Distributionally Robust Selection of the Best

Weiwei Fan* L. Jeff Hong† Xiaowei Zhang‡

Abstract. Specifying a proper input distribution is often a challenging task in simulation modeling. In practice, there may be multiple plausible distributions that can fit the input data reasonably well, especially when the data volume is not large. In this paper, we consider the problem of selecting the best from a finite set of simulated alternatives, in the presence of such input uncertainty. We model such uncertainty by an ambiguity set consisting of a finite number of plausible input distributions, and aim to select the alternative with the best worst-case mean performance over the ambiguity set. We refer to this problem as robust selection of the best (RSB). To solve the RSB problem, we develop a two-stage selection procedure and a sequential selection procedure; we then prove that both procedures can achieve at least a user-specified probability of correct selection under mild conditions. Extensive numerical experiments are conducted to investigate the computational efficiency of the two procedures. Finally, we apply the RSB approach to study a queueing system's staffing problem using synthetic data and an appointment-scheduling problem using real data from a large hospital in China. We find that the RSB approach can generate decisions significantly better than other widely used approaches.

Keywords. selection of the best; distributional robustness; input uncertainty; probability of correct selection

1. Introduction

Simulation is widely used to facilitate decision-making for stochastic systems. In general, the performance of a stochastic system depends on *design* variables and *environmental* variables. The former is controllable by the decision-maker, while the latter is not. By simulating the environmental variables, the decision-maker can estimate the system's mean performance for arbitrary values of the design variables. A crucial step for building a credible simulation model is to characterize the environmental variables with an appropriate probability distribution, typically referred to as the *input distribution* in simulation literature. This is often

*Advanced Institute of Business and School of Economics and Management, Tongji University, 200092 Shanghai, China
†School of Management and School of Data Science, Fudan University, 200433 Shanghai, China
‡Corresponding author. Department of Management Sciences, College of Business, City University of Hong Kong, Kowloong Tong, Hong Kong. Email: xiaowei.w.zhang@cityu.edu.hk

difficult, mainly because of lack of enough data, or measurement error in the data, either of which causes uncertainty concerning the input distribution, i.e., *input uncertainty*.

Input uncertainty has drawn substantial interest from the simulation community in the past two decades; see Henderson (2003) for a survey. The existing work usually assumes that the input distribution belongs to a particular parametric family, but the parameters of the distribution need to be estimated. This assumption reduces the input uncertainty to the so-called *parameter uncertainty* and the primary objective becomes to characterize the randomness of the simulation output that is amplified by the parameter uncertainty. For instance, Cheng and Holland (1997) use the delta method to approximate the variance of the simulation output and Barton and Schruben (2001) use the bootstrap method.

However, in practice it is non-trivial to determine the proper parametric family. Indeed, there may be several plausible parametric families that fit the input data reasonably well if the data volume is not large. For instance, in Section 7, we study an appointment-scheduling problem in a large hospital in China. The maximum number of operations of a particular type performed by a surgeon in the hospital in 2014 is 138, and goodness-of-fit tests reject neither the gamma distribution nor the lognormal distribution when fitting the data for the duration of operations. Notice that these two parametric families may result in qualitatively different performances of a stochastic system. For instance, a queueing system's behavior depends critically on whether its service times are light-tailed or heavy-tailed (Asmussen 2003). Therefore, in this paper we focus on the uncertainty in specifying the parametric family of the input distribution, instead of considering parameter uncertainty.

One approach to address this difficulty is Bayesian model averaging (Chick 2001). It measures the stochastic system by the weighted average of its mean performance under different plausible input distributions, where the weights are specified by prior estimation of the likelihood that a particular plausible distribution is the "true" distribution. This approach takes an "ambiguity-neutral" viewpoint concerning the input uncertainty.

In this paper, we take a robust approach that adopts an "ambiguity-averse" (Epstein 1999) viewpoint and uses the worst-case mean performance of all the plausible distributions to assess a stochastic system. Using the worst-case analysis to account for uncertainty has a long history in economic theory. Ellsberg (1961) argues that in a situation where probability distributions cannot be specified completely, considering the worst of all the plausible distributions might appeal to a conservative person. Gilboa and Schmeidler (1989) rationalize the ambiguity aversion by showing that an individual who considers multiple prior probability distributions and maximizes the minimum expected utility over these distributions would act in this conservative manner. However, we do not argue or suggest that worst-case analysis is better than the "model-averaging" approach. Instead, we believe that they are equally important and that decision-makers should consider different perspectives in order to be fully aware of the potential risks of a decision.

