An Efficient Budget Allocation Approach for Quantifying the Impact of Input Uncertainty in Stochastic Simulation

YUAN YI, WEI XIE, Rensselaer Polytechnic Institute

Simulations are often driven by input models estimated from finite real-world data. When we use simulations to assess the performance of a stochastic system, there exist two sources of uncertainty in the performance estimates: input and simulation estimation uncertainty. In this paper, we develop a budget allocation approach that can efficiently employ the potentially tight simulation resource to construct a percentile confidence interval quantifying the impact of the input uncertainty on the system performance estimates, while controlling the simulation estimation error. Specifically, non-parametric bootstrap is used to generate samples of input models quantifying both input distribution family and parameter value uncertainty. Then, the direct simulation is used to propagate the input uncertainty to the output response. Since each simulation run could be computationally expensive, given a tight simulation budget, we propose an efficient budget allocation approach that can balance the finite sampling error introduced by using finite bootstrapped samples to quantify the input uncertainty and the system response estimation error introduced by using finite replications to estimate the system response at each bootstrapped sample. Our approach is theoretically supported, and empirical studies also demonstrate that it has better and more robust performance than the direct bootstrapping.

Categories and Subject Descriptors: I.6.6 [**Simulation and Modeling**]: Simulation Output Analysis

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Confidence interval, budget allocation, input uncertainty, nonparametric bootstrap, percentile estimation

1. INTRODUCTION

When simulation is used to assess the performance of stochastic systems, the input models used to drive the simulations are often estimated from finite real-world data. Thus, there exist both input estimation uncertainty, called *the input uncertainty*, and the simulation estimation uncertainty in the system performance estimates. Ignoring either source of uncertainty could lead to unfounded confidence in the simulation assessment of system performance [Barton and Schruben 2001; Barton 2012; Xie et al. 2014a]. *Given a tight simulation budget, we want to efficiently estimate the impact of the input uncertainty on the system performance estimates, while controlling the simulation estimation error.* In this paper, we focus on the system mean response. However, our approach could be extended to other performance estimates, such as variances and probabilities.

There are various approaches proposed in the simulation literature to quantify the input uncertainty and they can be divided into frequentist and Bayesian approaches. The frequentist approaches typically use the sampling distributions of point estimators of input models to quantify the input uncertainty. Since it could be hard to get the exact sampling distributions in many situations, the asymptotic approximation,

Author's addresses: Y. Yi, email: yiy2@rpi.edu; Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180-3590. W. Xie (corresponding author), email: xiew3@rpi.edu; Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180-3590.

DOI: 0000001.0000001

including the normal approximation and the bootstrap, is often used to quantify the input uncertainty. The Bayesian approaches typically derive the posteriors of input models to quantify the input uncertainty. Notice that freqentist and Bayesian approaches have totally different perspective on quantifying uncertainty; see Xie et al. [2014a] for the detailed description.

When input distribution families are known and the input models can be specified by a finite number of parameters, the sampling distributions of parameters/moments or the posterior distributions of parameters can be developed to quantify the input uncertainty; see for example Cheng and Holland [1997], Barton et al. [2014], Ng and Chick [2006], Xie et al. [2014a; 2014b], Biller and Corlu [2011]. Then, either the direct simulation that runs simulations at each sample of input distributions or a metamodel can be used to propagate the input uncertainty to the output mean. Since metamodels explore the relationship between the mean response at different samples of input parameters/moments, they could efficiently propagate the input uncertainty to the output and reduce the simulation estimation uncertainty [Xie et al. 2014a]. However, a large number of input parameters could make it computationally expensive to construct a metamodel [Barton 2012]. In addition, unlike the direct simulation, metamodels require some prior information about the underlying mean response surface. It could be as strong as a global parametric trend [Chick 1997] or as weak as local smoothness and continuity [Xie et al. 2014a].

We typically do not know the families of input models. Various approaches, including Bayesian Model Average (BMA) [Chick 2001; Zouaoui and Wilson 2003], nonparametric Bayesian approaches [Xie et al. 2017], and nonparametric bootstrap [Barton and Schruben 1993; Barton 2007] were proposed in the simulation literature to account for both input distribution family and parameter value uncertainty. In BMA, the posterior probabilities of a few pre-specified candidate parametric families are developed to account for the model selection error. BMA is based the assumption that all data come from one of candidate distributions [Bishop 2006]. Without strong prior information about input models, it could be difficult to select an appropriate pool of candidate parametric families. In nonparametric bootstrap or Bayesian approaches, the sampling or posterior distribution of flexible input models can account for input uncertainty. They can avoid the potential issue of BMA. However, the nonparametric Bayesian approaches introduced in Xie et al. [2017] require Markov Chain Monte Carlo (MCMC) sampling to generate posterior samples of input models. In addition, for different priors, we need to derive the posterior distribution for input models.

In this paper, the nonparametric bootstrap is used to quantify the input uncertainty. Since bootstrapped empirical distributions can not be specified by a finite number of moments, it is challenging to construct a metamodel as a function of empirical distributions. Thus, the direct simulation is used to propagate the input uncertainty to the output mean.

Barton [2012] demonstrates that the t-based confident intervals (CIs) are not appropriate for quantifying the impact of the input uncertainty on the system performance estimates because of the skewness. *Thus, given a tight simulation budget, we focus on efficiently constructing a percentile CI to quantify the impact of the input uncertainty.* To have a precise estimation on the percentile CI, it is typically recommended to have a few thousands samples of input distributions quantifying the input uncertainty; see Xie et al. [2016b]. Without any prior information on which bootstrapped samples of input distributions contribute the most to the percentile CI estimation, the direct bootstrapping tends to *equally* allocate the simulation resource to the bootstrapped samples; see Barton and Schruben [1993; 2001] and Barton [2007]. Since the bootstrapped samples do not contribute equally to the percentile CI estimation, the equal allocation approach could not efficiently use the simulation resource to propagate the input un-

certainty to the output mean. For a complex stochastic system, since each simulation run could be computationally expensive, it is critical to develop a budget allocation approach that can efficiently use the simulation budget to estimate the percentile CI.

The percentile CI estimation can be formulated as *a nested simulation problem* where the outer level simulation is sampling from the nonparametric bootstrap to quantify the input uncertainty and the inner level is running the direct simulation to estimate the system mean response at each bootstrapped sample of input models. In this paper, a ranking and selection based approach is developed to efficiently estimate the percentile CI quantifying the impact of the input uncertainty. Before presenting our approach and contributions, we briefly review the studies in the simulation literature on the nested simulation and then ranking and selection.

Lee and Glynn [2003] considered the distribution function of conditional expectation in the discrete case. In order to find the optimal budget allocation to minimize the mean square error, they derived asymptotic bias and variance of nested simulation estimators with uniform and nonuniform inner sampling. Gordy and Juneja [2010] considered the probabilities of a large loss, value at risk and expected shortfall in the continuous case. To find the optimal allocation, they also studied asymptotic bias and variance of estimators with uniform inner sampling. To efficiently estimate the probabilities of a large loss, Broadie et al. [2011] proposed a myopic approach that sequentially allocates the inner simulations based on the marginal changes in the estimator. Sun et al. [2011] proposed an ANOVA-like estimator for the variance of the conditional expectation, and then found the optimal inner replications to minimize this estimator's variance.

Without strong prior information on the mean response of candidates, ranking and selection could be used to identify the systems with extreme mean performance. Nelson et al. [2001] and Boesel et al. [2003] proposed two-stage procedures to find the system with the largest mean from a finite number of candidates. One-sided screening is used to remove statistical inferior systems and then the indifference-zone (IZ) selection is used to assign the remaining simulation resource to the surviving systems. The restart is used to reduce the estimation bias introduced in the selection procedure [Boesel et al. 2003]. Instead of selecting the best system, Lesnevski et al. [2007] developed a multistage screening procedure to estimate the value of the maximum expected performance and further provide a CI for the estimation. The common random numbers (CRN) is used to efficiently screen out the inferior systems and control-variate estimators are employed to improve the system performance estimation. This multistage screening procedure was improved to be adaptive in Lesnevski et al. [2008]. In addition, ranking and selection was extended to efficiently estimate the conditional tail expectation in Lan et al. [2010]. They proposed a two-stage design for the expected shortfall estimation. In Stage I, a one-sided screening is developed to screen out samples that are not statistically likely to fall into the tail. Then, the restart is used to reduce the bias. In Stage II, the replication allocation is proportional to the sample variance so that we can minimize the width of the CI of expected shortfall accounting for both finite sampling and simulation estimation uncertainty.

Motivated by the ranking and selection literature, in this paper, given a tight simulation budget, we propose a budget allocation approach that can efficiently employ the simulation resource to estimate the percentile CI quantifying the impact of the input uncertainty. Specifically, we determine an appropriate number of bootstrapped samples of input distributions to balance the sampling uncertainty introduced by using finite bootstrapped samples to quantify the input uncertainty, and the system response estimation uncertainty introduced by using finite replications to estimate the system response at each bootstrapped sample. Further, when we allocate the simulation resource to bootstrapped samples of input distributions, since it is hard for a one-stage

approach to identify those samples that contribute the most to the percentile CI estimation, a sequential approach is developed to gradually find those important samples by two-sided screening and allocate more simulation resource there. Therefore, we could find the number of bootstrapped samples and replication allocation that can simultaneously control finite sampling and simulation estimation uncertainty. Our approach is theoretically supported. Empirical studies demonstrate that it can efficiently propagate the input uncertainty to the output mean, while reducing the simulation estimation uncertainty.

In sum, the main contributions of our paper are as follows.

1. We consider the situations where each simulation run is expensive. For example, a single run may take hours or even days. Given a tight simulation budget, our approach could efficiently employ the simulation resource to quantify the impact of input uncertainty. It can be applied to general situations where there is no strong prior information on the input models and the system response surface.
2. By following the framework in Lan et al. [2010], we propose a sequential procedure for the quantile estimation, including screening and estimation phases. In the screening phase, unlike ranking and selection approaches proposed in the literature that focus on finding the best candidate [Nelson et al. 2001; Boesel et al. 2003; Lesnevski et al. 2007] or a set of candidates that fall into the tail part [Lan et al. 2010], a two-sided screening introduced in our approach focuses on selecting the important samples that contribute the most to the quantile estimation. In the estimation phase, besides that the replication allocation is proportional to the sample variance (similar to Nelson et al. [2001]; Boesel et al. [2003]; Lesnevski et al. [2007]; Lan et al. [2010]), a variance reduction technique, the antithetic variance simulation algorithm, is used to improve the system performance estimation for the remaining candidates, which could obviously improve the percentile estimation.
3. Our approach can automatically adapt to the input uncertainty, the system mean and the simulation estimation uncertainty. The information obtained from the initial simulations could guide the search for the optimal parameters of the sequential procedure so that we can reduce the overall uncertainty of the quantile estimation.

The next section formally states our problem. In Section 3, we develop a sequential approach that can efficiently employ the simulation budget to estimate the percentile CI quantifying the impact of input uncertainty. In Section 4, we report numerical studies on an $M/M/1$ queue and a stochastic activity network. We conclude the paper in Section 5. All proofs are in the Appendix.

2. PROBLEM DESCRIPTION AND PROPOSED APPROACH

For the stochastic simulation driven by input models, denoted by F, the simulation output for the jth replication is

$$Y_j(F) = \mu(F) + \epsilon_j(F)$$

where $\mu(F)$ denotes the unknown system mean response and $\epsilon_j(F)$ represents the simulation error with mean zero. Notice that the simulation outputs depend on the choice of input models. The input models F could be composed of univariate and multivariate joint distributions. For the notation simplification, we only consider a single univariate input distribution.

Suppose that the simulation error follows a normal distribution, $\epsilon(F) \sim \mathrm{N}\left(0, \sigma_\epsilon^2(F)\right)$. This assumption holds for many situations where the simulation output is an average of a large number of more basic outputs. For example, when we study the steady state expected customer waiting time, the simulation output is the average of waiting time for many customers.

Denote the unknown true input model by F^c. It is estimated by finite real-world data, denoted by $\mathbf{X}_m \equiv \{X_1, X_2, \ldots, X_m\}$, with $X_i \overset{i.i.d.}{\sim} F^c$ and $i = 1, 2, \ldots, m$. Let the empirical distribution $\widehat{F}_m$ be the point estimator for F^c. Thus, the sampling distribution of $\widehat{F}_m$ can be used to quantify the input uncertainty. The impact of input uncertainty could be characterized by the induced sampling distribution of $\mu(\widehat{F}_m)$.

Since the t-based CIs are not appropriate for quantifying the impact of input uncertainty [Barton 2012], *in this paper, we are interested in constructing a percentile CI to quantify the impact of the input uncertainty on the system mean response.* Specifically, without loss of generality, we want to find a $(1-\beta)100\%$ one-sided percentile CI, denoted by $(-\infty, Q^c]$, such that

$$\Pr(\mu(F^c) \in (-\infty, Q^c]) = 1 - \beta.$$

In particular, if $\mu(\cdot)$ is known, we have the percentile $Q^c \equiv \inf\{q : \Pr[\mu(\widehat{F}_m) \leq q] \geq 1-\beta\}$. Notice that our approach could be easily extended to two-sided percentile CIs.

