22) is simplified as

$$\begin{aligned}
&\frac{m}{\epsilon^2} Var_q\left[\hat{P}_{1,n}(y) - \tilde{P}_{1,n}(y)\right] \\
&= \frac{m}{\epsilon^2}\mathbb{E}_q\left[\frac{1}{m^2}\sum_{i=1}^{m} s_y(\mathbf{X}_i)\left(1 - s_y(\mathbf{X}_i)\right)\left|\frac{1}{N_i} - \frac{1}{\tilde{N}_i}\right| L_i^2\right] \\
&= \frac{1}{\epsilon^2 m}\sum_{i=1}^{m}\mathbb{E}_q\left[s_y(\mathbf{X}_i)\left(1 - s_y(\mathbf{X}_i)\right)\left|\frac{1}{N_i} - \frac{1}{\tilde{N}_i}\right| L_i^2\right] \\
&= \frac{1}{\epsilon^2}\mathbb{E}_q\left[s_y(\mathbf{X}_1)\left(1 - s_y(\mathbf{X}_1)\right)\left|\frac{1}{N_1} - \frac{1}{\tilde{N}_1}\right| L_1^2\right], \qquad \text{(B.24)}
\end{aligned}$$

where the last equation in (B.24) holds because $X_1, \ldots, X_m$ are identically distributed.

We show that the expectation in (B.24) converges to zero as $m \to \infty$. By the continuous mapping theorem (*Van der Vaart*, 1998, Theorem 2.3) and Lemma III.5, we obtain

$$s_y(\mathbf{X}_1)\left(1 - s_y(\mathbf{X}_1)\right)\left|\frac{1}{N_1} - \frac{1}{\tilde{N}_1}\right| L_1^2 \xrightarrow{P} 0$$

as $m \to \infty$. Because

$$s_y(\mathbf{X}_1)\left(1 - s_y(\mathbf{X}_1)\right)\left|\frac{1}{N_1} - \frac{1}{\tilde{N}_1}\right| L_1^2 \leq 2 s_y(\mathbf{X}) L^2 \qquad \text{(B.25)}$$

and $\mathbb{E}_q\left[s_y(\mathbf{X})L^2\right] < \infty$ by Assumption III.2, the dominated convergence theorem (*Jiang*, 2010, Theorem 2.16) yields that the expectation in (B.24) converges to zero as $m \to \infty$. Because the right-hand side of (B.21) converges to zero, we complete the proof of (B.20), which implies (B.18).

**To prove the second main step's result** in (B.19),

$$\sqrt{\frac{m}{\sigma_{1,y}^2}}\left(\tilde{P}_{1,n}(y) - p_y\right) \xrightarrow{d} N(0,1),$$

we use the Lindeberg-Lévy central limit theorem (*Jiang*, 2010, Equation (4.23)). For

the theorem to hold, we verify its conditions as follows. First, $\tilde{P}_{1,n}(y)$ in (B.16) is the sample mean of

$$\tilde{Z}_i \equiv \left( \frac{1}{\tilde{N}_i} \sum_{j=1}^{\tilde{N}_i} \mathbb{I}\left(Y_j^{(i)} > y\right) \right) L_i, \quad i = 1, \ldots, m, \tag{B.26}$$

which are i.i.d. with

$$\begin{aligned}
\mathbb{E}_q\left[\tilde{Z}_i\right] &= \mathbb{E}_q\left[ \left( \frac{1}{\tilde{N}_i} \sum_{j=1}^{\tilde{N}_i} \mathbb{I}\left(Y_j^{(i)} > y\right) \right) L_i \right] \\
&= \mathbb{E}_q\left[ \frac{1}{\tilde{N}_i} \sum_{j=1}^{\tilde{N}_i} \mathbb{E}\left[ \mathbb{I}\left(Y_j^{(i)} > y\right) \mid \mathbf{X}_i \right] L_i \right] \\
&= \mathbb{E}_q\left[ \mathbb{P}\left(Y > y \mid \mathbf{X}_i\right) L_i \right] \\
&= p_y,
\end{aligned} \tag{B.27}$$

where the last equality holds by Assumption III.1.

Next, we obtain $Var_q\left[\tilde{Z}_i\right] = \sigma_{1,y}^2 < \infty$ because

$$\begin{aligned}
Var_q\left[\tilde{Z}_i\right] &= \mathbb{E}_q\left[\tilde{Z}_i^2\right] - \left(\mathbb{E}_q\left[\tilde{Z}_i\right]\right)^2 \\
&= \mathbb{E}_q\left[ \frac{1}{\tilde{N}_i^2} \left( \sum_{j=1}^{\tilde{N}_i} \mathbb{I}\left(Y_j^{(i)} > y\right)^2 + 2 \sum_{k=1}^{\tilde{N}_i} \sum_{l>k}^{\tilde{N}_i} \mathbb{I}\left(Y_k^{(i)} > y\right) \mathbb{I}\left(Y_l^{(i)} > y\right) \right) L_i^2 \right] - p_y^2 \\
&= \mathbb{E}_q\left[ \mathbb{E}\left[ \frac{1}{\tilde{N}_i^2} \left( \sum_{j=1}^{\tilde{N}_i} \mathbb{I}\left(Y_j^{(i)} > y\right) + 2 \sum_{k=1}^{\tilde{N}_i} \sum_{l>k}^{\tilde{N}_i} \mathbb{I}\left(Y_k^{(i)} > y\right) \mathbb{I}\left(Y_l^{(i)} > y\right) \right) L_i^2 \mid \mathbf{X}_i \right] \right] - p_y^2 \\
&= \mathbb{E}_q\left[ \frac{1}{\tilde{N}} s_y(\mathbf{X}) L^2 + \frac{\tilde{N}-1}{\tilde{N}} s_y(\mathbf{X})^2 L^2 \right] - p_y^2 \\
&= \mathbb{E}_q\left[ \frac{1}{\tilde{N}} s_y(\mathbf{X}) \left(1 - s_y(\mathbf{X})\right) L^2 \right] + \mathbb{E}_q\left[ s_y(\mathbf{X})^2 L^2 \right] - p_y^2 \\
&= \sigma_{1,y}^2
\end{aligned} \tag{B.28}$$

and Assumption III.2 ensures that the expectation terms in $\sigma_{1,y}^2$ are finite:

$$\mathbb{E}_q\left[\frac{1}{\tilde{N}}s_y(\mathbf{X})\left(1-s_y(\mathbf{X})\right)L^2\right] \leq \mathbb{E}_q\left[s_y(\mathbf{X})\left(1-s_y(\mathbf{X})\right)L^2\right] \leq \mathbb{E}_q\left[s_y(\mathbf{X})L^2\right] < \infty, \tag{B.29}$$

$$\mathbb{E}_q\left[s_y(\mathbf{X})^2L^2\right] \leq \mathbb{E}_q\left[s_y(\mathbf{X})L^2\right] < \infty. \tag{B.30}$$

Thus, $Var_q\left[\tilde{Z}_i\right] = \sigma_{1,y}^2 < \infty$ follows, completing the proof of (B.19) by the Lindeberg-Lévy central limit theorem.