We focus on an important class of simulation-based decision-making problems. We assume that the design variables of the stochastic system of interest take values from a finite set, each of which is referred to as an alternative. The mean performance of an alternative is estimated via simulation and we are interested in selecting the "best" alternative. This is known as the selection of the best (SB) problem in simulation literature. Due to statistical noise inherent in the simulation procedure, the probability that the best

alternative is not selected is nonzero regardless of the computational budget. Thus, the objective is to develop a selection procedure that selects the best alternative with some statistical guarantee; see Kim and Nelson (2006) for an overview. In this paper, we consider the SB problem in the presence of input uncertainty and solve it in a way that is robust with respect to input uncertainty.

1.1. Main Contributions

First, we model the input uncertainty as an ambiguity set consisting of finitely many plausible distribution families whose associated parameters are properly chosen. We then transform the SB problem in the presence of input uncertainty into a *robust selection of the best* (RSB) problem. Each alternative has a distinctive mean performance for each input distribution in the ambiguity set, and its worst-case mean performance is used as a measure of that alternative. The best alternative is defined as the one having the best worst-case mean performance.

Second, assuming the ambiguity set is given and fixed, we propose a new indifference-zone (IZ) formulation and design two selection procedures accordingly. The IZ formulation was proposed by Bechhofer (1954). However, to cope with our robust treatment of input uncertainty, we redefine the IZ parameter, denoted by δ, as the smallest difference between the *worst-case* mean performance of two alternatives that a decision-maker considers worth detecting. Then, the statistical evidence for designing a proper selection procedure can be expressed as the probability of selecting an alternative that is within δ of the best alternative in terms of their worst-case mean performance. We develop a two-stage procedure and a sequential procedure with statistical validity, i.e., they guarantee achieving a probability of correct selection (PCS) that is no less than a pre-specified level in a finite-sample regime and an asymptotic regime, respectively.

Third, we extend standard numerical tests for the SB problem to the new setting and demonstrate the computational efficiency of the two proposed RSB procedures. In particular, the sequential RSB procedure's efficiency in terms of the required total sample size is insensitive to the IZ parameter δ when δ is small enough. This is appealing to a practitioner, because it enables δ to be set as small as possible so that the unique best alternative can be selected instead of some "near-best" one without worrying computational burden. Besides, the proposed sequential RSB procedure is carefully designed so that it requires a much smaller total sample size than a plain-vanilla sequential RSB procedure as the problem scale increases.

Fourth, we assess the RSB approach in a queueing simulation environment where the input data, and thus the ambiguity set, is subject to random variation. Specifically, we consider a multi-server queue with abandonment, whose service time has an unknown distribution. The decision of interest is the staffing level, i.e., the number of servers. The cost of the queueing system depends on waiting and abandonment of the customers as well as the staffing level. We compare the RSB approach with a common approach for input modeling in practice, that is, the decision-maker fits a group of distribution families to the input data and uses the "best" fitted one as if it were the true distribution. An extensive numerical investigation reveals that the RSB approach can generate a staffing decision that has a significantly lower and more stable cost.

Finally, we apply the RSB approach to an appointment-scheduling problem using real data from a large

hospital in China. We show that in the presence of deep input uncertainty, the scheduling decision generated by the RSB approach incurs significantly lower operating costs than other widely used approaches, including a so-called "increasing order of variance" scheduling rule, one that is commonly viewed as a good heuristic in healthcare practice and was theoretically shown to be the optimal scheduling rule under some robust framework (Mak et al. 2015).