In general, it could be hard to have the closed-form sampling distribution for $\widehat{F}_m$. In this paper, we use the nonparametric bootstrap resampling to quantify the input uncertainty [Barton and Schruben 2001; Barton 2007]. Specifically, we draw with replacement from $\mathbf{X}_m$ to generate m bootstrapped samples, denoted by $\{X_1^{(1)}, X_2^{(1)}, \ldots, X_m^{(1)}\}$. Given these samples, we can construct a bootstrapped empirical distribution, denoted by $F^{(1)}$. By repeating this procedure, we can generate bootstrapped samples of the input model, denoted by $\{F^{(1)}, F^{(2)}, \ldots\}$, to quantify the input uncertainty. Denote the bootstrap distribution by $\widetilde{F}(\cdot|\mathbf{X}_m)$ with $F^{(b)} \sim \widetilde{F}(\cdot|\mathbf{X}_m)$ for $b = 1, 2, \ldots$. Therefore, the impact of input uncertainty on the system mean performance estimates can be quantified by the bootstrapped percentile CI, denoted by $(-\infty, Q]$, with the upper bound defined by

$$Q \equiv \inf\left\{q : \Pr\left[\mu\left(F^{(b)}\right) \leq q|\mathbf{X}_m\right] \geq 1-\beta \text{ with } F^{(b)} \sim \widetilde{F}(\cdot|\mathbf{X}_m)\right\}. \tag{1}$$

If $\mu(\cdot)$ is known and continuously differentiable in a neighborhood of F^c, the nonparametric bootstrap provides an asymptotically consistent estimation for the impact of input uncertainty. It also has a good finite-sample performance in many situations; see Hall [1992], Shao and Tu [1995] and Horowitz [2001].

When we use B bootstrapped samples of input distribution to quantify the input uncertainty, the percentile Q could be estimated with the order statistics $\mu_{[(1-\beta)B]}$, where $\mu_b \equiv \mu(F^{(b)})$ with $F^{(b)} \sim \widetilde{F}(\cdot|\mathbf{X}_m)$ for $b = 1, 2, \ldots, B$. The permutation, denoted by $[\cdot]$, is defined based on the system true mean response, $\mu_{[1]} \leq \mu_{[2]} \leq \cdots \leq \mu_{[B]}$. Suppose that $(1-\beta)B$ is an integer for simplicity. Given finite B, there exists *finite sampling estimation error* that characterizes the difference between Q and $\mu_{[(1-\beta)B]}$. It decreases to zero as the sample size B increases to infinity.

However, the underlying response surface $\mu(\cdot)$ is unknown in many situations. When we use the nonparametric bootstrap to quantify the input uncertainty, the bootstrapped empirical distribution could not be uniquely specified by finite moments. Thus, it could be challenging to build a metamodel to propagate the input uncertainty to the output. Hence, *the direct simulation approach* is used in this paper, which propagates the input uncertainty to the output mean by running simulations at each bootstrapped input distribution $F^{(b)}$ to estimate the mean response $\mu(F^{(b)})$ for $b = 1, 2, \ldots, B$. Specifically, suppose that we allocate n_b replications at $F^{(b)}$. For a given simulation budget, denoted by C, we have $\sum_{b=1}^{B} n_b = C$. Then, we could estimate Q by

the order statistics $\bar{Y}_{((1-\beta)B)}$, where $\bar{Y}_b = \sum_{j=1}^{n_b} Y_j(F^{(b)})/n_b$. The permutation, denoted by $(\cdot)$, is defined based on the estimated mean responses $\bar{Y}_{(1)} \leq \bar{Y}_{(2)} \leq \cdots \leq \bar{Y}_{(B)}$.

In the direct bootstrapping, we tend to equally allocate the simulation resource to all bootstrapped samples $\{F^{(1)}, F^{(2)}, \ldots, F^{(B)}\}$ [Barton 2007]. Since B is recommended to be at least one thousand [Xie et al. 2014a], given a tight simulation budget, the replications allocated to each bootstrapped sample of input distribution, $n = C/B$, could be small. For a given set of bootstrapped samples $\{F^{(1)}, F^{(2)}, \ldots, F^{(B)}\}$ used to quantify the input uncertainty, the difference between $\mu_{[(1-\beta)B]}$ and $\bar{Y}_{((1-\beta)B)}$ could be large. Since the percentile estimation is based on the *order statistics* $\bar{Y}_{((1-\beta)B)}$, we do not need to precisely estimate the mean response for the sample $F^{(b)}$ with μ_b far from Q. Thus, the equal allocation can not efficiently use the simulation resource.

To avoid the issue faced by the uniform allocation, in this paper, we develop a sequential approach that explores the system mean responses at the bootstrapped samples of input distribution, and gradually finds *important samples* that contribute the most to the percentile Q estimation or have high possibility to be selected as the $(1-\beta)B$-th smallest one. Then, we assign more simulation resource there. Thus, our approach can efficiently employ the computational resource to reduce the simulation estimation error introduced during propagating the input uncertainty to the output mean.

Therefore, when we use $\bar{Y}_{((1-\beta)B)}$ to estimate the percentile Q, there exist three sources of errors:

(1) The finite sampling error;
(2) The selection error, which means the sample we select for the percentile estimation based on $\mu(\cdot)$ and $\bar{Y}$ are different, $((1-\beta)B) \neq [(1-\beta)B]$;
(3) The system response estimation error, which means at each $F^{(b)}$, the system response is estimated with error, $\bar{Y}_b \neq \mu_b$, for $b = 1, 2, \ldots, B$.

Error (1) is introduced by using finite B bootstrapped samples of input distribution to quantify the input uncertainty. Errors (2) and (3) are introduced by using finite replications $\mathbf{n} \equiv \{n_1, n_2, \ldots, n_B\}$ to estimate the mean responses $\{\mu_1, \mu_2, \ldots, \mu_B\}$.

We quantify the impact of these errors by constructing a $(1-\alpha)100\%$ CI for the percentile Q estimation, denoted by $[C_L, C_U]$,

$$\Pr\left(Q \in [C_L, C_U] | \mathbf{X}_m\right) \geq 1 - \alpha. \tag{2}$$

This CI is conditional on the data $\mathbf{X}_m$ because the input uncertainty is quantified by the nonparametric bootstrap that takes the real-world data as the whole population. Notice that applying Equation (2), we have

$$\Pr(Q \in [C_L, C_U]) = \int \Pr(Q \in [C_L, C_U] | \mathbf{X}_m) d\mathbf{X}_m \geq 1 - \alpha.$$

The expected width of this CI can be used to quantify the percentile Q estimation error. Given a simulation budget C, we could solve the optimization problem

$$\begin{aligned} \min_{B, \mathbf{n}} \quad & \mathbf{E}(C_U - C_L | \mathbf{X}_m) \\ \text{S.t.} \quad & \textstyle\sum_{b=1}^{B} n_b = C \\ & \Pr(Q \in [C_L, C_U] | \mathbf{X}_m) \geq 1 - \alpha. \end{aligned} \tag{3}$$

to find an optimal allocation strategy.

Since this is a hard optimization problem, we develop an adaptive sequential approach that can automatically find B and $\mathbf{n}$ so that we could efficiently use the simulation budget to estimate the percentile Q, while controlling the simulation estimation

error. Specifically, B is selected to balance the finite sampling error with Errors (2) and (3). Then, given a fixed set $\{F^{(1)}, F^{(2)}, \ldots, F^{(B)}\}$ quantifying the input uncertainty, a ranking and selection based method is introduced to allocate the simulation resource, which can balance the selection error and the system response estimation error. Thus, our adaptive approach can simultaneously control all sources of errors and improve the quantile Q estimation.

3. AN ADAPTIVE SEQUENTIAL APPROACH

In this section, we develop an adaptive sequential approach that can efficiently use the simulation budget to estimate the percentile Q. Specifically, we first quantify the finite sampling error introduced by using finite B bootstrapped samples of input model to quantify the input uncertainty in Section 3.1. The impact of finite sampling error is called *the outer level uncertainty*. Then, given a set of bootstrapped samples, $\{F^{(1)}, F^{(2)}, \cdots, F^{(B)}\}$, quantifying the input uncertainty, we develop a ranking and selection based sequential procedure that can reduce the impact of the simulation estimation uncertainty, called *the inner level uncertainty*, by simultaneously controlling both selection and system response estimation errors in Section 3.2. This procedure gradually finds the important samples that contribute the most to the percentile Q estimation, and allocate more simulation resource there. Our approach returns an interval $[C_L, C_U]$ satisfying Equation (2) and accounting for both outer and inner level uncertainty. By solving an optimization problem, we can find the optimal B and parameters for the sequential procedure to minimize the expected width of $(C_U - C_L)$ in Section 3.3.

3.1. Quantifying Finite Sampling Error

Suppose that $\mu(\cdot)$ is known. When the order statistics $\mu_{[(1-\beta)B]}$ is used to estimate the percentile Q, the finite sampling error is introduced because we use finite B bootstrapped samples of input distribution $\{F^{(1)}, F^{(2)}, \cdots, F^{(B)}\}$ to quantify the input uncertainty. In this section, we construct a CI quantifying the finite sampling error.

Let $N(y) \equiv \sum_{b=1}^{B} \mathbf{1}(\mu_b \geq y)$ denote the number of bootstrapped samples with mean response greater or equal to a threshold y, where $\mathbf{1}(\cdot)$ represents the indicator function. Since $\sum_{b=1}^{B} \mathbf{1}(\mu_b \geq Q)|\mathbf{X}_m \sim \text{Binomial}(B, \beta)$, we can construct a $(1-\alpha_o)100\%$ two-sided CI for the percentile Q estimation

$$\Pr(Q \in [C_L^o, C_U^o]|\mathbf{X}_m) \geq 1 - \alpha_o$$

where α_o is the significant level assigned to control the outer level uncertainty of the percentile Q estimation. According to Baysal and Staum [2008], the lower and upper bounds of the interval are

$$C_L^o \equiv \inf\left\{y : \sum_{n=N(y)+1}^{B} \binom{B}{n} \beta^n (1-\beta)^{B-n} \geq \frac{\alpha_o}{2}\right\} \tag{4}$$

$$C_U^o \equiv \sup\left\{y : \sum_{n=0}^{N(y)} \binom{B}{n} \beta^n (1-\beta)^{B-n} \geq \frac{\alpha_o}{2}\right\}. \tag{5}$$

Theorem 3.1 shows that C_L^o and C_U^o in Equations (4) and (5) correspond to the k_1th and k_2th order statistics, denoted by $\mu_{[k_1]}$ and $\mu_{[k_2]}$, with k_1 defined as the smallest index such that $\mu_{[k_1+1]} = \min S_1$ and k_2 defined as the largest index such that $\mu_{[k_2]} =$

$\max S_2$, where

$$S_1 \equiv \left\{ \mu_{[b]} \text{ with } b = 1, 2, \ldots, B : \sum_{n=N(\mu_{[b]})+1}^{B} \binom{B}{n} \beta^n (1-\beta)^{B-n} \geq \frac{\alpha_o}{2} \right\}$$

$$S_2 \equiv \left\{ \mu_{[b]} \text{ with } b = 1, 2, \ldots, B : \sum_{n=0}^{N(\mu_{[b]})} \binom{B}{n} \beta^n (1-\beta)^{B-n} \geq \frac{\alpha_o}{2} \right\}.$$

The derivation of Theorem 3.1 is provided in the online appendix. If there is no simulation estimation error, by setting $\alpha_o = \alpha$, we get $[C_L, C_U] = [\mu_{[k_1]}, \mu_{[k_2]}]$.

THEOREM 3.1. *When we use $\mu_{[(1-\beta)B]}$ to estimate the percentile Q, the order statistics $\mu_{[k_1]}$ and $\mu_{[k_2]}$ define a $(1-\alpha_o)100\%$ CI quantifying the finite sampling error:*

$$\Pr(Q \in [\mu_{[k_1]}, \mu_{[k_2]}] | \mathbf{X}_m) \geq 1 - \alpha_o.$$

3.2. Quantifying Simulation Estimation Error

However, the system mean response $\mu(\cdot)$ is unknown. At a *given* set of bootstrapped samples, $\{F^{(1)}, F^{(2)}, \ldots, F^{(B)}\}$, used to quantify the input uncertainty, we estimate responses $\{\mu_1, \mu_2, \ldots, \mu_B\}$ with $\{\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_B\}$ obtained by using a finite simulation budget C. In this section, we develop a sequential procedure that can efficiently use the simulation resource to reduce the inner level uncertainty, and further deliver a CI quantifying the overall uncertainty of the percentile Q estimation.