By applying the Slutsky's theorem (*Jiang*, 2010, Theorem 2.13) to (B.17) based on (B.18) and (B.19), we complete the proof of (3.12). □

## B.5 Proof of Theorem III.8

The CLT for the SIS2 estimator, $\hat{P}_{2,n}(y)$, in (3.5) follows from the Lindeberg-Lévy central limit theorem (*Jiang*, 2010, Equation (4.23)), because $\hat{P}_{2,n}(y)$ is the sample mean of $Z_i \equiv \mathbb{I}\left(Y_i > y\right)L_i, i = 1, \ldots, n$, which are i.i.d. with

$$\begin{aligned}
\mathbb{E}_q\left[Z_i\right] &= \mathbb{E}_q\left[\mathbb{E}\left[Z_i \mid \mathbf{X}_i\right]\right] \\
&= \mathbb{E}_q\left[s_y(\mathbf{X}_i)L_i\right] \\
&= \mathbb{E}_f\left[s_y(\mathbf{X})\right] \\
&= p_y,
\end{aligned} \tag{B.31}$$

where the equality in (B.31) holds by Assumption III.1. Also, we have $Var_q[Z_i] = \sigma^2_{2,y} < \infty$ because

$$\begin{aligned} Var_q[Z_i] &= \mathbb{E}_q\left[Z_i^2\right] - \left(\mathbb{E}_q[Z_i]\right)^2 \\ &= \mathbb{E}_q\left[\mathbb{I}(Y_i > y) L_i^2\right] - p_y^2 \\ &= \mathbb{E}_q\left[\mathbb{E}\left[\mathbb{I}(Y_i > y) \mid \mathbf{X}_i\right] L_i^2\right] - p_y^2 \\ &= \mathbb{E}_q\left[s_y(\mathbf{X}) L^2\right] - p_y^2 \qquad \text{(B.32)} \\ &< \infty, \end{aligned}$$

where the last inequality follows from Assumption III.2. □

## B.6 Proof of Theorem III.9

*(a)* To prove $\hat{\sigma}^2_{1,y} \xrightarrow{P} \sigma^2_{1,y}$ in (3.16), we want to show

$$\mathbb{P}\left(\left|\hat{\sigma}^2_{1,y} - \sigma^2_{1,y}\right| > \epsilon\right) \to 0 \qquad \text{(B.33)}$$

for any $\epsilon > 0$, as $m \to \infty$. We bound the left-hand side of (B.33) from above as follows:

$$\begin{aligned} \mathbb{P}\left(\left|\hat{\sigma}^2_{1,y} - \sigma^2_{1,y}\right| > \epsilon\right) &= \mathbb{P}\left(\left|\hat{\sigma}^2_{1,y} - \tilde{\sigma}^2_{1,y} + \tilde{\sigma}^2_{1,y} - \sigma^2_{1,y}\right| > \epsilon\right) \\ &\leq \mathbb{P}\left(\left|\hat{\sigma}^2_{1,y} - \tilde{\sigma}^2_{1,y}\right| + \left|\tilde{\sigma}^2_{1,y} - \sigma^2_{1,y}\right| > \epsilon\right) \\ &\leq \mathbb{P}\left(\left|\hat{\sigma}^2_{1,y} - \tilde{\sigma}^2_{1,y}\right| > \epsilon/2\right) + \mathbb{P}\left(\left|\tilde{\sigma}^2_{1,y} - \sigma^2_{1,y}\right| > \epsilon/2\right), \qquad \text{(B.34)} \end{aligned}$$

where

$$\tilde{\sigma}^2_{1,y} \equiv \frac{1}{m-1} \sum_{i=1}^{m} \left( \frac{1}{\tilde{N}_i} \sum_{j=1}^{\tilde{N}_i} \mathbb{I}\left(Y_j^{(i)} > y\right) L_i - \hat{P}_{1,n}(y) \right)^2.$$

To prove (B.33), we show that the two terms in (B.34) converge to zeros as follows.

- Proof of $\mathbb{P}\big(\big|\hat{\sigma}^2_{1,y} - \tilde{\sigma}^2_{1,y}\big| > \epsilon/2\big) \to 0$: To simplify $\big|\hat{\sigma}^2_{1,y} - \tilde{\sigma}^2_{1,y}\big|$, we first define

$$\hat{s}_y(\mathbf{X}_i) \equiv \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbb{I}\left(Y_j^{(i)} > y\right)$$

and

$$\tilde{s}_y(\mathbf{X}_i) \equiv \frac{1}{\tilde{N}_i} \sum_{j=1}^{\tilde{N}_i} \mathbb{I}\left(Y_j^{(i)} > y\right).$$

Also, we simplify $\hat{\sigma}^2_{1,y}$ by algebraic operations as follows:

$$\begin{aligned}
\hat{\sigma}^2_{1,y} &= \frac{1}{m-1} \sum_{i=1}^{m} \left(\hat{s}_y(\mathbf{X}_i) L_i - \hat{P}_{1,n}(y)\right)^2 \\
&= \frac{1}{m-1} \sum_{i=1}^{m} \left(\hat{s}_y(\mathbf{X}_i)^2 L_i^2 - 2\hat{s}_y(\mathbf{X}_i) L_i \hat{P}_{1,n}(y) + \hat{P}^2_{1,n}(y)\right) \\
&= \frac{m}{m-1} \left(\frac{1}{m} \sum_{i=1}^{m} \hat{s}_y(\mathbf{X}_i)^2 L_i^2 - 2\hat{P}^2_{1,n}(y) + \hat{P}^2_{1,n}(y)\right) \\
&= \frac{m}{m-1} \left(\frac{1}{m} \sum_{i=1}^{m} \hat{s}_y(\mathbf{X}_i)^2 L_i^2 - \hat{P}^2_{1,n}(y)\right).
\end{aligned}$$

Similarly, we can simplify $\tilde{\sigma}^2_{1,y}$ as

$$\tilde{\sigma}^2_{1,y} = \frac{m}{m-1} \left(\frac{1}{m} \sum_{i=1}^{m} \tilde{s}_y(\mathbf{X}_i)^2 L_i^2 - \hat{P}^2_{1,n}(y)\right).$$

Then, we obtain

$$
\begin{aligned}
&\mathbb{P}\left(\left|\hat{\sigma}_{1,y}^2 - \tilde{\sigma}_{1,y}^2\right| > \epsilon/2\right) \\
&= \mathbb{P}\left(\left|\left(\frac{1}{m-1}\sum_{i=1}^{m}\hat{s}_y(\mathbf{X}_i)^2 L_i^2 - \hat{P}_{1,n}^2(y)\right) - \left(\frac{1}{m-1}\sum_{i=1}^{m}\tilde{s}_y(\mathbf{X}_i)^2 L_i^2 - \hat{P}_{1,n}^2(y)\right)\right| > \epsilon/2\right) \\
&= \mathbb{P}\left(\left|\frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{s}_y(\mathbf{X}_i)^2 - \tilde{s}_y(\mathbf{X}_i)^2\right) L_i^2\right| > \epsilon/2\right) \\
&\leq \frac{2}{\epsilon}\mathbb{E}_q\left[\left|\frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{s}_y(\mathbf{X}_i)^2 - \tilde{s}_y(\mathbf{X}_i)^2\right) L_i^2\right|\right] \qquad \text{(B.35)} \\
&\leq \frac{2}{\epsilon}\mathbb{E}_q\left[\frac{1}{m-1}\sum_{i=1}^{m}\left|\hat{s}_y(\mathbf{X}_i)^2 - \tilde{s}_y(\mathbf{X}_i)^2\right| L_i^2\right] \\
&= \frac{2}{\epsilon}\frac{1}{m-1}\sum_{i=1}^{m}\mathbb{E}_q\left[\left|\hat{s}_y(\mathbf{X}_i)^2 - \tilde{s}_y(\mathbf{X}_i)^2\right| L_i^2\right] \\
&= \frac{2}{\epsilon}\frac{m}{m-1}\mathbb{E}_q\left[\left|\hat{s}_y(\mathbf{X}_1)^2 - \tilde{s}_y(\mathbf{X}_1)^2\right| L_1^2\right] \qquad \text{(B.36)} \\
&= \frac{2}{\epsilon}\frac{m}{m-1}\mathbb{E}_q\left[\left|\hat{s}_y(\mathbf{X}_1) - \tilde{s}_y(\mathbf{X}_1)\right| \left(\hat{s}_y(\mathbf{X}_1) + \tilde{s}_y(\mathbf{X}_1)\right) L_1^2\right] \\
&\leq \frac{2}{\epsilon}\frac{m}{m-1}\sqrt{\mathbb{E}_q\left[\left(\hat{s}_y(\mathbf{X}_1) - \tilde{s}_y(\mathbf{X}_1)\right)^2 L_1^2\right]}\sqrt{\mathbb{E}_q\left[\left(\hat{s}_y(\mathbf{X}_1) + \tilde{s}_y(\mathbf{X}_1)\right)^2 L_1^2\right]} \qquad \text{(B.37)} \\
&= \frac{2}{\epsilon}\frac{m}{m-1}\sqrt{\mathbb{E}_q\left[s_y(\mathbf{X}_1)\left(1 - s_y(\mathbf{X}_1)\right)\left|\frac{1}{N_1} - \frac{1}{\tilde{N}_1}\right| L_1^2\right]}\sqrt{\mathbb{E}_q\left[\left(\hat{s}_y(\mathbf{X}_1) + \tilde{s}_y(\mathbf{X}_1)\right)^2 L_1^2\right]}, \qquad \text{(B.38)} \\
&\to 0 \qquad \text{(B.39)}
\end{aligned}
$$