1.2. Related Literature

This paper is related to three streams of literature, i.e., simulation input uncertainty, robust optimization, and selection of the best. Studies of input uncertainty in simulation literature have focused on the impact of input uncertainty on simulation output analysis; for instance, constructing confidence intervals to reflect input uncertainty. A preferred approach is resampling, consisting of macro-replications, in each of which the input data is first resampled to construct an empirical distribution as the input distribution. The sampled empirical distribution is then used to drive the simulation model to estimate the performance of the involved stochastic system. Finally, the performance estimate is collected as a bootstrap statistic from each macro-replication and a dynamic confidence interval is constructed for the performance measure of interest. Representative articles include Cheng and Holland (1997) and Barton and Schruben (2001). Bayesian model averaging also relies on macro-replication, but each macro-replication begins with sampling from the posterior (based on the input data) of the plausible input distributions and then uses the sampled input distribution to drive the simulation model; see Chick (2001). Recently, Barton et al. (2014) and Xie et al. (2014) have both studied the propagation of input uncertainty to the estimated performance of a stochastic system, using nonparametric bootstrapping and Bayesian analysis, respectively.

The above research essentially takes an ambiguity-neutral attitude to the input uncertainty rather than ambiguity averse attitude as we do. In addition, it concentrates on the performance analysis of a stochastic system for a fixed value of its design variables. Unlike our paper, they do not include optimizing the performance over the design variables. Optimizing performance in the presence of distributional uncertainty is a theme of robust optimization; see Ben-Tal et al. (2009) for an introduction to this broad area. However, robust optimization literature generally does not consider cases in which an objective function is embedded in a black-box simulation model and can only be evaluated using random samples; an exception is Hu et al. (2012) but they focus on parameter uncertainty of the input distribution.

There is also a vast literature regarding the SB problem. Selection procedures can be categorized into frequentist procedures or Bayesian procedures depending on the viewpoint adopted for interpreting the unknown mean performance of an alternative. The former treats it as a constant and can be estimated through repeated sampling. Representative frequentist selection procedures include Rinott (1978), Kim and Nelson (2001), Chick and Wu (2005), Frazier (2014), and Fan et al. (2016). All these adopt an IZ formulation and use PCS as a selection criterion, except Chick and Wu (2005) in which the selection criterion is set to be expected opportunity cost, and Fan et al. (2016) in which an IZ-free formulation that can save users from the burden of specifying an appropriate IZ parameter is proposed. The present paper follows a frequentist viewpoint as well. Bayesian procedures, on the other hand, view the unknown mean of an alternative as a posterior distribution conditionally on samples calculated by Bayes' rule. The main approaches used in

the Bayesian framework include (i) optimal computing budget allocation (He et al. 2007), (ii) knowledge gradient (Frazier et al. 2009), (iii) expected value of information (Chick et al. 2010), and (iv) economics of selection procedures (Chick and Frazier 2012).

Few prior papers in SB literature address input uncertainty, except Corlu and Biller (2013, 2015), which focuses on the subset-selection formulation instead of the IZ formulation, and Song et al. (2015), which finds that in the presence of input uncertainty, IZ selection procedures designed for the SB problems may fail to deliver a valid statistical guarantee of correct selection for some configurations of the competing alternatives. Unlike our paper, these three papers all take an ambiguity-neutral viewpoint.

The rest of paper is organized as follows. Section 2 formulates the RSB problem. Sections 3 and 4 develop the two-stage and sequential RSB procedures, respectively, and show their statistical validity. Section 5 presents numerical experiments to demonstrate the computational efficiency of the two RSB procedures. In Section 6, we verify statistical validity of the proposed RSB procedures and demonstrate usefulness of the RSB approach in the context of queueing simulation, a more realistic setting than that of Section 5. In Section 7, we apply the RSB approach to address an appointment-scheduling problem using real data from a large hospital in China. We conclude in Section 8 and collect all the proofs and additional numerical results in Appendix.

2. Robust Selection of the Best

Suppose that a decision-maker needs to decide among k competing alternatives, i.e., $\mathcal{S} = \{s_1, s_2, \ldots, s_k\}$. For each s_i, $i = 1, \ldots, k$, let $g(s_i, \xi)$ denote its performance given an input variable ξ. In practice, ξ is typically random and follows probability distribution P_0. Notice that P_0 may differ between the alternatives, but we suppress its dependence on s_i for the purpose of notational simplicity. Each alternative is then measured by its mean performance $\mathbb{E}_{P_0}[g(s_i, \xi)]$, $i = 1, \ldots, k$. The decision-maker aims to select the best alternative from $\mathcal{S}$, which is defined as the one having the smallest mean performance,

$$\min_{s \in \mathcal{S}} \mathbb{E}_{P_0}[g(s, \xi)].$$

This is known as the SB problem, and a great variety of selection procedures have been developed, aiming to provide a desirable statistical guarantee on the probability of selecting the best.