Accounting for the finite sampling error described in Section 3.1, the B bootstrapped samples $\{F^{(1)}, F^{(2)}, \ldots, F^{(B)}\}$ are divided into three classes: $\mathcal{F}_L \equiv \{F^{(b)} : \mu_b < \mu_{[k_1]} \text{ for } b = 1, 2, \ldots, B\}$, $\mathcal{F}_U \equiv \{F^{(b)} : \mu_b > \mu_{[k_2]} \text{ for } b = 1, 2, \ldots, B\}$ and $\mathcal{F}_C \equiv \{F^{(b)} : \mu_{[k_1]} \leq \mu_b \leq \mu_{[k_2]} \text{ for } b = 1, 2, \ldots, B\}$. The samples in the set $\mathcal{F}_C$ are statistically indifferent with the order statistics $\mu_{[(1-\beta)B]}$. To reduce the impact of the inner level uncertainty, we want to efficiently identify the important samples included in the set $\mathcal{F}_C$, which contribute the most to the percentile Q estimation. Then, assign more simulation resource there to reduce the expected width of $(C_U - C_L)$.

We first show that if we can construct CIs for the order statistics $\mu_{[k_1]}$ and $\mu_{[k_2]}$, we also obtain a CI for the mean responses of all samples in the set $\mathcal{F}_C$. Specifically, suppose we can construct a $(1-\alpha_I/2)100\%$ one-sided CI, denoted by $[C_L, +\infty)$ and $(-\infty, C_U]$, for $\mu_{[k_1]}$ and $\mu_{[k_2]}$

$$\Pr(\mu_{[k_1]} \in [C_L, +\infty)) \geq 1 - \frac{\alpha_I}{2}$$
$$\Pr(\mu_{[k_2]} \in (-\infty, C_U]) \geq 1 - \frac{\alpha_I}{2}$$

where α_I is the significant level assigned to control the inner level uncertainty. By Theorem 3.2, we can show $\Pr(\mu_{[b]} \in [C_L, C_U]) \geq 1 - \alpha_I$ for all $b \in \mathcal{F}_C$. The proof of Theorem 3.2 is provided in the online appendix.

THEOREM 3.2. *If $\Pr(\mu_{[k_1]} \in [C_L, +\infty)) \geq 1 - \alpha_I/2$ and $\Pr(\mu_{[k_2]} \in (-\infty, C_U]) \geq 1 - \alpha_I/2$, then $\Pr(\mu_{[b]} \in [C_L, C_U]) \geq 1 - \alpha_I$ for all $b \in \mathcal{F}_C$.*

Then, by applying Theorems 3.1, 3.2 and the Bonferroni inequality, Theorem 3.3 shows that if α is decomposed into outer and inner significant levels, $\alpha = \alpha_o + \alpha_I$, then the interval $[C_L, C_U]$ accounting for both finite sampling and simulation estimation uncertainty satisfies Equation (2). The proof of Theorem 3.3 is provided in the online appendix.

THEOREM 3.3. *If $\alpha = \alpha_o + \alpha_I$ and the conditions in Theorems 3.1–3.2 hold, then* $\Pr(Q \in [C_L, C_U]|\mathbf{X}_m) \geq 1 - \alpha$.

Therefore, given B bootstrapped samples $\{F^{(1)}, F^{(2)}, \ldots, F^{(B)}\}$ quantifying the input uncertainty, to reduce the impact of simulation estimation uncertainty, we need to find the optimal budget allocation specified by $\mathbf{n} = (n_1, n_2, \ldots, n_B)$ to minimize the estimation error of order statistics $\mu_{[k_1]}$ and $\mu_{[k_2]}$ or the expected width of $(C_U - C_L)$. The optimal $\mathbf{n}$ depends on the unknown response mean $\mu(\cdot)$ and simulation estimation variance $\sigma^2_\epsilon(\cdot)$ at these bootstrapped samples. Further, since B could be hundreds and thousands, it is difficult to directly solve for the optimal $\mathbf{n}$.

Motivated by ranking and selection in Lesnevski et al. [2007; 2008], we develop a sequential procedure to simultaneously estimate order statistics $\mu_{[k_1]}$ and $\mu_{[k_2]}$. Specifically, let the surviving set, denoted by I^q, contain all the samples that are statistically indifferent to be the order statistics $\mu_{[k_q]}$ for $q = 1, 2$. Without any prior information on the system mean responses, we start with classifying all samples $\{F^{(1)}, F^{(2)}, \ldots, F^{(B)}\}$ to the surviving set I^q. By running simulations at samples in I^q, we could gradually screen out samples that are statistically smaller or larger than $\mu_{[k_q]}$, and then allocate more simulation budget to surviving samples in I^q that contribute the most to estimating $\mu_{[k_q]}$.

The screening process could introduce the selection bias, especially when the number of bootstrapped samples is large and the simulation estimation uncertainty is high. For example, if we want to find the system with the maximum mean response $\mu_{[B]}$, the selected sample with the maximum sample mean $\bar{Y}_{(B)}$ is biased high, $\mathrm{E}[\bar{Y}_{(B)}] \geq \mu_{[B]}$; see Lan et al. [2010], Boesel et al. [2003], Nelson and Goldsman [2001]. The "restart" is used to reduce the bias, which divides the sequential procedure into screening and estimation phases.

3.2.1. Phase I: Screening with Common Random Numbers. The screening phase is to eliminate bootstrapped samples of input model that are statistically impossible to be the order statistics $\mu_{[k_q]}$ for $q = 1, 2$. Since $\{\mu_1, \mu_2, \ldots, \mu_B\}$ represent the responses of *the same system* derived by different input model estimates, the samples with similar input models also have similar mean responses. Thus, CRN is used to efficiently screen out samples having mean responses close to order statistics $\mu_{[k_q]}$.

Suppose that there are n replications at each sample of $\{F^{(1)}, F^{(2)}, \ldots, F^{(B)}\}$. We introduce *a two-sided screening* to screen out samples with mean response statistically smaller or larger than $\mu_{[k_q]}$ for $q = 1, 2$. Given a screening significant level, denoted by α_S, Theorem 3.4 shows that the surviving set I^q defined in Equation (6) includes the k_qth smallest sample with probability greater or equal to α_S, $\Pr([k_q] \in I^q) \geq 1 - \alpha_S$, for $q = 1, 2$. The detailed proof of Theorem 3.4 is provided in the online appendix.

THEOREM 3.4. *Define the surviving set*

$$I^q = \left\{ i \in \{1, 2, \ldots, B\} : \sum_{j \in \{1,2,\ldots,B\}: j \neq i} \mathbf{1}\left(\bar{Y}_j \leq \bar{Y}_i + W_{ij}\right) \geq k_q - 1 \right.$$
$$\left. \text{and} \sum_{j \in \{1,2,\ldots,B\}: j \neq i} \mathbf{1}\left(\bar{Y}_j \geq \bar{Y}_i - W_{ij}\right) \geq B - k_q \right\} \quad (6)$$

where $W_{ij} = t_{n-1,1-\frac{\alpha_S}{B-1}} \cdot \frac{S_{ij}}{\sqrt{n}}$, $S^2_{ij} = \frac{1}{n-1}\sum_{h=1}^{n}(Y_{ih} - Y_{jh} - (\bar{Y}_i - \bar{Y}_j))^2$ *and* n *denotes the number of replications assigned to each sample. Then,* $\Pr([k_q] \in I^q) \geq 1 - \alpha_S$ *for* $q = 1, 2$.

Therefore, to efficiently find the important bootstrapped samples that contribute the most to the estimation of order statistics $\mu_{[k_1]}$ and $\mu_{[k_2]}$, built on the screening rule in Equation (6), we develop *a sequential screening procedure* as follows. It is specified by parameters: (1) the initial number of replications allocated to each bootstrapped sample, denoted by n_0, (2) the growth factor characterizing the replication increase rate for samples in the surviving sets, denoted by R, and (3) the number of screening iterations, denoted by M. Given a fixed simulation budget, large n_0 could lead to global exploration. Small R and large M could lead to local exploitation.

In the ℓth iteration, let I_ℓ^q denote the surviving set, and let $I_\ell^{L_q}$ and $I_\ell^{U_q}$ denote sets with mean response statistically smaller and larger than order statistics $\mu_{[k_q]}$ for $q = 1, 2$. Let $N_1^q(\ell)$ denote the number of replications accumulated in Phase I until the ℓth iteration at each sample in the surviving set I_ℓ^q. Without any prior information on the system mean response, we start with putting all bootstrapped samples of input distribution to the surviving sets I_0^q. In Step (2), we equally divide the screening significant level α_S to M iterations and set $\alpha_S' = \alpha_S/M$. Then, assign n_0 replications to each sample $N_1^q(1) = n_0$, and run simulations. Based on the simulation results, screen out samples that are statistically impossible to be the order statistics $\mu_{[k_q]}$ in Step (3.a), where $|I_{\ell-1}^{L_q}|$, $|I_{\ell-1}^{U_q}|$ and $|I_{\ell-1}^q|$ denote the numbers of samples in the sets $I_{\ell-1}^{L_q}$, $I_{\ell-1}^{U_q}$ and $I_{\ell-1}^q$. After that, check the stopping criteria in Step (3.b). If the stopping criteria hold, stop screening. Otherwise, allocate additional $n_0 R^{\ell-1}(R-1)$ replications to each sample in the surviving sets, run simulations and do further screening. In iteration $\ell+1$, each sample in the surviving sets has the accumulated replications $N_1^q(\ell+1) = n_0 R^\ell$.

(1) Let $I_0^q \leftarrow \{F^{(1)}, F^{(2)}, \ldots, F^{(B)}\}$ and $I_0^{L_q} = I_0^{U_q} = \emptyset$ for $q = 1, 2$.
(2) Let $\alpha_S' = \alpha_S/M$ be the screening significant level for each iteration in Phase I. Assign n_0 replications to each $F^{(b)}$ with $b = 1, 2, \ldots, B$ and run simulations.
(3) For $\ell = 1$ to M
   (a) Do screening. For $q = 1, 2$, set

$$I_\ell^{L_q} = I_{\ell-1}^{L_q} \bigcup A_\ell^q,\ I_\ell^{U_q} = I_{\ell-1}^{U_q} \bigcup B_\ell^q,\ I_\ell^q = \bar{A_\ell^q} \bigcap \bar{B_\ell^q}$$

   where $A_\ell^q \equiv \left\{ i \in I_{\ell-1}^q : \sum_{j \in I_{\ell-1}^q : j \neq i} \mathbf{1}(\bar{Y}_j \leq \bar{Y}_i + W_{\ell,ij}^q) < k_q - 1 - |I_{\ell-1}^{L_q}| \right\}$, $B_\ell^q \equiv \left\{ i \in I_{\ell-1}^q : \sum_{j \in I_{\ell-1}^q : j \neq i} \mathbf{1}(\bar{Y}_j \geq \bar{Y}_i - W_{\ell,ij}^q) < B - k_q - |I_{\ell-1}^{U_q}| \right\}$, $W_{\ell,ij}^q = t_{N_1^q(\ell)-1, 1-\frac{\alpha_S'}{|I_{\ell-1}^q|-1}} \cdot \frac{S_{ij}^q}{\sqrt{N_1^q(\ell)}}$ and $S_{ij}^q = \left[\frac{1}{N_1^q(\ell)-1} \sum_{h=1}^{N_1^q(\ell)} \left(Y_{ih} - Y_{jh} - (\bar{Y}_i - \bar{Y}_j) \right)^2 \right]^{1/2}$.
   (b) Check stopping condition. If $|I_\ell^q| = 1$, then stops updating I_ℓ^q for $q = 1, 2$. If either both surviving sets I_ℓ^1 and I_ℓ^2 have a single sample left or $\ell = M$, the procedure moves to Phase II. Otherwise, assign additional $n_0 R^{\ell-1}(R-1)$ replications to each sample in the surviving set I_ℓ^q, and run simulations.
   Next ℓ.

By applying Theorem 3.4 and Bonferroni inequality, we can show that the surviving set obtained from this screening procedure, denoted by I_M^q, includes the k_qth smallest sample with probability greater or equal to $1 - \alpha_S$: $\Pr([k_q] \in I_M^q) \geq 1 - \alpha_S$ for $q = 1, 2$. It can be obtained by following the similar proof with Lemma 1 in Nelson et al. [2001].

3.2.2. Phase II: Estimation with Variance Reduction Technique. The sequential screening described in Section 3.2.1 could introduce the selection bias. For each sample in the surviving sets, $b \in \{I_M^1 \cup I_M^2\}$, the sample means are biased, $\mathrm{E}(\bar{Y}_b) \neq \mu_b$. Thus, the restart

is used to reduce the bias. Only sample variances S_b^2 are used to guide the remaining simulation budget allocation.