where the inequality in (B.35) holds by the Chebyshev's inequality (*Jiang*, 2010, Equation (5.77)). The equality in (B.36) holds because $|\hat{s}_y(\mathbf{X}_i)^2 - \tilde{s}_y(\mathbf{X}_i)^2| L_i^2$, $i = 1, \ldots, m$ are identically distributed. The inequality in (B.37) holds by the Cauchy–Schwarz inequality (*Jiang*, 2010, Equation (5.60)). The equality in (B.38) holds by (B.23). The convergence in (B.39) holds by the following three facts:

- The ratio, $\frac{m}{m-1}$, in (B.38) goes to one as $m \to \infty$.
- The first square-rooted expectation in (B.38),

$$\sqrt{\mathbb{E}_q\left[s_y(\mathbf{X}_1)\left(1 - s_y(\mathbf{X}_1)\right)\left|\frac{1}{N_1} - \frac{1}{\tilde{N}_1}\right| L_1^2\right]},$$

goes to zero as $m \to \infty$, as it was shown that (B.24) goes to zero as $m \to \infty$ based on Assumption III.2, Lemma III.5, and the dominated convergence theorem (*Jiang*, 2010, Theorem 2.16).

– The second square-rooted expectation in (B.38) is finite:

$$\begin{aligned}
&\sqrt{\mathbb{E}_q\left[\left(\hat{s}_y(\mathbf{X}_1)+\tilde{s}_y(\mathbf{X}_1)\right)^2 L_1^2\right]} \\
&= \sqrt{\mathbb{E}_q[(\hat{s}_y(\mathbf{X}_1)^2 + 2\hat{s}_y(\mathbf{X}_1)\tilde{s}_y(\mathbf{X}_1) + \tilde{s}_y(\mathbf{X}_1)^2)\, L_1^2]} \\
&\leq \sqrt{\mathbb{E}_q[(\hat{s}_y(\mathbf{X}_1) + 2\tilde{s}_y(\mathbf{X}_1) + \tilde{s}_y(\mathbf{X}_1))\, L_1^2]} \qquad \text{(B.40)} \\
&= \sqrt{\mathbb{E}_q[(s_y(\mathbf{X}_1) + 2s_y(\mathbf{X}_1) + s_y(\mathbf{X}_1))\, L_1^2]} \\
&= 2\sqrt{\mathbb{E}_q[s_y(\mathbf{X}_1)L_1^2]} \\
&< \infty, \qquad \text{(B.41)}
\end{aligned}$$

where the inequality in (B.40) holds because of $0 \leq \hat{s}_y(\mathbf{X}_1) \leq 1$ and $0 \leq \tilde{s}_y(\mathbf{X}_1) \leq 1$. The inequality in (B.41) holds by Assumption III.2.

- Proof of $\mathbb{P}\left(\left|\tilde{\sigma}_{1,y}^2 - \sigma_{1,y}^2\right| > \epsilon/2\right) \to 0$: By definition, we want to show

$$\tilde{\sigma}_{1,y}^2 \xrightarrow{P} \sigma_{1,y}^2. \qquad \text{(B.42)}$$

Because

$$\tilde{\sigma}_{1,y}^2 = \frac{m}{m-1}\left(\frac{1}{m}\sum_{i=1}^{m} \tilde{s}_y(\mathbf{X}_i)^2 L_i^2 - \hat{P}_{1,n}^2(y)\right).$$

and

$$\sigma_{1,y}^2 = \mathbb{E}_q\left[\frac{1}{\tilde{N}} s_y(\mathbf{X})\left(1 - s_y(\mathbf{X})\right) L^2\right] + \mathbb{E}_q\left[s_y(\mathbf{X})^2 L^2\right] - p_y^2,$$

the convergence in probability in (B.42) follows if

$$\frac{1}{m}\sum_{i=1}^{m}\tilde{s}_y(\mathbf{X}_i)^2L_i^2 \xrightarrow{P} \mathbb{E}_q\left[\frac{1}{\tilde{N}}s_y(\mathbf{X})\left(1-s_y(\mathbf{X})\right)L^2\right]+\mathbb{E}_q\left[s_y(\mathbf{X})^2L^2\right] \tag{B.43}$$

and

$$\hat{P}_{1,n}^2(y) \xrightarrow{P} p_y^2. \tag{B.44}$$

– Proof of the convergence in probability in (B.43): This convergence holds by the weak law of large numbers (*Jiang*, 2010, Theorem 6.1) because $\tilde{s}_y(\mathbf{X}_i)^2L_i^2, i=1,\ldots,m$ are i.i.d. and

$$\mathbb{E}_q\left[\tilde{s}_y(\mathbf{X})^2L^2\right] = \mathbb{E}_q\left[\frac{1}{\tilde{N}}s_y(\mathbf{X})\left(1-s_y(\mathbf{X})\right)L^2\right]+\mathbb{E}_q\left[s_y(\mathbf{X})^2L^2\right] \tag{B.45}$$

$$< \infty, \tag{B.46}$$

where the equation in (B.45) is derived in (B.28). The inequality in (B.46) holds by Assumption III.2 based on (B.29) and (B.30).

– Proof of the convergence in probability in (B.44): We want to show

$$\mathbb{P}\Big(\Big|\hat{P}_{1,n}^2(y)-p_y^2\Big|>\epsilon'\Big)\to 0$$

for any $\epsilon'>0$ as $m\to\infty$. Note that

$$\begin{aligned}
\mathbb{P}\Big(\Big|\hat{P}_{1,n}^2(y)-p_y^2\Big|>\epsilon'\Big) &= \mathbb{P}\Big(\Big|\Big(\hat{P}_{1,n}(y)-p_y\Big)\Big(\hat{P}_{1,n}(y)+p_y\Big)\Big|>\epsilon'\Big)\\
&\le \mathbb{P}\Big(2\Big|\hat{P}_{1,n}(y)-p_y\Big|>\epsilon'\Big)\\
&= \mathbb{P}\Big(\Big|\hat{P}_{1,n}(y)-\tilde{P}_{1,n}(y)+\tilde{P}_{1,n}(y)-p_y\Big|>\epsilon'/2\Big)\\
&\le \mathbb{P}\Big(\Big|\hat{P}_{1,n}(y)-\tilde{P}_{1,n}(y)\Big|>\epsilon'/4\Big)+\mathbb{P}\Big(\Big|\tilde{P}_{1,n}(y)-p_y\Big|>\epsilon'/4\Big),
\end{aligned}$$

where the right-hand side of the last inequality goes to zero because the

first term,

$$\mathbb{P}\Big(\Big|\hat{P}_{1,n}(y) - \tilde{P}_{1,n}(y)\Big| > \epsilon'/4\Big) \to 0$$

as $m \to \infty$ by (B.20) and the second term,

$$\mathbb{P}\Big(\Big|\tilde{P}_{1,n}(y) - p_y\Big| > \epsilon'/4\Big) \to 0$$

as $m \to \infty$ by the weak law of large numbers (*Jiang*, 2010, Theorem 6.1) because $\tilde{P}_{1,n}(y)$ is a sample mean of i.i.d. random variables with the finite mean of $p_y$ as shown in (B.27) based on Assumption III.1.

Because (B.43) and (B.44) hold, the convergence in probability in (B.42) holds.

By (B.39) and (B.42), the right-hand side of the inequality in (B.34) goes to zero, completing the proof of (B.33) and, equivalently, (3.16).

*(b)* The statement in (3.18) follows from the Slutsky's theorem (*Jiang*, 2010, Theorem 2.13) based on (3.12) and (3.16). □

## B.7 Proof of Theorem III.10

*(a)* To prove $\hat{\sigma}^2_{2,y} \xrightarrow{P} \sigma^2_{2,y}$ in (3.19), we first simplify the expression of $\hat{\sigma}^2_{2,y}$ in (3.20) as follows:

$$\hat{\sigma}^2_{2,y} = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(Y_i > y\right)L_i^2 - \hat{P}^2_{2,n}(y)\right). \tag{B.47}$$

Because $n/(n-1)$ in (B.47) converges to one as $n \to \infty$, we consider the convergences of the two terms within the outermost parentheses in (B.47). The first term is the average of i.i.d. random variables, $\mathbb{I}\left(Y_i > y\right)L_i^2, i = 1, \ldots, n$, which have the mean of

$\mathbb{E}_q\left[s_y(\mathbf{X})L^2\right] < \infty$ from (B.32) and Assumption III.2. Thus, by the weak law of large numbers (*Jiang*, 2010, Theorem 6.1), we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(Y_i > y\right)L_i^2 \xrightarrow{P} \mathbb{E}_q\left[s_y(\mathbf{X})L^2\right].$$

Similarly, $\hat{P}_{2,n}(y) \xrightarrow{P} p_y$ by the weak law of large numbers. Therefore, it follows that

$$\begin{aligned}\hat{\sigma}_{2,y}^2 &= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(Y_i > y\right)L_i^2 - \hat{P}_{2,n}^2(y)\right)\\ &\xrightarrow{P} \sigma_{2,y}^2\end{aligned}$$

as $n \to \infty$ by the continuous mapping theorem (*Van der Vaart*, 1998, Theorem 2.3), completing the proof of (3.19).

*(b)* The statement in (3.21) follows from the Slutsky's theorem (*Jiang*, 2010, Theorem 2.13) based on (3.14) and (3.19). □

## B.8 Proof of Corollary III.11

*(a)* Among the conditions in Theorem III.9, only Assumptions III.1 and III.2 involve $y$. We show that the conditions in Assumptions III.1 and III.2 hold when $y$ is replaced by $\tilde{y}$ for $\tilde{y} > y$. Then, it follows that Theorem III.9 where $y$ is replaced by $\tilde{y}$ holds.

- Assumption III.1 with $\tilde{y}$ in place of $y$: If we replace $y$ in Assumption III.1 with

$\tilde{y}$, the condition still holds because if $q(\mathbf{x}) = 0$, then

$$\begin{aligned} 0 &\leq \mathbb{P}(Y > \tilde{y} \mid \mathbf{X} = \mathbf{x})\, f(\mathbf{x}) \\ &\leq \mathbb{P}(Y > y \mid \mathbf{X} = \mathbf{x})\, f(\mathbf{x}) \\ &= 0 \end{aligned}$$

for any $\mathbf{x}$.

- Assumption III.2 with $\tilde{y}$ in place of $y$: If we substitute $\tilde{y}$ for $y$ in Assumption III.2, the condition remains satisfied because

$$\mathbb{E}_q\left[\mathbb{I}(Y > \tilde{y})L^2\right] \leq \mathbb{E}_q\left[\mathbb{I}(Y > y)L^2\right] < \infty$$

for $\tilde{y} > y$.

Therefore, it follows that Theorem III.9 with $\tilde{y}$ in place of $y$ holds for $\tilde{y} > y$. That is, $\mathbb{P}\Big(p_{\tilde{y}} \in \Big(\hat{P}_{1,n}(\tilde{y}) \pm z_{\alpha/2}\hat{\sigma}_{1,\tilde{y}}/\sqrt{m}\Big)\Big) \to 1-\alpha$ holds for $\alpha \in (0,1)$ as $m \to \infty$.
*(b)* Similarly, because the conditions in Assumptions III.1 and III.2 hold when $y$ is substituted by $\tilde{y}$ for $\tilde{y} > y$, it follows that Theorem III.10 with $\tilde{y}$ in place of $y$ holds. Thus, $\mathbb{P}\Big(p_{\tilde{y}} \in \Big(\hat{P}_{2,n}(\tilde{y}) \pm z_{\alpha/2}\hat{\sigma}_{2,\tilde{y}}/\sqrt{n}\Big)\Big) \to 1-\alpha$ holds for $\alpha \in (0,1)$ as $n \to \infty$. $\square$

# APPENDIX C

# Appendix for Chapter IV

This Appendix provides the proof of Theorem IV.3 and the detail of approximating $N_i$ for SIS.

## C.1 Proof of Theorem IV.3

Theorem IV.3 states the asymptotic bias of $\bar{\mathcal{C}}_{\hat{\boldsymbol{\theta}}}$ in estimating $\mathcal{C}_{\hat{\boldsymbol{\theta}}}$, the CE of $q\left(\mathbf{x};\hat{\boldsymbol{\theta}}\right)$ from $q^*$, $\mathcal{C}\left(q^*(\mathbf{x}), q\left(\mathbf{x};\hat{\boldsymbol{\theta}}\right)\right)$. Based on this key result, CIC guides us to correct the asymptotic bias.

We note that AIC (*Akaike*, 1974) similarly corrects the asymptotic bias in estimating the log-likelihood under the key condition that MLE should be consistent and asymptotically normal (*Cavanaugh and Neath*, 2014). Analogously, to prove Theorem IV.3, we show that MCE is consistent and asymptotically normal in Lemmas C.1 and C.2 below, respectively. Because MLE is a special case of MCE, we can derive parallel results for MCE with those established for MLE under similar regularity conditions.

We assume the regularity conditions (*Keener*, 2010; *Cavanaugh and Neath*, 2014), such as identifiability, continuity, and differentiability, that are necessary to prove the

consistency and asymptotic normality of MLE. Also, we assume $\mathbb{E}\left[h^2(\mathbf{X})w^2(\mathbf{X})\right] < \infty$ to ensure that the IS estimator has a finite variance. Additionally, Assumption IV.1 in Chapter IV is necessary to simplify the model complexity penalty. Note that the similar assumption is also made to derive AIC (*Akaike*, 1974), namely, the true data generating density belongs to the parametric family of the density whose parameters are being estimated by MLE.

Recall that MCE is defined as the minimizer of (4.7) as follows:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \bar{\mathcal{C}}_{\boldsymbol{\theta}} \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, -\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i)\log q(\mathbf{X}_i;\boldsymbol{\theta}).\end{aligned}$$

This MCE converges in probability to the optimal parameter, $\boldsymbol{\theta}^*$, as stated in the following lemma.

**Lemma C.1.** *MCE, $\hat{\boldsymbol{\theta}}$, is a consistent estimator of the parameter, $\boldsymbol{\theta}^*$, of the density, $q\left(\mathbf{X};\boldsymbol{\theta}^*\right)$ that minimizes the cross-entropy, $\mathbb{C}(q^*, q)$.*

***Proof.*** *When $\boldsymbol{\Theta}$ is compact, the consistency of MCE is proved in Theorem A1 in Rubinstein and Shapiro (1993). Extending the result to $\boldsymbol{\Theta} = \mathbb{R}^d$ follows from Theorem 9.11 in Keener (2010).*

The intuition behind this lemma is as follows: (a) $\bar{\mathcal{C}}_{\boldsymbol{\theta}}(q^*, q)$ is a consistent estimator of $\mathbb{C}(q^*, q)$ by the weak law of large numbers (*Keener*, 2010) and (b) the MCE minimizes $\bar{\mathcal{C}}_{\boldsymbol{\theta}}(q^*, q)$ by definition. Therefore, the minimizer, $\hat{\boldsymbol{\theta}}$, of $\bar{\mathcal{C}}_{\boldsymbol{\theta}}(q^*, q)$ also converges in probability to the minimizer, $\boldsymbol{\theta}^*$, of $\mathbb{C}(q^*, q)$.