The number of replications allocated to surviving bootstrapped samples in Phase II, denoted by $N_2(b)$, is proportional to the sample variance

$$N_2(b) = (C - T)\frac{S_b^2}{\sum_{i\in\{I_M^1 \cup I_M^2\}} S_i^2} \tag{7}$$

where T is the total budget used in Phase I, which is a function of (B, n_0, R, M) and also depends on unknown $\mu(\cdot)$ and $\sigma_\epsilon(\cdot)$ at $\{F^{(1)}, F^{(2)}, \ldots, F^{(B)}\}$. Then, run simulations at each sample in the surviving sets. Based on the simulation results, we construct a $(1-\alpha)100\%$ CI for the percentile Q estimation with lower and upper bounds

$$C_L = \min_{b\in I_M^1}\left(\bar{Y}_b - t_{N_2(b)-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_b}{\sqrt{N_2(b)}}\right) \tag{8}$$

$$C_U = \max_{b\in I_M^2}\left(\bar{Y}_b + t_{N_2(b)-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_b}{\sqrt{N_2(b)}}\right) \tag{9}$$

where α_E is the significant level for the estimation phase. Since the CRN used in the screening phase could efficiently screen out the samples with mean responses slightly different with the order statistics $\mu_{[k_q]}$, we assume that the mean responses at samples in I_M^q are close to $\mu_{[k_q]}$ for $q = 1, 2$. Thus, the budget allocation in Equation (7) could minimize the width of $(C_U - C_L)$.

When the simulation budget is tight and the mean responses at bootstrapped samples are also close to each other, there could exist many surviving samples in Phase II. To efficiently estimate the mean response μ_b for $b \in \{I_M^1 \cup I_M^2\}$ and reduce the estimation variance, a variance reduction technique, *the antithetic variate simulation algorithm* [Hammersley and Morton 1956], is employed in the simulations. Specifically, let $N_2'(b) = N_2(b)/2$ be an integer. At each $F^{(b)}$, let Y_{bi} and $\widetilde{Y}_{bi}$ denote the simulation outputs generated by the random streams following uniform distribution, denoted by $\mathbf{U}_{bi}$ and $1 - \mathbf{U}_{bi}$, for $i = 1, 2, \ldots, N_2'(b)$. Then, Equations (8) and (9) become

$$C_L = \min_{b\in I_M^1}\left(\bar{Y}_b - t_{N_2'(b)-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_{b,a}}{\sqrt{N_2'(b)}}\right) \tag{10}$$

$$C_U = \max_{b\in I_M^2}\left(\bar{Y}_b + t_{N_2'(b)-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_{b,a}}{\sqrt{N_2'(b)}}\right) \tag{11}$$

where $S_{b,a}^2 = \sum_{i=1}^{N_2'(b)}\left(\frac{Y_{bi}+\widetilde{Y}_{bi}}{2} - \bar{Y}_b\right)^2 / [N_2'(b) - 1]$. Our empirical study indicates that it could obviously improve the percentile Q estimation.

We can show that the interval $[C_L, C_U]$ obtained from Equations (10) and (11) satisfies Equation (2). We decompose the inner uncertainty significant level α_I into the parts for screening and estimation. Since the significant level used to screen for each $\mu_{[k_q]}$ with $q = 1, 2$ is α_S, the total screening significant level is $2\alpha_S$. Theorem 3.5 shows that if $\alpha_I = 2\alpha_S + \alpha_E$, we can get $\Pr(\mu_{[b]} \in [C_L, C_U]|\mathbf{X}_m) \geq 1 - \alpha_I$ for any $b \in \mathcal{F}_C$. Then, $\Pr(Q \in [C_L, C_U]|\mathbf{X}_m) \geq 1 - \alpha$ follows by applying Theorems 3.2 and 3.3. The detailed proof of Theorem 3.5 is provided in the online appendix.

THEOREM 3.5. *If $\alpha_I = 2\alpha_S + \alpha_E$, then the interval $[C_L, C_U]$ defined by Equations (10) and (11) satisfies $\Pr(\mu_{[b]} \in [C_L, C_U]|\mathbf{X}_m) \geq 1 - \alpha_I$ for any $b \in \mathcal{F}_C$.*

3.3. An Adaptive Sequential Procedure

Given the simulation budget C, the sequential procedure described in Sections 3.1 and 3.2 delivers a CI for the percentile Q estimation. This procedure is parameterized by (B, n_0, R, M). The optimal parameters, denoted by $(B^\star, n_0^\star, R^\star, M^\star)$, minimizing $\mathrm{E}(C_U - C_L|\mathbf{X}_m)$ depend on the input uncertainty and also unknown $\mu(\cdot)$, $\sigma_\epsilon^2(\cdot)$ at bootstrapped samples quantifying the input uncertainty. In this section, we introduce an adaptive sequential procedure that could find the optimal parameters. Specifically, we use part of simulation budget to run the initial simulations that could provide the information on the input uncertainty and $\mu(\cdot), \sigma_\epsilon^2(\cdot)$ at bootstrapped samples of input model. Then, by following the procedure in Sections 3.1 and 3.2, we can estimate the expected interval width $\mathrm{E}(C_U - C_L|\mathbf{X}_m)$ at any feasible (B, n_0, R, M), which allows us to search for the optimal parameters $(B^\star, n_0^\star, R^\star, M^\star)$.

Our adaptive sequential procedure includes main steps as follows. In Step (1), we specify the significant levels $\alpha, \alpha_o, \alpha_S, \alpha_E$ and the ranges or the design space of (B, n_0, R, M). The sensitivity study in Section 4.1.2 indicates that the performance of our approach is not sensitive to the choice of significant levels. In Step (2), for the initial simulations, we generate B_0 bootstrapped samples, run n_{00} replications at each sample, and calculate the sample means and variances of simulation outputs, denoted by $\bar{Y}_{0i}$ and S_{0i}^2 with $i = 1, 2, \ldots, B_0$. Suppose that these B_0 samples provide the representative behaviors of the simulation outputs at bootstrapped samples so that we can estimate $\mu(\cdot)$ and $\sigma_\epsilon^2(\cdot)$ at any new generated bootstrapped sample. In Step (3), to estimate the performance of our sequential approach with any feasible parameters (B, n_0, R, M), suppose that there are B *virtual* bootstrapped samples of input model. We assign them to B_0 groups defined by sampling from the initial simulations with equal probabilities. Then, without running any additional simulations, we estimate the response mean and simulation estimation uncertainty at each virtual bootstrapped sample, and further construct the interval $[C_L, C_U]$ defined by Equations (10) and (11) by following the sequential procedure in Sections 3.1 and 3.2. Through repeating Step (3), we can estimate the mean and variance of the interval width $(C_U - C_L)$ at any (B, n_0, R, M), which could be used to search for the optimal parameters $(B^\star, n_0^\star, R^\star, M^\star)$ in Step (5). After that, we generate additional $B^\star - B_0$ bootstrapped samples of input models, and then by following the sequential procedure in Section 3.2 with parameters $(n_0^\star, R^\star, M^\star)$, we can construct the interval $[C_L, C_U]$ accounting for both finite sampling and system response estimation uncertainty, which could efficiently use the simulation budget to improve the percentile Q estimation.

(1) Specify the significant levels $\alpha, \alpha_o, \alpha_S$ and α_E. Specify the ranges for (B, n_0, R, M).
(2) Generate B_0 bootstrapped empirical distributions, assign n_{00} replications to each distribution and run simulations. Then, record response means and variances, $\bar{Y}_{0i}$, and S_{0i}^2, with $i = 1, 2, \ldots, B_0$.
(3) For any feasible setting (B, n_0, M, R), assign B virtual samples into B_0 groups by using a multinomial sampling with probability parameters $p_i = 1/B_0$ for $i = 1, 2, \ldots, B_0$. The group index of the jth sample associated to is $c_j \overset{i.i.d.}{\sim}$ Multinomial$(p_1, p_2, \ldots, p_{B_0})$, for $j = 1, 2, \ldots, B$. Draw the response mean and variance for the jth sample, $\bar{Y}_j \sim \mathrm{N}\left(\bar{Y}_{0c_j}, S_{0c_j}^2/n_0\right)$ and $S_j^2 \sim S_{0c_j}^2 \cdot \chi_{n_0-1}^2/(n_0-1)$. Following the screening and estimation procedure in Section 3.2, construct $[C_L, C_U]$ defined by Equations (10) and (11), and record the interval width.
(4) Repeat Step (3) to estimate the mean and variance of $(C_U - C_L)$.
(5) Find the optimal parameters $(B^\star, n_0^\star, R^\star, M^\star)$ to minimize $\mathrm{E}(C_U - C_L|\mathbf{X}_m)$.

(6) Use the budget allocation obtained by the sequential procedure in Sections 3.2.1 and 3.2.2 with $(B^\star, n_0^\star, R^\star, M^\star)$ to estimate the percentile Q by $\bar{Y}_{((1-\beta)B)}$ and return the interval $[C_L, C_U]$ quantifying the overall estimation uncertainty.
   (a) Generate $B^\star - B_0$ new bootstrapped samples of input model. Run $n_0^\star$ replications at each new generated sample and $n_0^\star - n_{00}$ additional replications at each existing sample.
   (b) By following the sequential screening in Section 3.2.1 with parameters $(n_0^\star, R^\star, M^\star)$ and then the estimation step in Section 3.2.2, construct the interval $[C_L, C_U]$ defined by Equations (10) and (11).

Here, we provide some guidance on choosing n_{00} and B_0 for the initial simulations. For the situations with a tight simulation budget, we set them to be the smallest values of n_0 and B. The choice of n_0 should make sure that the sample variance of simulation outputs provides a reasonable estimation of the simulation estimation uncertainty. Based on the literature on ranking and selection, the minimal number of replications is recommended to be 10 [Kim and Nelson 2007; Tsai et al. 2009]. Thus, we use $n_{00} = 10$ in the empirical study. Since $\sum_{b=1}^{B} \mathbf{1}(\mu_b \geq Q)|\mathbf{X}_m \sim \text{Binomial}(B, \beta)$, in Section 3.1, we construct the interval $[\mu_{[k_1]}, \mu_{[k_2]}]$ to quantify the finite sampling error. If $[k_2] = B$, there is no bootstrapped sample μ_b statistically larger than the percentile Q. To have the representative behavior of the simulation outputs at bootstrapped samples of input model, the choice of B should guarantee that at least one bootstrapped samples fall in the $\alpha_o/2$ right tail part,

$$\sum_{b=0}^{1} \binom{B}{b} \beta^b (1-\beta)^{B-b} < \frac{\alpha_o}{2}. \tag{12}$$

Thus, B_0 should be the smallest B satisfying Equation (12).

For the discrete optimization via simulation (DOvS) problem in Step (5), there are a variety of optimization algorithms that can be incorporated into our procedure to find the optimal parameters $(B^\star, n_0^\star, R^\star, M^\star)$ minimizing $L(B, n_0, R, M) \equiv \mathbf{E}(C_U - C_L|\mathbf{X}_m)$, e.g., approaches proposed in Brooks [1958], Hong and Nelson [2006], Xu et al. [2010], Xu et al. [2013], Sun et al. [2014], and Hong et al. [2014]. To illustrate that the performance of our approach is robust to the choice of the optimization approaches, two of which are employed in our empirical study in Section 4. The first algorithm is the pure random search algorithm (PRS) [Brooks 1958]. It is simple and makes no assumption on the surface $L(B, n_0, R, M)$. Specifically, we randomly generate N_s samples of (B, n_0, R, M) covering the design space. At each sample, the expected CI width is estimated by following Steps (3)-(4) and the best one is recorded as the optimal solution. However, PRS may require large N_s to find a good solution except when $L(B, n_0, R, M)$ is relatively flat around $(B^\star, n_0^\star, R^\star, M^\star)$ or the simulation budget C is tight. The second approach is Gaussian-Process based search (GPS) [Sun et al. 2014]. Unlike PRS, GPS requires some smoothness assumption on the unknown response surface $L(B, n_0, R, M)$. We can estimate the expected CI width at a few well chosen design points by following the sequential procedure in Sections 3.2.1 and 3.2.2, and then build the fitted surface to predict the CI width at other untried points in the design space. This information could guide further search to the promising subregion. Thus, GPS includes main steps as follows [Sun et al. 2014].

(1) Generate the initial design points evenly covering the design space for (B, n_0, R, M). To have all prediction points located close to design points, a space-filling design, the orthogonal max-min Latin Hypercube Design (LHD), is used to select the design points [Liu and Staum 2010].
(2) Fit or update the surface $L(B, n_0, R, M)$.

(3) Calculate the posterior probability of being better than the current optimal solution, use it as the sampling distribution to draw new samples of (B, n_0, R, M), and choose the point with the smallest expected CI width as the current optimal.
(4) Repeat Steps (2-3) and terminate the process if the number of iterations reaches the predetermined threshold or the maximum expected improvement is sufficiently small.

Since GPS exploits the relationship between the expected CI width at different settings (B, n_0, R, M), it could make an effective use of the computational budget. However, there exist fitting issues when the response surface $L(B, n_0, R, M)$ is relatively flat near the optimal; see Sun et al. [2014]. It is worth noting that other efficient algorithms can also apply in our procedure, such as the COMPASS algorithm. Interested readers are referred to Hong and Nelson [2006], Xu et al. [2010] and Xu et al. [2013] for further details.

The overhead cost introduced by our approach mainly comes from the sequential screening and the optimization search. It is polynomial in the simulation budget C; see the derivation in the Appendix. The empirical study in Section 4 demonstrates that the average overhead cost takes about a few seconds for each simulation run. Since we consider the situations where each simulation run could be computationally expensive and the simulation budget is tight, the overhead cost is not dominating.

4. EMPIRICAL STUDY

In this section, we use an $M/M/1$ queue and a stochastic activity network to study the performance of our sequential approach. To demonstrate our approach is compatible with various optimization techniques, we use GPS for the $M/M/1$ queue in Section 4.1 and use PRS for the stochastic activity network in Section 4.2 to search for the optimal parameters $(B^\star, n_0^\star, R^\star, M^\star)$ for our sequential procedure.

4.1. An $M/M/1$ Queue

We first use a tractable $M/M/1$ queue to illustrate the performance of our approach. The customer arrival rate is λ^c and the service rate is θ^c. We are interested in the average number of customers in the system.

To evaluate our approach, we pretend that the distribution of service time is unknown and estimated from m i.i.d. service time data $\mathbf{X}_m = (X_1, X_2, \ldots, X_m)$, which are generated from the underlying true service distribution, $\exp(\theta^c)$. We refer the observations as "real-world data." Notice that the system under the bootstrapped empirical distribution for the service is in fact an $M/G/1$ queue. By the Pollaczek-Khinchine formula, we can calculate the expected number of customers in the system at any bootstrapped sample of service distribution $F^{(b)} \sim \tilde{F}(\cdot|\mathbf{X}_m)$,

$$\mu\left(F^{(b)}\right) = \rho + \frac{\rho^2 + (\lambda^c)^2 \mathrm{Var}(X)}{2(1-\rho)} \tag{13}$$

where X denotes the service time with $X \sim F^{(b)}$ and $\rho = \lambda^c \mathrm{E}(X)$ is the traffic intensity. Hence, the impact of the input uncertainty on the system performance estimates can be quantified by the $(1-\beta)100\%$ percentile Q of the induced distribution of $\mu(F^{(b)})$ satisfying Equation (1).

Given a fixed simulation budget C, we study the finite-sample performance of our approach to estimate the percentile Q. In our experiments, at any $F^{(b)}$, we estimate the mean response $\mu(F^{(b)})$ by running simulations. The simulations start with an empty system. We set the warm-up length to be 500. To consider the situation where the simulation budget is tight and the simulation estimation uncertainty could be significant,

Table I. The maximum absolute relative difference for percentile Q estimation when $m = 100$ (in the unit $\%$).

Number of bootstrapped samples	10^3	5×10^3	10^4	5×10^4	10^5	5×10^5
$\lambda^c = 0.5, 1-\beta = 90\%$	0.67	0.44	0.23	0.15	0.09	0.04
$\lambda^c = 0.5, 1-\beta = 95\%$	0.75	0.67	0.50	0.17	0.09	0.06
$\lambda^c = 0.7, 1-\beta = 90\%$	1.36	0.81	0.46	0.33	0.15	0.10
$\lambda^c = 0.7, 1-\beta = 95\%$	2.28	2.12	1.27	0.43	0.19	0.14

we set a short run length after the warm-up to be 50 for each replication. Here, the unit of warm-up and run length is defined based on the number of arrivals.

To evaluate the robustness of our approach, we examine the effects of the traffic intensity, the amount of real-world data, the percentile level for Q and the simulation budget. We fix the service rate $\theta^c = 1$ in the experiments. We consider different levels of arrival rate $\lambda^c = 0.5, 0.7$, the amount of real-world data $m = 100, 500$, the significant level for input uncertainty $1-\beta = 90\%, 95\%$, and the total simulation budget in terms of replications $C = 5000, 10000, 50000$.

We use the Monte Carlo approach to estimate the true percentile Q. To find the number of bootstrapped samples required to get a precise percentile estimation, denoted by B_1, we did a side experiment by running 10 macro-replications for arrival rate $\lambda^c = 0.5, 0.7$. Since $m = 500$ has less input uncertainty, we focus on cases with $m = 100$. In each macro-replication, we draw $m = 100$ independent real-world observations from underlying true distribution F^c, generate B_1 bootstrapped samples of service distribution, calculate the mean response at each bootstrapped sample by using Equation (13), and then estimate the percentile Q by using the order statistics $\widehat{Q}_{B_1} = \mu_{[(1-\beta)B_1]}$. We consider different number of bootstrapped samples and record the relative difference compared to the benchmark with 10^6 bootstrapped samples, error $= |\widehat{Q}_{B_1} - Q|/Q$, where Q denotes the percentile estimated by using 10^6 bootstrapped samples. Suppose 10^6 is large enough and the finite sample estimation error is negligible. The maximum relative error for different choices of B_1 obtained from 10 macro-replications is recorded in Table I with the unit to be percentage ($\%$). For simplicity, we suppress the percentage sign in all tables presented in the paper. We observe that 10^5 achieves accuracy with the maximum relative error less than 0.2%. Balancing precision and computational cost, we use $B_1 = 10^5$ bootstrapped samples to estimate the true percentile Q in our experiments.

It is worth mentioning that we occasionally encounter unstable systems with $\rho \geq 1$. To estimate the possibility getting the unstable bootstrapped system, we did a side experiment. We run 100 macro-replications for all combinations of $\lambda^c = 0.5, 0.7$ and $m = 100, 500$. In each macro-replication, we first draw m independent real-world observations from the underlying true distribution, generate 10^5 bootstrapped samples of input distribution and calculate the percentage of unstable bootstrapped samples. Based on results from 100 macro-replications, we can have mean and standard deviation (SD) of the percentage. When $m = 100$ and $\lambda^c = 0.7$, we have the average 0.24% unstable bootstrapped samples with SD equal to 0.8%, which is much smaller than β. The unstable issue is negligible for other cases. Given finite warm-up and run-length, the simulation outputs at unstable bootstrapped samples of input model are finite and they tend to larger than those at stable bootstrapped samples. Thus, we keep all unstable bootstrapped samples in the empirical study.

4.1.1. Comparing Our Approach with Direct Bootstrapping. Given a fixed simulation budget, we compare the performance of our approach and the direct bootstrapping. The absolute relative error of percentile Q estimated by using our approach and direct bootstrapping is shown in Tables II-III when $1-\beta = 90\%, 95\%$. For our approach, we

Table II. The relative error of the percentile Q estimates when $1-\beta = 90\%$ (in the unit %).

$\lambda^c = 0.5$	our approach		direct bootstrap	
	mean	SE	mean	SE
$m = 100, C = 5000$	6.6	0.35	18.8	1.85
$m = 100, C = 10000$	6.2	0.3	15.9	1.74
$m = 100, C = 50000$	5.3	0.31	6.6	0.49
$m = 500, C = 5000$	9.3	0.21	15.8	1.38
$m = 500, C = 10000$	7.7	0.17	12.3	1.07
$m = 500, C = 50000$	5.5	0.36	6.0	0.47
$\lambda^c = 0.7$	our approach		direct bootstrap	
	mean	SE	mean	SE
$m = 100, C = 5000$	10.9	0.84	32.6	3.18
$m = 100, C = 10000$	9.8	0.66	18.9	1.55
$m = 100, C = 50000$	9.1	0.53	10.5	0.82
$m = 500, C = 5000$	14.1	0.36	26.3	2.04
$m = 500, C = 10000$	10.7	0.31	21.9	1.94
$m = 500, C = 50000$	7.8	0.37	9.8	0.75

set $\alpha = 0.05$ and split α into $\alpha_o = 0.02$, $\alpha_S = \alpha_E = 0.01$. Following the study in Nelson et al. [2001] and Boesel et al. [2003], we use the equal significant level for the screening and estimation. The study in Section 4.1.2 further indicates that the performance of our approach is not sensitive to the value of α and the decomposition to the inner and outer uncertainty. For the direct bootstrapping, we set $B = 1000$ and equally allocate the total budget C to bootstrapped samples of input distribution.

In our approach, the parameters (B, n_0, M, R) are optimized based on the information obtained from the initial simulations. The setting for initial simulations is: $B_0 = 200$ and $n_{00} = 10$. When we search for the optimal (B, n_0, M, R), we set that n_0 can be chosen from $\{10, 20, 30, 40, 50\}$. Larger n_0 is infeasible for a tight budget. M can be selected from $\{1, 2, \ldots, 10\}$. For a tight budget, it is infeasible to do more than 10 screening iterations. Even for sufficient budget cases, if M is too large, fewer systems can be screened out at each stage since the screening significant level for each stage $\alpha'_S = \alpha_S/M$ decreases as M becomes larger. And R is chosen from $\{1.1, 1.2, \ldots, 2\}$. Too large growth factor can lead to the screening procedure that may consume more than necessary in each stage. For B, the available choices are from $\{200, 225, \ldots, C/10\}$ since $B \geq B_0$. And any feasible choice B cannot exceed $B = C/10$ because 10 is the smallest n_0 value.

GPS is used to search for the parameters of our sequential procedure. By following the "$10d$" rule in Jones et al. [1998], we use LHD to generate 40 initial design points to evenly cover the design space. We then follow the steps described in Section 3.3 to update the meta-model. The GPS search terminates when the number of feasible (B, n_0, R, M) visited reaches 100, or the maximum expected improvement is less than 1%.

The results in Tables II-III are based on 100 macro-replications. In each macro-replication, we first generate m real-world data from the underlying true service distribution, estimate the true percentile Q by using B_1 bootstrapped samples and Equation (13). Then, run simulations and estimate the impact of input uncertainty by using our approach and the direct bootstrapping. We record the relative error for percentile Q estimation: error $= |\bar{Y}_{((1-\beta)B)} - Q|/Q$. Based on the results from 100 macro-replications, we can estimate the mean and standard error (SE) of relative error.

The conclusions obtained from Tables II-III are similar. Given a fixed amount of real-world data, as the simulation budget increases, the mean and SE of percentile Q estimation error decrease by using either our approach or direct bootstrapping. Compared

Table III. The relative error of the percentile Q estimates when $1-\beta = 95\%$ (in the unit %).

$\lambda^c = 0.5$	our approach		direct bootstrap	
	mean	SE	mean	SE
$m = 100, C = 5000$	6.8	0.37	19.0	1.89
$m = 100, C = 10000$	6.3	0.36	21.5	1.31
$m = 100, C = 50000$	4.2	0.31	7.5	0.62
$m = 500, C = 5000$	11.1	0.31	20.3	1.84
$m = 500, C = 10000$	7.4	0.2	14.7	1.37
$m = 500, C = 50000$	4.6	0.15	6.6	0.52
$\lambda^c = 0.7$	our approach		direct bootstrap	
	mean	SE	mean	SE
$m = 100, C = 5000$	12.6	1.06	27.0	2.03
$m = 100, C = 10000$	9.5	0.96	23.4	1.89
$m = 100, C = 50000$	7.9	0.89	11.9	1.19
$m = 500, C = 5000$	15.5	0.39	28.4	2.67
$m = 500, C = 10000$	11.2	0.37	18.4	1.43
$m = 500, C = 50000$	6.6	0.38	10.2	0.84

to the direct bootstrapping, our approach has better and more robust performance. This advantage is more obvious when the simulation budget is tight. As the budget increases, the performance of both approaches becomes closer. For our approach, when the simulation budget is tight, e.g., $C = 5000, 10000$, the percentile estimation is better when $m = 100$ than when $m = 500$ because it is relatively easy to screen out samples when the input uncertainty is large and the system response at bootstrapped samples spreads out.

4.1.2. Sensitivity Analysis. In this section, we first study the effect of α on the performance of our approach. Larger $1-\alpha$ provides higher statistic guarantee that the true percentile Q is covered by the interval $[C_L, C_U]$. Smaller $1-\alpha$ could lead to screening out samples of input model more easily so that surviving ones can receive more simulation resource. Here, we study the performance of our approach when $1-\alpha = 90\%, 95\%, 99\%$. The decomposition rule remains fixed with $\alpha_o = 2\alpha/5, \alpha_S = \alpha_E = \alpha/5$. We set $1-\beta = 95\%$. The results shown in Table IV are based on 100 macro-replications. They indicate that the performance of our approach is robust to the value of $1-\alpha$. When $C = 5000, 10000$, all three settings with $1-\alpha = 90\%, 95\%, 99\%$ have similar performance. When $C = 50000$, we observe that the setting with $1-\alpha = 95\%$ performs slightly better than other two settings.