To establish the asymptotic normality of MCE in the following lemma, we define a few notations. Let $\nabla_{\boldsymbol{\theta}}$ and $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}$ denote the gradient with respect to $\boldsymbol{\theta}$ and the Hessian matrix of second order derivatives, respectively. Define $\boldsymbol{\theta}'$ as the MCE at the last CE iteration such that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are sampled from $q(\mathbf{x};\boldsymbol{\theta}')$. The expectation operator, $\mathbb{E}_q$, is taken with respect to $\mathbf{X}$ that follows $q(\mathbf{x};\boldsymbol{\theta}')$.

**Lemma C.2.** *MCE is asymptotically normal, i.e.,*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) \xrightarrow{d} N(0, J^{-1}IJ^{-1}),$$

*where*

$$J = -\mathbb{E}_q\left[h(\mathbf{X})w(\mathbf{X})\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \log q\left(\mathbf{X}; \boldsymbol{\theta}^*\right)\right]$$

*and*

$$I = \mathbb{E}_q\left[h^2(\mathbf{X})w^2(\mathbf{X})\nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X}; \boldsymbol{\theta}^*\right)\left(\nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X}; \boldsymbol{\theta}^*\right)\right)^T\right].$$

***Proof.*** *By definition of MCE, we know*

$$0 = \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i)\nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X}_i; \hat{\boldsymbol{\theta}}\right).$$

*By using the mean value theorem on the right-hand side, we have*

$$\begin{aligned} 0 = &\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i)\nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X}_i; \boldsymbol{\theta}^*\right) \\ &+ \left(\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i)\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \log q\left(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}\right)\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right), \end{aligned}$$

*where $\tilde{\boldsymbol{\theta}}$ is an intermediate value between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$. Rearranging the terms leads to*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) = J_n^{-1}S_n,$$

*where*

$$J_n = -\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i)\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \log q\left(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}\right)$$

*and*

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(\mathbf{X}_i) w(\mathbf{X}_i) \nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X}_i; \boldsymbol{\theta}^*\right).$$

*By the weak law of large numbers (Keener, 2010) and the consistency of* $\hat{\boldsymbol{\theta}}$ *in Lemma C.1, it follows that*

$$J_n \xrightarrow{P} J.$$

*By the central limit theorem (Keener, 2010), we also have*

$$S_n \xrightarrow{d} N(0, I),$$

*where*

$$\begin{aligned} I &= Var_q\left[h(\mathbf{X}) w(\mathbf{X}) \nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X}; \boldsymbol{\theta}^*\right)\right] \\ &= \mathbb{E}_q\left[h^2(\mathbf{X}) w^2(\mathbf{X}) \nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X}; \boldsymbol{\theta}^*\right) \left(\nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X}; \boldsymbol{\theta}^*\right)\right)^T\right] \\ &\quad - \left(\mathbb{E}_q\left[h(\mathbf{X}) w(\mathbf{X}) \nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X}; \boldsymbol{\theta}^*\right)\right]\right)^2 \\ &= \mathbb{E}_q\left[h^2(\mathbf{X}) w^2(\mathbf{X}) \nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X}; \boldsymbol{\theta}^*\right) \left(\nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X}; \boldsymbol{\theta}^*\right)\right)^T\right]. \end{aligned}$$

*Here, we used*

$$\begin{aligned}
\mathbb{E}_q\left[h(\mathbf{X})w(\mathbf{X})\nabla_{\boldsymbol{\theta}}\log q\left(\mathbf{X};\boldsymbol{\theta}^*\right)\right] &= \mathbb{E}_q\left[h(\mathbf{X})w(\mathbf{X})\frac{\nabla_{\boldsymbol{\theta}}q\left(\mathbf{X};\boldsymbol{\theta}^*\right)}{q\left(\mathbf{X};\boldsymbol{\theta}^*\right)}\right] \\
&= \int h(\mathbf{x})\frac{f(\mathbf{x})}{q(\mathbf{x};\boldsymbol{\theta}')}\frac{\nabla_{\boldsymbol{\theta}}q\left(\mathbf{x};\boldsymbol{\theta}^*\right)}{q\left(\mathbf{x};\boldsymbol{\theta}^*\right)}q(\mathbf{x};\boldsymbol{\theta}')\ \mathrm{d}\mathbf{x} \\
&= \int h(\mathbf{x})f(\mathbf{x})\ \frac{\nabla_{\boldsymbol{\theta}}q\left(\mathbf{x};\boldsymbol{\theta}^*\right)}{q\left(\mathbf{x};\boldsymbol{\theta}^*\right)}\,\mathrm{d}\mathbf{x} \\
&= \left(\int h(\mathbf{x})f(\mathbf{x})\ \mathrm{d}\mathbf{x}\right)\ \nabla_{\boldsymbol{\theta}}\int q\left(\mathbf{x};\boldsymbol{\theta}^*\right)\ \mathrm{d}\mathbf{x} \\
&= 0, \qquad \text{(C.1)}
\end{aligned}$$

*where*

$$q\left(\mathbf{x};\boldsymbol{\theta}^*\right) = \frac{h(\mathbf{x})f(\mathbf{x})}{\int h(\mathbf{x})f(\mathbf{x})\ \mathrm{d}\mathbf{x}}$$

*by Assumption IV.1. The equation in* (C.1) *holds because the regularity conditions allow the interchange of integration and differentiation and* $q\left(\mathbf{x};\boldsymbol{\theta}^*\right)$ *is integrated to one. Therefore, it follows that* $\mathbb{E}\left[S_n\right] = 0$.

*Note that*

$$Var\left[J^{-1}S_n\right] = J^{-1}Var\left[S_n\right]J^{-1}.$$

*By Slutsky's theorem (Jiang, 2010, Theorem 2.13), it follows that*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) \xrightarrow{d} N(0, J^{-1}IJ^{-1}).$$

□

With Lemmas C.1 and C.2, we now prove Theorem IV.3. Specifically, we want to

derive the bias, $\mathbb{E}\left[\bar{\mathcal{C}}_{\hat{\boldsymbol{\theta}}} - \mathcal{C}_{\hat{\boldsymbol{\theta}}}\right]$, where

$$\bar{\mathcal{C}}_{\hat{\boldsymbol{\theta}}} = -\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i)\log q\Big(\mathbf{X}_i;\hat{\boldsymbol{\theta}}\Big)$$

and

$$\mathcal{C}_{\hat{\boldsymbol{\theta}}} = -\mathbb{E}_q\left[h(\mathbf{X})w(\mathbf{X})\log q\left(\mathbf{X};\hat{\boldsymbol{\theta}}\right)\right].$$

Note that $\hat{\boldsymbol{\theta}}$ is a function of $\mathbf{X}_1,\ldots,\mathbf{X}_n$ and considered as a constant by the expectation operator, $\mathbb{E}_q$, that is taken with respect to $\mathbf{X}$ following $q(\mathbf{x};\boldsymbol{\theta}')$. Also, recall that the normalizing constant of $q^*$ is

$$K_{q^*} = \int h(\mathbf{x})f(\mathbf{x})\ \mathrm{d}\mathbf{x}.$$

**Theorem IV.3**

$$\mathbb{E}\left[\bar{\mathcal{C}}_{\hat{\boldsymbol{\theta}}} - \mathcal{C}_{\hat{\boldsymbol{\theta}}}\right] = -K_{q^*}\frac{d}{n} + o\left(\frac{1}{n}\right),$$

*where the expectation is taken with respect to the data,* $\mathbf{X}_1,\ldots,\mathbf{X}_n$.

***Proof.*** *The Taylor expansion (Keener, 2010) leads to*

$$\begin{aligned}
\mathcal{C}_{\hat{\boldsymbol{\theta}}} &= -\mathbb{E}_q\left[h(\mathbf{X})w(\mathbf{X})\log q\left(\mathbf{X};\hat{\boldsymbol{\theta}}\right)\right] \\
&= -\mathbb{E}_q\left[h(\mathbf{X})w(\mathbf{X})\left(\log q\left(\mathbf{X};\boldsymbol{\theta}^*\right) + \left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)^T\nabla_{\boldsymbol{\theta}}\log q\left(\mathbf{X};\boldsymbol{\theta}^*\right)\right.\right. \\
&\quad \left.\left. + \frac{1}{2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)^T\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\log q\left(\mathbf{X};\boldsymbol{\theta}^*\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right)\right] + o_p\left(\frac{1}{n}\right).
\end{aligned}$$

*The error bound holds by Lemma C.2 and the regularity condition on the third derivative of log-likelihood, similar to the condition necessary to establish AIC (Cavanaugh*

*and Neath, 2014). Recall that* $\hat{\boldsymbol{\theta}}$ *is not a function of* $\mathbf{X}$ *and that the expectation operator,* $\mathbb{E}_q$*, is taken with respect to* $\mathbf{X}$*. The expectation of the first-order term involving the score function is zero in a similar manner that leads to* (C.1)*. Define* $\delta_n = \sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)$*. Then, we can simplify the expression of* $\mathcal{C}_{\hat{\boldsymbol{\theta}}}$ *as*

$$\mathcal{C}_{\hat{\boldsymbol{\theta}}} = \mathcal{C}_{\boldsymbol{\theta}^*} + \frac{1}{2n}\delta_n^T J \delta_n + o_p\left(\frac{1}{n}\right).$$

*Similarly, we can express*

$$\begin{aligned}
\bar{\mathcal{C}}_{\hat{\boldsymbol{\theta}}} &= -\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i) w(\mathbf{X}_i) \log q\left(\mathbf{X}_i; \hat{\boldsymbol{\theta}}\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i) w(\mathbf{X}_i) \left(\log q\left(\mathbf{X}_i; \boldsymbol{\theta}^*\right) + \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)^T \nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X}_i; \boldsymbol{\theta}^*\right)\right. \\
&\quad \left. + \frac{1}{2}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)^T \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \log q\left(\mathbf{X}_i; \boldsymbol{\theta}^*\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)\right) + o_p\left(\frac{1}{n}\right) \\
&= \bar{\mathcal{C}}_{\boldsymbol{\theta}^*} - \frac{1}{n}\delta_n^T S_n + \frac{1}{2n}\delta_n^T J_n^* \delta_n + o_p\left(\frac{1}{n}\right),
\end{aligned}$$

*where* $J_n^*$ *is* $-\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i) w(\mathbf{X}_i) \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \log q\left(\mathbf{X}_i; \boldsymbol{\theta}^*\right)$*. By the weak law of large numbers (Keener, 2010), we have* $J_n^* - J = o_p(1)$*, so we can further express*

$$\bar{\mathcal{C}}_{\hat{\boldsymbol{\theta}}} = \bar{\mathcal{C}}_{\boldsymbol{\theta}^*} - \frac{1}{n}\delta_n^T S_n + \frac{1}{2n}\delta_n^T J \delta_n + o_p\left(\frac{1}{n}\right).$$

*Note that* $\mathbb{E}\left[\bar{\mathcal{C}}_{\boldsymbol{\theta}^*}\right] = \mathcal{C}_{\boldsymbol{\theta}^*}$*. Thus, the bias of interest is*

$$\begin{aligned}
\mathbb{E}\left[\bar{\mathcal{C}}_{\hat{\boldsymbol{\theta}}} - \mathcal{C}_{\hat{\boldsymbol{\theta}}}\right] &= \mathbb{E}\left[\bar{\mathcal{C}}_{\boldsymbol{\theta}^*} - \frac{1}{n}\delta_n^T S_n + \frac{1}{2n}\delta_n^T J \delta_n - \left(\mathcal{C}_{\boldsymbol{\theta}^*} + \frac{1}{2n}\delta_n^T J \delta_n\right)\right] + o\left(\frac{1}{n}\right) \\
&= \mathbb{E}\left[\bar{\mathcal{C}}_{\boldsymbol{\theta}^*} - \frac{1}{n}\delta_n^T S_n - \mathcal{C}_{\boldsymbol{\theta}^*}\right] + o\left(\frac{1}{n}\right) \\
&= -\mathbb{E}\left[\frac{1}{n}\delta_n^T S_n\right] + o\left(\frac{1}{n}\right)
\end{aligned}$$

*Because* $\delta_n = J_n^{-1} S_n$ *and* $J_n - J = o_p(1)$*, we have*

$$\begin{aligned}
\mathbb{E}\left[\bar{\mathcal{C}}_{\hat{\boldsymbol{\theta}}} - \mathcal{C}_{\hat{\boldsymbol{\theta}}}\right] &= -\mathbb{E}\left[\frac{1}{n} S_n^T J^{-1} S_n\right] + o\left(\frac{1}{n}\right) \\
&= -\frac{1}{n}\mathbb{E}\left[tr\left(S_n^T J^{-1} S_n\right)\right] + o\left(\frac{1}{n}\right) \\
&= -\frac{1}{n}\mathbb{E}\left[tr\left(J^{-1} S_n S_n^T\right)\right] + o\left(\frac{1}{n}\right) \\
&= -\frac{1}{n} tr\left(J^{-1}\mathbb{E}\left[S_n S_n^T\right]\right) + o\left(\frac{1}{n}\right) \\
&= -\frac{1}{n} tr\left(J^{-1} I\right) + o\left(\frac{1}{n}\right). \qquad \text{(C.2)}
\end{aligned}$$

*To further simplify the trace,* $tr(J^{-1}I)$*, we derive the simpler expressions of* $J$ *and* $I$ *as follows:*

$$\begin{aligned}
J &= -\mathbb{E}_q\left[h(\mathbf{X}) w(\mathbf{X}) \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \log q\left(\mathbf{X}; \boldsymbol{\theta}^*\right)\right] \\
&= -\mathbb{E}_q\left[h(\mathbf{X}) w(\mathbf{X}) \left(-\frac{1}{q^2\left(\mathbf{X}; \boldsymbol{\theta}^*\right)} \nabla_{\boldsymbol{\theta}} q\left(\mathbf{X}; \boldsymbol{\theta}^*\right) \left(\nabla_{\boldsymbol{\theta}} q\left(\mathbf{X}; \boldsymbol{\theta}^*\right)\right)^T \right.\right. \\
&\qquad \left.\left. + \frac{1}{q\left(\mathbf{X}; \boldsymbol{\theta}^*\right)} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} q\left(\mathbf{X}; \boldsymbol{\theta}^*\right)\right)\right],
\end{aligned}$$

*where the second term is zero because*

$$\begin{aligned}
&-\mathbb{E}_q\left[h(\mathbf{X}) w(\mathbf{X}) \left(\frac{1}{q\left(\mathbf{X}; \boldsymbol{\theta}^*\right)} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} q\left(\mathbf{X}; \boldsymbol{\theta}^*\right)\right)\right] \\
&= -\int h(\mathbf{x}) \frac{f(\mathbf{x})}{q(\mathbf{x}; \boldsymbol{\theta}')} \left(\frac{1}{q\left(\mathbf{x}; \boldsymbol{\theta}^*\right)} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} q\left(\mathbf{x}; \boldsymbol{\theta}^*\right)\right) q(\mathbf{x}; \boldsymbol{\theta}')\, \mathrm{d}\mathbf{x} \\
&= -K_{q^*} \int \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} q\left(\mathbf{x}; \boldsymbol{\theta}^*\right) \mathrm{d}\mathbf{x} \\
&= -K_{q^*} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \int q\left(\mathbf{x}; \boldsymbol{\theta}^*\right) \mathrm{d}\mathbf{x} \\
&= 0
\end{aligned}$$

*by Assumption IV.1 and*

$$q\left(\mathbf{x};\boldsymbol{\theta}^*\right) = \frac{1}{K_{q^*}} h(\mathbf{x}) f(\mathbf{x})\,. \tag{C.3}$$