Then, we study the effect of significant level decomposition to the inner and outer uncertainty, $\alpha = \alpha_o + \alpha_I$, on the performance of our approach. We consider two decompositions: Case (1) has $\alpha_o = 2\alpha/5$ and $\alpha_I = 3\alpha/5$; Case (2) has $\alpha_o = 3\alpha/5$ and $\alpha_I = 2\alpha/5$. Following the study in Nelson et al. [2001] and Boesel et al. [2003], we set $\alpha_S = \alpha_E$. Here, we set $1-\alpha = 95\%$ and $1-\beta = 95\%$. Table V records the performance of our approach under these two cases. When $C = 5000, 10000$, the performances for both cases are not distinguishable. When $C = 50000$, Case (2) gives slightly worse performance. In sum, the performance of our approach is robust to the significant level decomposition.

4.2. A Stochastic Activity Network

In this section, we use a stochastic activity network in Fig.1 to study the performance of our approach. Suppose that the time required to complete arc (task) j is denoted by X_j with $X_j \sim \exp(\theta_j^c)$ for $j = 1, 2, \ldots, 5$ and parameters $\boldsymbol{\theta}^c = (10, 5, 12, 11, 5)$. The time to finish the project, denoted by Y, is defined as the longest path of the network,

Table IV. The performance of our approach when $1-\alpha = 90\%, 95\%, 99\%$ (in the unit $\%$).

$\lambda^c = 0.5$	$1-\alpha = 90\%$		$1-\alpha = 95\%$		$1-\alpha = 99\%$	
	mean	SE	mean	SE	mean	SE
$m = 100, C = 5000$	7.1	0.38	6.8	0.37	6.5	0.36
$m = 100, C = 10000$	6.2	0.38	6.3	0.36	6.3	0.43
$m = 100, C = 50000$	6.0	0.46	4.2	0.31	5.5	0.45
$m = 500, C = 5000$	11.1	0.26	11.1	0.31	11.6	0.27
$m = 500, C = 10000$	8.7	0.23	7.4	0.2	7.6	0.21
$m = 500, C = 50000$	6.4	0.49	4.6	0.15	5.4	0.28
$\lambda^c = 0.7$	$1-\alpha = 90\%$		$1-\alpha = 95\%$		$1-\alpha = 99\%$	
	mean	SE	mean	SE	mean	SE
$m = 100, C = 5000$	12.7	1.23	12.6	1.06	10.3	0.85
$m = 100, C = 10000$	11.6	0.99	9.5	0.96	8.7	1.04
$m = 100, C = 50000$	8.3	0.72	7.9	0.89	10.2	0.9
$m = 500, C = 5000$	15.2	0.4	15.5	0.39	15.9	0.45
$m = 500, C = 10000$	11.5	0.38	11.2	0.37	10.6	0.35
$m = 500, C = 50000$	7.8	0.39	6.6	0.38	6.6	0.31

Table V. The performance of our approach for different significant level decomposition (in the unit $\%$).

$\lambda^c = 0.5$	Case (1)		Case (2)	
	mean	SE	mean	SE
$m = 100, C = 5000$	6.8	0.37	7.0	0.36
$m = 100, C = 10000$	6.3	0.36	6.2	0.43
$m = 100, C = 50000$	4.2	0.31	5.5	0.35
$m = 500, C = 5000$	11.1	0.31	11.4	0.28
$m = 500, C = 10000$	7.4	0.2	8.5	0.24
$m = 500, C = 50000$	4.6	0.15	5.9	0.45
$\lambda^c = 0.7$	Case (1)		Case (2)	
	mean	SE	mean	SE
$m = 100, C = 5000$	12.6	1.06	11.1	1.15
$m = 100, C = 10000$	9.5	0.96	9.4	0.87
$m = 100, C = 50000$	7.9	0.89	8.3	0.75
$m = 500, C = 5000$	15.5	0.39	15.9	0.41
$m = 500, C = 10000$	11.2	0.37	10.7	0.34
$m = 500, C = 50000$	6.6	0.38	7.1	0.38

$Y = \max(X_1 + X_2 + X_5, X_1 + X_4, X_3 + X_5)$. We are interested in the expected time to finish the project $\mathbf{E}[Y]$.

To evaluate our approach, we pretend that the distributions for all five tasks are unknown and they are estimated by m "real-world data" $\mathbf{X}_m$ from underlying true distributions. The impact of the input uncertainty on the system performance estimates can be quantified by the $(1-\beta)100\%$ percentile Q of the induced distribution of $\mathbf{E}[Y(F^{(b)})]$ with $F^{(b)} \sim \tilde{F}(\cdot|\mathbf{X}_m)$. To evaluate the robustness of our approach on the percentile Q estimation, we consider the amount of real-world data $m = 100, 500$, the significant level for input uncertainty $1-\beta = 90\%, 95\%$, and the simulation budget $C = 5000, 10000, 20000$.

Monte Carlo simulation is used to estimate the true percentile Q in the experiments and a side experiment of 10 macro-replications is conducted to find B_1, the number of bootstrapped samples required to get a precise percentile estimation. Although $m = 500$ has less input uncertainty, it is more likely to generate the extreme observations, which could have a large impact on the system mean performance estimates. Thus, both $m = 100$ and $m = 500$ are considered. Specifically, in each macro-

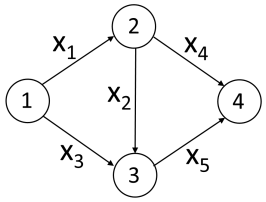

Fig. 1. A stochastic activity network

Table VI. The maximum absolute relative difference for percentile Q estimation (in the unit $\%$).

Number of bootstrapped samples B_1	1000			5000			10000	
Amount of simulations n_{B_1}	10^3	10^4	10^5	10^3	10^4	10^5	10^3	10^4
$m = 100, 1-\beta = 90\%$	7.8	1.4	0.8	4.0	1.9	1.0	7.0	1.8
$m = 100, 1-\beta = 95\%$	4.1	2.0	0.9	7.2	1.9	0.5	4.0	1.9
$m = 500, 1-\beta = 90\%$	8.5	2.5	0.9	7.2	1.9	0.8	8.2	2.6
$m = 500, 1-\beta = 95\%$	7.3	2.1	1.6	6.7	1.2	0.9	6.5	1.2

replication, we draw m real-world data, generate B_1 bootstrapped samples of input distributions, run n replications at each bootstrapped sample, and use the sample mean $\bar{Y}_b = \bar{Y}(F^{(b)})$ to estimate $\mathbf{E}[Y(F^{(b)})]$ for $b = 1, 2, \ldots, B_1$. The percentile Q is estimated by the order statistics $\widehat{Q}_{B_1} = \bar{Y}_{((1-\beta)B_1)}$. We select the combination of $B_1 = 10,000$ and $n = 100,000$ as the benchmark. Then, for different combinations of B_1 and n, we record the maximum relative error with the unit to be percentage ($\%$) obtained from 10 macro-replications in Table VI. Balancing precision and computational cost, we use $B_1 = 5,000$ and $n = 100,000$ to estimate the true percentile Q in our experiments because it achieves accuracy with the maximum relative error less than 1% for situations that we consider.

4.2.1. Comparing Our Approach with Direct Bootstrapping. We first study the performance of our approach and the direct bootstrapping under a variety of budgets. The experimental settings are very similar to the $M/M/1$ queueing example in Section 4.1. For our approach, we let $\alpha = 0.05$ which is further decomposed into $\alpha_o = 0.02$, $\alpha_S = \alpha_E = 0.01$. We employ the pure random search method to find good parameters (B, n_0, M, R) based on the initial simulations with $B_0 = 200$ and $n_{00} = 10$. The candidate pool is as follows: $n_0 \in \{10, 20, 30, 40, 50\}$, $M \in \{1, 2, \ldots, 10\}$ and $R \in \{1.1, 1.2, \ldots, 2\}$. Since the slight change in B does not have a significant impact on the accuracy, we select $B \in \{200, 250, \ldots, C/10\}$. The pure random search terminates and reports the best option it finds after 200 random searches. For the direct bootstrapping, we equally allocate the total budget C to $B = 1000$ bootstrapped samples of input distributions.

The results of mean and SE of the absolute relative error for the percentile Q estimation obtained from our approach and direct bootstrapping are shown in Tables VII-VIII. They are based on 100 macro-replications. We can observe that our approach performs much better than the direct bootstrapping in all cases. The precision of our approach improves quickly as the budget C increases. It achieves remarkable accuracy with tight budget $C = 20,000$. Notice both our approach and the direct bootstrapping attain better accuracy when $m = 100$ due to the fact that $m = 500$ has a larger chance to obtain extreme observations.

4.2.2. Impact of the Number of Searches N_s. In this section, we study the effect of the number of searches N_s used in the pure random search. In Section 4.2.1, we set the number of searches equal to 200, and here we set $N_s = 100$. Since $1-\beta = 95\%$ quantile is more extreme than $1-\beta = 90\%$ and more challenging to get accurate results, we only

Table VII. The relative error of the percentile Q estimates when $1-\beta = 90\%$ (in the unit $\%$).

	our approach		direct bootstrap	
	mean	SE	mean	SE
$m = 100, C = 5000$	5.9	0.26	32.8	2.98
$m = 100, C = 10000$	2.5	0.16	26.3	1.92
$m = 100, C = 20000$	1.7	0.12	15.1	1.15
$m = 500, C = 5000$	14.4	0.31	40.0	3.12
$m = 500, C = 10000$	9.9	0.22	24.8	1.76
$m = 500, C = 20000$	6.3	0.18	16.0	1.09

Table VIII. The relative error of the percentile Q estimates when $1-\beta = 95\%$ (in the unit $\%$).

	our approach		direct bootstrap	
	mean	SE	mean	SE
$m = 100, C = 5000$	9.3	0.33	33.2	2.64
$m = 100, C = 10000$	4.6	0.26	26.3	1.79
$m = 100, C = 20000$	1.7	0.15	15.6	1.13
$m = 500, C = 5000$	19.9	0.41	38.2	2.55
$m = 500, C = 10000$	13.5	0.25	24.9	1.66
$m = 500, C = 20000$	9.1	0.27	15.6	1.15

Table IX. The relative error of the percentile Q estimates when $1-\beta = 95\%$ (in the unit $\%$).

	$N_s = 200$		$N_s = 100$	
	mean	SE	mean	SE
$m = 100, C = 5000$	9.3	0.33	9.5	0.37
$m = 100, C = 10000$	4.6	0.26	4.9	0.28
$m = 100, C = 20000$	1.7	0.15	2.3	0.18
$m = 500, C = 5000$	19.9	0.41	20.3	0.44
$m = 500, C = 10000$	13.5	0.25	14.5	0.35
$m = 500, C = 20000$	9.1	0.27	10.0	0.36

consider $1-\beta = 95\%$. Table IX provides results of mean and SE of absolute relative error of the percentile Q estimation when $N_s = 100, 200$. The results with $N_s = 100$ are only slightly worse than those with $N_s = 200$, which indicate that a good but not optimal (B, n_0, R, M) could be sufficient to make our approach efficient and accurate.

4.3. The Overhead Cost

Here, we only report the overhead cost introduced by our approach since it is negligible for the direct bootstrapping. For the $M/M/1$ queue, the average overhead cost is 0.6, 1 and 2.4 seconds per simulation run for $C = 5000, 10000$ and 50000 respectively. For the stochastic activity network example, the average overhead cost is 0.76, 0.8 and 1.2 seconds per simulation run for $C = 5000, 10000$ and 20000 respectively. Thus, our approach requires a few seconds overhead cost per simulation run. In addition, as C increases, the overhead cost also increases. This matches well with the conclusion that the overhead cost is $O(C^3)$; see the Appendix for the derivation.

For both $M/M/1$ queue and stochastic activity network examples, our approach provides a better estimation of the percentile Q than the direct bootstrapping with the simulation budget doubled according to results in Tables II-III and VII-VIII. This indicates a clear advantage of our budget allocation approach even when its overhead cost takes half of the total computational resource. *Thus, our approach is more preferable under the situations when each simulation run is computationally expensive (takes more than seconds) and the simulation budget is tight.*

5. CONCLUSIONS

When we use simulation to assess the stochastic system performance, there exist both input and simulation estimation uncertainty in the performance estimates. The nonparametric bootstrap is used to quantify the input uncertainty, including both input distribution family and parameter value uncertainty. In this paper, we develop a sequential approach to efficiently propagate the input uncertainty to the output mean, while reducing the simulation estimation uncertainty. It can gradually explore the system performance at bootstrapped samples of input models, find the important samples that contribute the most to the percentile estimation and allocate more simulation resource there. Compared to the direct bootstrapping that equally allocates the simulation budget to all bootstrapped samples of input distributions, our approach demonstrates better and more robust performance, especially when the simulation budget is tight.

ACKNOWLEDGMENTS

The authors acknowledge helpful advice from Barry L. Nelson, Enlu Zhou and Cheng Li. They thank the associate editor, three anonymous referees for helpful comments. Portions of this paper were published in Yi et al. [2015].

REFERENCES

Russell R. Barton. 2007. Presenting A More Complete Characterization of Uncertainty: Can It Be Done?. In *Proceedings of the 2007 INFORMS Simulation Society Research Workshop*. INFORMS Simulation Society, Fontainebleau.

Russell R. Barton. 2012. Tutorial: Input Uncertainty in Output Analysis. In *Proceedings of the 2012 Winter Simulation Conference*, C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher (Eds.). IEEE Computer Society, Washington, DC, 67–78.

Russel R. Barton, Barry L. Nelson, and Wei Xie. 2014. Quantifying Input Uncertainty via Simulation Confidence Intervals. *INFORMS Journal on Computing* 26, 1 (2014), 74–87.

Russell R. Barton and Lee W. Schruben. 1993. Uniform And Bootstrap Resampling of Input Distributions. In *Proceedings of the 1993 Winter Simulation Conference*, G. W. Evans, M. Mollaghasemi, E. C. Russell, and W. E. Biles (Eds.). IEEE Computer Society, Washington, DC, 503–508.

Russell R. Barton and Lee W. Schruben. 2001. Resampling Methods for Input Modeling. In *Proceedings of the 2001 Winter Simulation Conference*, B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer (Eds.). IEEE Computer Society, Washington, DC, 372–378.

Evren R. Baysal and Jeremy Staum. 2008. Empirical Likelihood for Value-at-Risk and Expected Shortfall. *The Journal of Risk* 11, 1 (2008), 3–32.

Bahar Biller and Canan G. Corlu. 2011. Accounting for Parameter Uncertainty in Large-Scale Stochastic Simulations with Correlated Inputs. *Operations Research* 59 (2011), 661–673.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, New York.

Justin Boesel, Barry L. Nelson, and Seong-Hee Kim. 2003. Using Ranking and Selection to 'Clean Up' after Simulation Optimization. *Operations Research* 51, 5 (2003), 814–825.

Mark Broadie, Yiping Du, and Ciamac C. Moallemi. 2011. Efficient Risk Estimation via Nested Sequential Simulation. *Management Science* 57, 6 (2011), 1172–1194.

S. H. Brooks. 1958. A Discussion of Random Methods for Seeking Maxima. *Operations Research* 6 (1958), 244–251.
Russell C. H. Cheng and Wayne Holland. 1997. Sensitivity of Computer Simulation Experiments to Errors in Input Data. *Journal of Statistical Computation and Simulation* 57 (1997), 219–241.
Stephen E. Chick. 1997. Bayesian Analysis for Simulation Input and Output. In *Proceedings of the 1997 Winter Simulation Conference*, S. Andradóttir, K. J. Healy, D. H. Withers. and B. L. Nelson (Ed.). IEEE, 253–260.
Stephen E. Chick. 2001. Input Distribution Selection for Simulation Experiments: Accounting for Input Uncertainty. *Operations Research* 49 (2001), 744–758.
Michael B. Gordy and Sandeep Juneja. 2010. Nested Simulation in Portfolio Risk Measurement. *Management Science* 56, 10 (2010), 1833–1848.
Peter Hall. 1992. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag New York.
J. M. Hammersley and K. W. Morton. 1956. A New Monte Carlo Technique: Antithetic Variates. *Mathematical Proceedings of the Cambridge Philosophical Society* 52, 3 (1956), 449–475.
L. Jeff Hong and Barry L. Nelson. 2006. Discrete Optimization via Simulation Using COMPASS. *Operations Research* 54, 1 (2006), 115–129.
L. Jeff Hong, Barry L. Nelson, and Jie Xu. 2014. Discrete Optimization via Simulation. In *Handbook on Simulation Optimization*, M. C. Fu (Ed.). Springer, New York, 9–44.
Joel L. Horowitz. 2001. *The Bootstrap*. Handbook of Econometrics, Vol. 5. North Holland, Oxford, UK.
D. Jones, M. Schonlau, and W Welch. 1998. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization* 13 (1998), 455–492.
Seong-Hee Kim and Barry L. Nelson. 2007. Recent Advances in Ranking and Selection. In *Proceedings of the 2007 Winter Simulation Conference*. IEEE Computer Society, Washington, DC, 162–172.
Hai Lan, Barry L. Nelson, and Jeremy Staum. 2010. A Confidence Interval Procedure for Expected Shortfall Risk Measurement via Two-Level Simulation. *Operations Research* 58, 5 (2010), 1481–1490.
Shing-Hoi Lee and Peter W. Glynn. 2003. Computing the Distribution Function of a Conditional Expectation via Monte Carlo: Discrete Conditioning Spaces. *ACM Transactions on Modeling and Computer Simulation* 13, 3 (2003), 238–258.
Vadim Lesnevski, Barry L. Nelson, and Jeremy Staum. 2007. Simulation of Coherent Risk Measures Based on Generalized Scenarios. *Management Science* 53, 11 (2007), 1756–1769.
Vadim Lesnevski, Barry L. Nelson, and Jeremy Staum. 2008. An Adaptive Procedure for Estimating Coherent Risk Measures Based on Generalized Scenarios. *Journal of Computational Finance* 11, 4 (2008), 1–31.
Ming Liu and Jeremy Staum. 2010. Stochastic Kriging for Efficient Nested Simulation of Expected Shortfall. *The Journal of Risk* 12 (2010), 3–27.
Barry L. Nelson and David Goldsman. 2001. Comparisons with a Standard in Simulation Experiments. *Management Science* 47, 3 (2001), 449–463.
Barry L. Nelson, Julie Swann, David Goldsman, and Wheyming Song. 2001. Simple Procedures for Selecting the Best Simulated System when the Number of Alternatives is Large. *Operations Research* 49 (2001), 950–963.
Szu H. Ng and Stephen E. Chick. 2006. Reducing Parameter Uncertainty for Stochastic Systems. *ACM Transactions on Modeling and Computer Simulation* 16 (2006), 26–51.
Jun Shao and Dongsheng Tu. 1995. *The Jackknife and Bootstrap*. Springer-Verlag.
Lihua Sun, Jeff L. Hong, and Zhaolin Hu. 2014. Balancing Exploitation and Explo-

ration in Discrete Optimization via Simulation Through a Gaussian Process-Based Search. *Operations Research* 62 (2014), 1416–1438.

Yunpeng Sun, Daniel W. Apley, and Jeremy Staum. 2011. Efficient Nested Simulation for Estimating the Variance of a Conditional Expectation. *Operations Research* 59, 4 (2011), 998–1007.

Shing Chih Tsai, Barry L. Nelson, and Jeremy Staum. 2009. *Combined Screening and Selection of the Best with Control Variates*. International Series in Operations Research & Management Science, Vol. 133. Springer, New York.

Wei Xie, Cheng Li, and Pu Zhang. 2017. A Bayesian Nonparametric Hierarchical Framework for Uncertainty Quantification in Simulation. (2017). Under Review.

Wei Xie, Barry L. Nelson, and Russell. R. Barton. 2014a. A Bayesian Framework for Quantifying Uncertainty in Stochastic Simulation. *Operations Research* 62, 6 (2014), 1439–1452.

Wei Xie, Barry L. Nelson, and Russell R. Barton. 2014b. Statistical Uncertainty Analysis for Stochastic Simulation with Dependent Input Models. In *Proceedings of the 2014 Winter Simulation Conference*, A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller (Eds.). Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey, 674–685.

Wei Xie, Barry L. Nelson, and Russell R. Barton. 2016b. Statistical Uncertainty Analysis for Stochastic Simulation. (2016b). Working Paper, Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY.

Jie Xu, Barry L. Nelson, and L. Jeff Hong. 2010. Industrial Strength COMPASS: A Comprehensive Algorithm and Software for Optimization via Simulation. *ACM Transactions on Modeling and Computer Simulation* 20, 1 (2010), 1–29.

Jie Xu, Barry L. Nelson, and L. Jeff Hong. 2013. An Adaptive Hyperbox Algorithm for High-Dimensional Discrete Optimization via Simulation Problems. *INFORMS Journal on Computing* 25, 1 (2013), 133–146.

Yuan Yi, Wei Xie, and Enlu Zhou. 2015. A Sequential Experiment Design for Input Uncertainty Quantification in Stochastic Simulation. In *Proceedings of the 2015 Winter Simulation Conference*. IEEE Computer Society, Washington, DC.

Faker Zouaoui and James R. Wilson. 2003. Accounting for Parameter Uncertainty in Simulation Input Modeling. *IIE Transactions* 35 (2003), 781–792.

Online Appendix to: An Efficient Budget Allocation Approach for Quantifying the Impact of Input Uncertainty in Stochastic Simulation

YUAN YI, WEI XIE, Rensselaer Polytechnic Institute

PROOF. (THEOREM 3.1) Let $N(y) \equiv \sum_{b=1}^{B} \mathbf{1}(\mu_b \geq y)$. We first prove that the CI lower bound equals to the order statistics $\mu_{[k_1]}$. Define the set

$$S_L \equiv \left\{ y : \sum_{n=N(y)+1}^{B} \binom{B}{n} \beta^n (1-\beta)^{B-n} \geq \frac{\alpha_o}{2} \right\}. \tag{14}$$

By Baysal and Staum [2008], the CI lower bound is $C_L^o = \inf S_L$. Define the set of sample order statistics satisfying the inequality in (14) as

$$S_1 \equiv \left\{ \mu_{[b]} \text{ with } b = 1, 2, \ldots, B : \sum_{n=N(\mu_{[b]})+1}^{B} \binom{B}{n} \beta^n (1-\beta)^{B-n} \geq \frac{\alpha_o}{2} \right\}$$

and $S_1 \subseteq S_L$. Let k_1 to be the smallest index such that the order statistics $\mu_{[k_1+1]} = \min S_1$. We claim that $C_L^o = \mu_{[k_1]}$.

We prove $C_L^o = \mu_{[k_1]}$ by contradiction. Since $\mu_{[k_1]}$ does not satisfy the inequality in (14), we have $C_L^o \geq \mu_{[k_1]}$. To prove $C_L^o = \mu_{[k_1]}$, we only need to show that $C_L^o > \mu_{[k_1]}$ does not hold. Assume $C_L^o > \mu_{[k_1]}$. Then, there exists $\bar{q}_1 \in \Re$ such that $\mu_{[k_1]} < \bar{q}_1 < C_L^o \leq \mu_{[k_1+1]}$. Since $N(\bar{q}_1) = N(\mu_{[k_1+1]}) = B - k_1$ and $\mu_{[k_1+1]} \in S_L$, we have $\bar{q}_1 \in S_L$, which contradicts with $C_L^o > \bar{q}_1$. Thus, $C_L^o = \mu_{[k_1]}$.

Similarly, we prove that the CI upper bound equals to the order statistics $\mu_{[k_2]}$. Define

$$S_U \equiv \left\{ y : \sum_{n=0}^{N(y)} \binom{B}{n} \beta^n (1-\beta)^{B-n} \geq \frac{\alpha_o}{2} \right\}.$$

By Baysal and Staum [2008], we have $C_U^o = \sup S_U$. Define the set of sample order statistics as

$$S_2 \equiv \left\{ \mu_{[b]} \text{ with } b = 1, 2, \ldots, B : \sum_{n=0}^{N(\mu_{[b]})} \binom{B}{n} \beta^n (1-\beta)^{B-n} \geq \frac{\alpha_o}{2} \right\}$$

and $S_2 \subseteq S_U$. Let k_2 to be the largest index such that the order statistics $\mu_{[k_2]} = \max S_2$ and we also have $\mu_{[k_2]} \in S_U$. We claim that $C_U^o = \mu_{[k_2]}$.

Since $\mu_{[k_2]} \in S_U$, we have $C_U^o \geq \mu_{[k_2]}$. To show $C_U^o = \mu_{[k_2]}$, we only need to show that $C_U^o > \mu_{[k_2]}$ does not hold. We prove it by contradiction. Assume $C_U^o > \mu_{[k_2]}$. Then, there exists $\bar{q}_2 \in S_U$ such that $C_U^o > \bar{q}_2 > \mu_{[k_2]}$. If $\bar{q}_2 \geq \mu_{[k_2+1]}$, it is evident that $N(\bar{q}_2) \leq N(\mu_{[k_2+1]})$. If $\mu_{[k_2]} < \bar{q}_2 < \mu_{[k_2+1]}$, we have $N(\bar{q}_2) = N(\mu_{[k_2+1]}) = B - k_2$. For both cases, since $\bar{q}_2 \in S_U$, $\sum_{n=0}^{N(\mu_{[k_2+1]})} \binom{B}{n} \beta^n (1-\beta)^{B-n} \geq \alpha_o/2$ also holds. Thus, we have $\mu_{[k_2+1]} \in S_U$ and $\mu_{[k_2+1]} \in S_2$. Since k_2 is the largest index in S_2, there is a contradiction. Therefore, $\mu_{[k_2]} = C_U^o$. □

DOI: 0000001.0000001

PROOF. (THEOREM 3.2) Since $\mu_{[k_2]} \geq \mu_{[b]} \geq \mu_{[k_1]}$ for $b = k_1, k_1+1, \ldots, k_2$, we have $P(\mu_{[b]} \in [C_L, +\infty)) \geq 1 - \alpha_I/2$. Similarly, we have $P(\mu_{[b]} \in (-\infty, C_U]) \geq 1 - \alpha_I/2$. Thus, $P(\mu_{[b]} \in [C_L, C_U]) \geq 1 - \alpha_I$ for all $b \in \mathcal{F}_C$.