*The remaining first term can be rewritten as*

$$\begin{aligned}
J &= -\mathbb{E}_q\left[h(\mathbf{X})w(\mathbf{X})\left(-\frac{1}{q^2\left(\mathbf{X};\boldsymbol{\theta}^*\right)}\nabla_{\boldsymbol{\theta}} q\left(\mathbf{X};\boldsymbol{\theta}^*\right)\left(\nabla_{\boldsymbol{\theta}} q\left(\mathbf{X};\boldsymbol{\theta}^*\right)\right)^T\right)\right] \\
&= -\int h(\mathbf{x}) f(\mathbf{x})\left(-\frac{1}{q^2\left(\mathbf{X};\boldsymbol{\theta}^*\right)}\nabla_{\boldsymbol{\theta}} q\left(\mathbf{X};\boldsymbol{\theta}^*\right)\left(\nabla_{\boldsymbol{\theta}} q\left(\mathbf{X};\boldsymbol{\theta}^*\right)\right)^T\right) \mathrm{d}\mathbf{x} \\
&= K_{q^*}\int \frac{1}{q\left(\mathbf{X};\boldsymbol{\theta}^*\right)}\nabla_{\boldsymbol{\theta}} q\left(\mathbf{X};\boldsymbol{\theta}^*\right)\left(\nabla_{\boldsymbol{\theta}} q\left(\mathbf{X};\boldsymbol{\theta}^*\right)\right)^T \mathrm{d}\mathbf{x}.
\end{aligned}$$

*On the other hand,*

$$\begin{aligned}
I &= \mathbb{E}_q\left[h^2(\mathbf{X})w^2(\mathbf{X})\nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X};\boldsymbol{\theta}^*\right)\left(\nabla_{\boldsymbol{\theta}} \log q\left(\mathbf{X};\boldsymbol{\theta}^*\right)\right)^T\right] \\
&= \int h^2(\mathbf{x})\frac{f^2(\mathbf{x})}{q^2(\mathbf{x};\boldsymbol{\theta}')}\frac{1}{q^2\left(\mathbf{x};\boldsymbol{\theta}^*\right)}\nabla_{\boldsymbol{\theta}} q\left(\mathbf{x};\boldsymbol{\theta}^*\right)\left(\nabla_{\boldsymbol{\theta}} q\left(\mathbf{x};\boldsymbol{\theta}^*\right)\right)^T q(\mathbf{x};\boldsymbol{\theta}')\,\mathrm{d}\mathbf{x} \\
&= K_{q^*}^2\int \frac{1}{q(\mathbf{x};\boldsymbol{\theta}')}\nabla_{\boldsymbol{\theta}} q\left(\mathbf{x};\boldsymbol{\theta}^*\right)\left(\nabla_{\boldsymbol{\theta}} q\left(\mathbf{x};\boldsymbol{\theta}^*\right)\right)^T \mathrm{d}\mathbf{x} \\
&= K_{q^*}^2\int \left(\frac{1}{q\left(\mathbf{x};\boldsymbol{\theta}^*\right)} + o(1)\right)\nabla_{\boldsymbol{\theta}} q\left(\mathbf{x};\boldsymbol{\theta}^*\right)\left(\nabla_{\boldsymbol{\theta}} q\left(\mathbf{x};\boldsymbol{\theta}^*\right)\right)^T \mathrm{d}\mathbf{x} \\
&= K_{q^*} J + o(1)\int \nabla_{\boldsymbol{\theta}} q\left(\mathbf{x};\boldsymbol{\theta}^*\right)\left(\nabla_{\boldsymbol{\theta}} q\left(\mathbf{x};\boldsymbol{\theta}^*\right)\right)^T \mathrm{d}\mathbf{x},
\end{aligned}$$

*where we used the fact that $\boldsymbol{\theta}'$ is the MCE at the last CE outer iteration. Thus, $\boldsymbol{\theta}'$ is also a consistent estimator of $\boldsymbol{\theta}^*$ by Assumption IV.2 and Lemma C.1. Plugging this result to* (C.2) *leads to*

$$\begin{aligned}
\mathbb{E}\left[\bar{\mathcal{C}}_{\hat{\boldsymbol{\theta}}} - \mathcal{C}_{\hat{\boldsymbol{\theta}}}\right] &= -\frac{1}{n} tr\left(J^{-1} K_{q^*} J\right) + o\left(\frac{1}{n}\right) \\
&= -K_{q^*}\frac{d}{n} + o\left(\frac{1}{n}\right).
\end{aligned}$$

□

## C.2 Approximation of $N_i$ for SIS

Chapter II uses a metamodel to approximate the optimal $N_i$ in (2.13). To circumvent the need for building a metamodel, we use an asymptotic approximation of the optimal allocation size,

$$N_i = \frac{\sqrt{s(\mathbf{X}_i)\left(1-s(\mathbf{X}_i)\right)}f(\mathbf{X}_i)\,/q(\mathbf{X}_i)}{\sum_{j=1}^{m}\sqrt{s(\mathbf{X}_j)\left(1-s(\mathbf{X}_j)\right)}f(\mathbf{X}_j)\,/q(\mathbf{X}_j)}\cdot n, \quad i=1,\ldots,m,$$

in (2.10).

First, for a large $n \gg \max_{i=1}^{m}\left(1-s(\mathbf{X}_i)\right)/s(\mathbf{X}_i)$, we can approximate

$$\begin{aligned} q(\mathbf{X}_i) &= \frac{1}{C_{q1}} f(\mathbf{X}_i)\sqrt{\frac{1}{n}s(\mathbf{X}_i)\left(1-s(\mathbf{X}_i)\right)+s(\mathbf{X}_i)^2} \\ &\approx \frac{1}{C_{q1}} f(\mathbf{X}_i)\sqrt{s(\mathbf{X}_i)^2} \\ &= \frac{1}{C_{q1}} f(\mathbf{X}_i)\, s(\mathbf{X}_i) \end{aligned}$$

for any $i=1,\ldots,m$. This asymptotic approximation may be not good for some $N_i$ if $s(\mathbf{X}_i)$ is close to zero. However, in that case, $q(\mathbf{X}_i)$ is small too, and such $\mathbf{X}_i$ is unlikely to be sampled in the first place. Therefore, we can approximate

$$s(\mathbf{X}_i) \approx C_{q1}\frac{q(\mathbf{X}_i)}{f(\mathbf{X}_i)},$$

where $f(\mathbf{X}_i)$ and $q(\mathbf{X}_i)$ are known.

Furthermore, for a large $n \gg \max_{i=1}^{m} (1 - s(\mathbf{X}_i))/s(\mathbf{X}_i)$, we can also approximate

$$\begin{aligned} C_{q1} &= \int_{\mathcal{X}_f} f(\mathbf{x}) \sqrt{\frac{1}{n} s(\mathbf{x}) \cdot (1 - s(\mathbf{x})) + s(\mathbf{x})^2} \, \mathrm{d}\mathbf{x} \\ &\approx \int_{\{\mathbf{x}:\, s(\mathbf{x}) > 1/(n+1)\}} f(\mathbf{x}) \sqrt{\frac{1}{n} s(\mathbf{x}) \cdot (1 - s(\mathbf{x})) + s(\mathbf{x})^2} \, \mathrm{d}\mathbf{x} \\ &\approx \int_{\{\mathbf{x}:\, s(\mathbf{x}) > 1/(n+1)\}} f(\mathbf{x})\, s(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ &\approx \hat{P}_{SIS}. \end{aligned}$$

Thus, it follows that

$$s(\mathbf{X}_i) \approx \frac{\hat{P}_{SIS}}{w(\mathbf{X}_i)}.$$

Therefore,

$$\begin{aligned} N_i &\propto \sqrt{s(\mathbf{X}_i)\,(1 - s(\mathbf{X}_i))} f(\mathbf{X}_i) \,/ q(\mathbf{X}_i) \\ &\approx \sqrt{\frac{\hat{P}_{SIS}}{w(\mathbf{X}_i)} \left(1 - \frac{\hat{P}_{SIS}}{w(\mathbf{X}_i)}\right)} w(\mathbf{X}_i) \\ &\propto \sqrt{w(\mathbf{X}_i) - \hat{P}_{SIS}}. \end{aligned}$$

Although it does not happen frequently, if $w(\mathbf{X}_i) - \hat{P}_{SIS} \leq 0$, then we set the corresponding $N_i$ as 1, the smallest allocation possible to maintain the unbiasedness of the SIS estimator.