□

PROOF. (THEOREM 3.3)

$$
\begin{aligned}
& P\Big(Q \in [C_L, C_U] | \mathbf{X}_m\Big) \\
& \geq P\Big(Q \in [\mu_{[k_1]}, \mu_{[k_2]}] \text{ and } \mu_{[b]} \in [C_L, C_U] \text{ for } b = k_1, k_1+1, \ldots, k_2 | \mathbf{X}_m\Big) \\
& \geq 1 - P\Big(Q \notin [\mu_{[k_1]}, \mu_{[k_2]}] | \mathbf{X}_m\Big) - P\Big(\mu_{[b]} \notin [C_L, C_U] \text{ for } b = k_1, k_1+1, \ldots, k_2 | \mathbf{X}_m\Big) \quad (15) \\
& \geq 1 - \alpha_o - \alpha_I = 1 - \alpha. \quad (16)
\end{aligned}
$$

Step (15) follows by applying the Bonferroni inequality and Step (16) follows by applying Theorems 3.1 and 3.2. □

PROOF. (THEOREM 3.4) Define

$$
\begin{aligned}
A_1 &= \{ \text{ all } j \neq [k_q] \text{ with } \mu_j \leq \mu_{[k_q]} : \bar{Y}_j \leq \bar{Y}_{[k_q]} + W_{[k_q]j} \} \\
A_2 &= \{ \text{ all } j \neq [k_q] \text{ with } \mu_j \geq \mu_{[k_q]} : \bar{Y}_j \geq \bar{Y}_{[k_q]} - W_{[k_q]j} \}.
\end{aligned}
$$

Let $A = A_1 \cap A_2$. Since $A \subseteq \{[k_q] \in I^q\}$, to prove $\Pr([k_q] \in I^q) \geq 1 - \alpha_S$, we only need to show $\Pr(A) \geq 1 - \alpha_S$. We have

$$
\begin{aligned}
& \Pr(A_1^c) = \Pr\left(\exists j \neq [k_q] \text{ with } \mu_j \leq \mu_{[k_q]} : \bar{Y}_j > \bar{Y}_{[k_q]} + W_{[k_q]j}\right) \\
& = \Pr\left(\bigcup_{j \neq [k_q]: \mu_j \leq \mu_{[k_q]}} \left\{\bar{Y}_j > \bar{Y}_{[k_q]} + W_{[k_q]j}\right\}\right) \\
& \leq \sum_{j \neq [k_q]: \mu_j \leq \mu_{[k_q]}} \Pr\left(\bar{Y}_j > \bar{Y}_{[k_q]} + W_{[k_q]j}\right) \\
& = \sum_{j \neq [k_q]: \mu_j \leq \mu_{[k_q]}} \Pr\left(\frac{\bar{Y}_j - \bar{Y}_{[k_q]} - (\mu_j - \mu_{[k_q]})}{S_{[k_q]j}/\sqrt{n}} > \frac{W_{[k_q]j} + (\mu_{[k_q]} - \mu_j)}{S_{[k_q]j}/\sqrt{n}}\right) \\
& \leq \sum_{j \neq [k_q]: \mu_j \leq \mu_{[k_q]}} \Pr\left(\frac{\bar{Y}_j - \bar{Y}_{[k_q]} - (\mu_j - \mu_{[k_q]})}{S_{[k_q]j}/\sqrt{n}} > t_{n-1, 1-\frac{\alpha_S}{B-1}}\right) \\
& \leq (k_q - 1) \cdot \frac{\alpha_S}{B-1}.
\end{aligned}
$$

Similarly, we can have $\Pr(A_2^c) = (B - k_q) \cdot \alpha_S/(B-1)$. Therefore,

$$
\begin{aligned}
& \Pr([k_q] \in I^q) \geq \Pr(A) \\
& \geq 1 - \Pr(A_1^c) - \Pr(A_2^c) \quad (17) \\
& = 1 - (k_q - 1)\frac{\alpha_S}{B-1} - (B - k_q)\frac{\alpha_S}{B-1} \\
& = 1 - \alpha_S.
\end{aligned}
$$

Step (17) follows by applying the Bonferroni inequality. Thus, $\Pr([k_q] \in I^q) \geq 1-\alpha_S$ for $q = 1, 2$. $\square$

PROOF. (THEOREM 3.5) In Section 3.2.2, we construct the CI for the quantile Q,

$$[C_L, C_U] = \left[\min_{b \in I_M^1} \left(\bar{Y}_b - t_{N_2'(b)-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_{b,a}}{\sqrt{N_2'(b)}} \right), \max_{b \in I_M^2} \left(\bar{Y}_b + t_{N_2'(b)-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_{b,a}}{\sqrt{N_2'(b)}} \right) \right].$$

For the notation simplification, we drop conditional on $\mathbf{X}_m$ in the following proof. We first consider $\mu_{[k_1]}$ and $\mu_{[k_2]}$,

$$\begin{aligned}
&\Pr\left\{ \mu_{[k_1]} \geq \min_{b \in I_M^1} \left(\bar{Y}_b - t_{N_2'(b)-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_{b,a}}{\sqrt{N_2'(b)}} \right) \right\} \\
&\geq \Pr\left\{ [k_1] \in I_M^1 \text{ and } \mu_{[k_1]} \geq \min_{b \in I_M^1} \left(\bar{Y}_b - t_{N_2'(b)-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_{b,a}}{\sqrt{N_2'(b)}} \right) \right\} \\
&\geq \Pr\left([k_1] \in I_M^1 \text{ and } \mu_{[k_1]} \geq \bar{Y}_{[k_1]} - t_{N_2'([k_1])-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_{[k_1],a}}{\sqrt{N_2'([k_1])}} \right) \\
&\geq 1 - \Pr\left([k_1] \notin I_M^1 \right) - \Pr\left(\mu_{[k_1]} < \bar{Y}_{[k_1]} - t_{N_2'([k_1])-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_{[k_1],a}}{\sqrt{N_2'([k_1])}} \right) \\
&\geq 1 - \alpha_S - \frac{\alpha_E}{2},
\end{aligned}$$

and

$$\begin{aligned}
&\Pr\left\{ \mu_{[k_2]} \leq \max_{b \in I_M^2} \left(\bar{Y}_b + t_{N_2'(b)-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_{b,a}}{\sqrt{N_2'(b)}} \right) \right\} \\
&\geq \Pr\left\{ [k_2] \in I_M^2 \text{ and } \mu_{[k_2]} \leq \max_{b \in I_M^2} \left(\bar{Y}_b + t_{N_2'(b)-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_{b,a}}{\sqrt{N_2'(b)}} \right) \right\} \\
&\geq \Pr\left([k_2] \in I_M^2 \text{ and } \mu_{[k_2]} \leq \bar{Y}_{[k_2]} + t_{N_2'([k_2])-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_{[k_2],a}}{\sqrt{N_2'([k_2])}} \right) \\
&\geq 1 - \Pr\left([k_2] \notin I_M^2 \right) - \Pr\left(\mu_{[k_2]} > \bar{Y}_{[k_2]} + t_{N_2'([k_2])-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_{[k_2],a}}{\sqrt{N_2'([k_2])}} \right) \\
&\geq 1 - \alpha_S - \frac{\alpha_E}{2}.
\end{aligned}$$

Thus, for any $b = k_1, k_1+1, \ldots, k_2$, since $\mu_{[k_1]} \leq \mu_{[b]} \leq \mu_{[k_2]}$,

$$\begin{aligned}
&\Pr\left\{ \mu_{[b]} \geq \min_{i \in I_M^1} \left(\bar{Y}_i - t_{N_2'(i)-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_{i,a}}{\sqrt{N_2'(i)}} \right) \right\} \\
&\geq \Pr\left\{ \mu_{[k_1]} \geq \min_{i \in I_M^1} \left(\bar{Y}_i - t_{N_2'(i)-1,1-\frac{\alpha_E}{2}} \cdot \frac{S_{i,a}}{\sqrt{N_2'(i)}} \right) \right\} \\
&\geq 1 - \alpha_S - \frac{\alpha_E}{2},
\end{aligned}$$

and

$$\begin{aligned}
&\Pr\left\{\mu_{[b]} \le \max_{i\in I_M^2}\left(\bar{Y}_i + t_{N_2'(i)-1,1-\frac{\alpha_E}{2}}\cdot\frac{S_{i,a}}{\sqrt{N_2'(i)}}\right)\right\}\\
&\ge \Pr\left\{\mu_{[k_2]} \le \max_{i\in I_M^2}\left(\bar{Y}_i + t_{N_2'(i)-1,1-\frac{\alpha_E}{2}}\cdot\frac{S_{i,a}}{\sqrt{N_2'(i)}}\right)\right\}\\
&\ge 1-\alpha_S-\frac{\alpha_E}{2}.
\end{aligned}$$

Thus, for any $b = k_1, k_1+1, \ldots, k_2$, we have $\Pr(\mu_{[b]} \in [C_L, C_U]|\mathbf{X}_m) \ge 1-\alpha_S-\frac{\alpha_E}{2}-\alpha_S-\frac{\alpha_E}{2} = 1-\alpha_I$. □

PROOF. (OVERHEAD COST OF OUR SEQUENTIAL APPROACH)
Screening and optimization search play dominant roles in the overhead computational cost introduced by our sequential approach. We first calculate the cost caused by the sequential screening. Let c_S be the total number of operations incurred from two-sided statistical tests,

$$c_S \le n_t \cdot c_t \tag{18}$$

where c_t is the maximum number of operations incurred from a single test and n_t is the maximum number of tests performed during the screening procedure. In the ℓth iteration, to determine if sample $i \in I_\ell^q$ should be remained in the surviving set I_ℓ^q, we do $|I_\ell^q|-1$ pairwise comparisons with all other samples in I_ℓ^q for $q = 1, 2$. The number of pairwise comparison is bounded by B. For each pairwise comparison, the dominant cost is to calculate the sample covariance $S_{ij}^2 = \frac{1}{N_1^q(\ell)-1}\sum_{h=1}^{N_1^q(\ell)}\left(Y_{ih}-Y_{jh}-(\bar{Y}_i-\bar{Y}_j)\right)^2$ with $i, j \in I_\ell^q$, which requires $O(C)$ multiplications. Therefore, we can get $c_t = O(BC)$.

We further calculate n_t. The total number of statistical tests performed in the ℓth iteration is $|I_{\ell-1}^1| + |I_{\ell-1}^2|$. Before these tests, $N_1^q(\ell) - N_1^q(\ell-1)$ new simulation replications are allocated to each surviving sample. It indicates that the total number of replications consumed is $(N_1^q(\ell) - N_1^q(\ell-1))(|I_{\ell-1}^1| + |I_{\ell-1}^2|)$, and for a new statistical test, $(N_1^q(\ell) - N_1^q(\ell-1))$ new replications would be used, which is at least one. Thus, $n_t \le C$. Plugging these results into (18), we get $c_S = O(BC^2)$. Since $B \le C/10$, the total cost required in the screening procedure is $c_S = O(C^3)$ multiplications.

Next, we consider the overhead cost introduced by our adaptive sequential approach including the optimization search. For simplification, we only consider the pure random search algorithm. The overhead cost required by the optimization search, denoted by c_O, is

$$c_O \le n_o \cdot c_S \le n_c \cdot c_S$$

where n_o is the number of candidates (B, n_0, R, M) visited and n_c is the total number of candidates under consideration given a tight budget C. Since it is desirable to keep the initial allocation n_0 small so that we can quickly screen out bootstrapped samples that are extremely unlikely to contribute to the percentile Q estimation [Lesnevski et al. 2007], we let $n_0 \in \{10, 20, 30, 40, 50\}$. Since [Lesnevski et al. 2008] recommended the growth factor in the range 1.2 to 2, and the performance does not improve much by altering the choice, we slightly extend the range and let $R \in \{1.1, 1.2, \ldots, 2\}$. Since our numerical study indicates that it is either infeasible or detrimental to have too large M, we let $M \in \{1, 2, \ldots, 10\}$. The choice of B is at most $C/10$. Thus, n_c is at most polynomial in C. Combining with $n_S = O(C^3)$, the overhead cost c_O introduced by our adaptive sequential approach is also polynomial in C. □