# BIBLIOGRAPHY

# BIBLIOGRAPHY


Ackley, D. H. (1987), *A connectionist machine for genetic hillclimbing*, Boston: Kluwer Academic Publishers.

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

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

Au, S. K., and J. L. Beck (2003), Important sampling in high dimensions, *Structural Safety*, *25*(2), 139–163.

Botev, Z. I., D. P. Kroese, and T. Taimre (2007), Generalized cross-entropy methods with applications to rare-event simulation and optimization, *Simulation*, *83*(11), 785–806.

Botev, Z. I., D. P. Kroese, R. Y. Rubinstein, P. L'Ecuyer, et al. (2013), The cross-entropy method for optimization, *Machine Learning: Theory and Applications, V. Govindaraju and C. R. Rao, Eds, Chennai: Elsevier*, *31*, 35–59.

Byon, E., L. Ntaimo, and Y. Ding (2010), Optimal maintenance strategies for wind power systems under stochastic weather conditions, *IEEE Transactions on Reliability*, *59*(2), 393–404.

Byon, E., Y. Choe, and N. Yampikulsakul (2015), Adaptive learning in time-variant processes with application to wind power systems, *IEEE Transactions on Automation Science and Engineering*, forthcoming. Available at http://dx.doi.org/10.1109/TASE.2015.2440093.

Cannamela, C., J. Garnier, and B. Iooss (2008), Controlled stratification for quantile estimation, *Annals of Applied Statistics*, *2*(4), 1554–1580.

Casella, G., and R. L. Berger (2002), *Statistical inference*, Australia: Thomson Learning.

Cavanaugh, J. E., and A. A. Neath (2014), Akaike's information criterion: Background, derivation, properties, and refinements, in *International Encyclopedia of Statistical Science*, pp. 26–29, Springer.

Choe, Y., and E. Byon (2015), Computationally efficient uncertainty minimization in wind turbine extreme load assessments, *Submitted*.

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

Choulakian, V., and M. A. Stephens (2001), Goodness-of-fit tests for the generalized Pareto distribution, *Technometrics*, *43*(4), 478–484.

Coles, S. G. (2001), *An introduction to statistical modeling of extreme values*, London: Springer.

Courant, R., and D. Hilbert (1989), *Methods of Mathematical Physics.*, New York: Wiley.

D'Auria, F., A. Bousbia-Salah, A. Petruzzi, and A. Del Nevo (2006), State of the art in using best estimate calculation tools in nuclear technology, *Nuclear Engineering and Technology*, *38*(1), 11–32.

De Boer, P., D. P. Kroese, S. Mannor, and R. Y. Rubinstein (2005), A tutorial on the cross-entropy method, *Annals of Operations Research*, *134*(1), 19–67.

Dubourg, V., B. Sudret, and F. Deheeger (2013), Metamodel-based importance sampling for structural reliability analysis, *Probabilistic Engineering Mechanics*, *33*, 47–57.

Efron, B., and R. Tibshirani (1993), *An introduction to the bootstrap*, CRC Press, LLC.

Figueiredo, M. A., and A. K. Jain (2002), Unsupervised learning of finite mixture models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(3), 381–396.

Geweke, J. (2005), *Contemporary Bayesian econometrics and statistics*, Hoboken, New Jersey: John Wiley & Sons.

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

Graf, P. A., G. Stewart, M. Lackner, K. Dykes, and P. Veers (2015), High-throughput computation and the applicability of monte carlo integration in fatigue load estimation of floating offshore wind turbines, *Wind Energy*, doi:10.1002/we.1870.

Green, P. J., and B. W. Silverman (1994), *Nonparametric regression and generalized linear models*, London: Chapman and Hall.

Hastie, T. J., and R. J. Tibshirani (1990), *Generalized additive models*, London: Chapman and Hall.

Heidelberger, P. (1995), Fast simulation of rare events in queueing and reliability models, *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *5*(1), 43–85.

Hesterberg, T. C. (1995), Weighted average importance sampling and defensive mixture distributions, *Technometrics*, *37*(2), 185–194.

Huang, D., T. T. Allen, W. I. Notz, and N. Zeng (2006), Global optimization of stochastic black-box systems via sequential kriging meta-models, *Journal of Global Optimization*, *34*(3), 441–466.

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

Jiang, J. (2010), *Large sample techniques for statistics*, New York: Springer–Verlag.

Jonkman, B. J. (2009), TurbSim user's guide: version 1.50, *Tech. Rep. NREL/TP-500-46198*, National Renewable Energy Laboratory, Golden, Colorado.

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

Jonkman, J. M., S. Butterfield, W. Musial, and G. Scott (2009), Definition of a 5-MW reference wind turbine for offshore system development, *Tech. Rep. NREL/TP-500-38060*, National Renewable Energy Laboratory, Golden, Colorado.

Joseph, V. R. (2006), Limit kriging, *Technometrics*, *48*(4), 458–466.

Kahn, H., and A. W. Marshall (1953), Methods of reducing sample size in Monte Carlo computations, *Journal of the Operations Research Society of America*, *1*(5), 263–278.

Keener, R. W. (2010), *Theoretical Statistics: Topics for a Core Course*, New York: Springer-Verlag.

Koopman, S. J., N. Shephard, and D. Creal (2009), Testing the assumptions behind importance sampling, *Journal of Econometrics*, *149*(1), 2–11.

Kroese, D. P., T. Taimre, and Z. I. Botev (2011), *Handbook of Monte Carlo Methods*, New York: John Wiley and Sons.

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

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

Manuel, L., H. H. Nguyen, and M. F. Barone (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*, Grapevine, Texas.

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

Plumlee, M., and R. Tuo (2014), Building accurate emulators for stochastic simulations via quantile kriging, *Technometrics*, *56*(4), 466–473.

Rigby, R. A., and D. M. Stasinopoulos (2005), Generalized additive models for location, scale and shape, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507–554.

Rubinstein, R. (1999), The cross-entropy method for combinatorial and continuous optimization, *Methodology and Computing in Applied Probability*, *1*(2), 127–190.

Rubinstein, R. (2005), A stochastic minimum cross-entropy method for combinatorial optimization and rare-event estimation, *Methodology and Computing in Applied Probability*, *7*(1), 5–50.

Rubinstein, R. Y., and A. Shapiro (1993), *Discrete event systems: Sensitivity analysis and stochastic optimization by the score function method*, Chichester: John Wiley & Sons Ltd.

Schwarz, G. (1978), Estimating the dimension of a model, *The Annals of Statistics*, *6*(2), 461–464.

Seber, G. A. F., and A. J. Lee (2003), *Linear regression analysis*, Hoboken: John Wiley & Sons.

Shampine, L. F. (2008), Vectorized adaptive quadrature in MATLAB, *Journal of Computational and Applied Mathematics*, *211*(2), 131–140.

Staid, A., S. D. Guikema, R. Nateghi, S. M. Quiring, and M. Z. Gao (2014), Simulation of tropical cyclone impacts to the us power system under climate change scenarios, *Climatic Change*, *127*(3-4), 535–546.

Stephens, M. A. (1974), EDF statistics for goodness of fit and some comparisons, *Journal of the American Statistical Association*, *69*(347), 730–737.

Van der Vaart, A. W. (1998), *Asymptotic statistics*, Cambridge: Cambridge University Press.

Wang, H., and X. Zhou (2015), A cross-entropy scheme for mixtures, *ACM Transactions on Modeling and Computer Simulation*, *25*(1), 6:1–6:20.

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

Zhang, N., and D. W. Apley (2014), Fractional brownian fields for response surface metamodeling, *Journal of Quality Technology*, *46*(4), 285.

Zhang, N., and D. W. Apley (2015), Brownian integrated covariance functions for gaussian process modeling: Sigmoidal versus localized basis functions, *Journal of the American Statistical Association*, forthcoming. Available at http://dx.doi.org/10.1080/01621459.2015.1077711.