To date, the SB problem has been studied primarily under the premise that the distribution P_0 is known and fixed. However, this is hardly the case in real-world applications. We assume that the distribution P_0 belongs to an *ambiguity set* $\mathcal{P}$ that consists of a finite number of plausible distributions, i.e., $\mathcal{P} = \{P_1, P_2, \ldots, P_m\}$. The form of $\mathcal{P}$ is determined by the following common scenario in input modeling: modern simulation software, e.g., Input Analyzer of Arena (Kelton et al. 2009), typically has a built-in functionality to fit input data to a specified parametric distribution family and to perform some goodness-of-fit tests (e.g., Kolmogorov-Smirnov test and chi-squared test). A preliminary exploration of the input data may suggest a set of plausible distribution families and they are then examined by the software one at a time. Hence, a typical example of $\mathcal{P}$ is such that each P_j belongs to a distinctive parametric family, whose parameters are estimated from the data and which is not rejected by the goodness-of-fit tests. Notice that

the ambiguity set $\mathcal{P}$ constructed in this way will converge to the true input distribution as the data volume increases, provided that the true distribution family is included in the set of plausible distribution families.

Given the ambiguity set $\mathcal{P}$, we measure an alternative by its worst-case mean performance over $\mathcal{P}$ and denote the best alternative as the alternative with the smallest worst-case mean performance. Then, the SB problem in the presence of input uncertainty is formulated as

$$\min_{s\in\mathcal{S}} \max_{P\in\mathcal{P}} \mathbb{E}_P[g(s,\xi)], \tag{1}$$

which we call the RSB problem. Our goal is to develop selection procedures that, upon termination, select the best alternative with a probability of at least a user-specified value $1-\alpha$, $(0<\alpha<1)$.

Remark 1. The formulation (1) assumes that $\mathcal{P}$ is given and fixed. On its own, it does not address the issue of *statistical consistency* in the sense that $\mathcal{P}$ converges to the unit set that contains only the true distribution P_0 as the size of the input data grows to infinity. Thus, this issue is not addressed by the RSB methodology developed here. For our methodology to perform correctly, certain mechanism needs to be implemented to ensure that all plausible distributions in $\mathcal{P}$ that are not P_0 would be discarded eventually as more input data becomes available. Using a goodness-of-fit test is one possible approach. But further theoretical work on this issue would be of interest.

Remark 2. There is a subtle but critical difference between the conventional SB context and the context of the present paper with regard to the concept of "random sample". In the former context, P_0 is known and the mean performance of each s_i is estimated by a random sample of size N of the simulation output $g(s_i,\xi)$ with ξ generated from P_0, so the estimate depends on N. In the RSB context, however, P_0 is unknown and the distributions in $\mathcal{P}$ all try to estimate P_0 based on a sample of it of size ℓ (i.e., the input data), so each $P_j\in\mathcal{P}$ depends on ℓ. Therefore, in a RSB procedure the estimate of each alternative's mean performance under each P_j generally depends on both N and ℓ. By assuming $\mathcal{P}$ is given and fixed, we essentially ignore the dependence on ℓ. A more complete treatment would account for the fact that each $P_j\in\mathcal{P}$ estimated from the input data is actually random and make ℓ a possible factor for designing a RSB procedure. But this is beyond the scope of the present paper.

Before moving to next section, we first introduce some necessary notations and assumptions. Let "system (i,j)" represent the pair of decision s_i and probability scenario P_j, and $g(s_i,\xi)$ with ξ following distribution P_j denote the random observation from system (i,j); further, let $\mu_{ij}=\mathbb{E}_{P_j}[g(s_i,\xi)]$ and $\sigma^2_{ij}=\mathrm{Var}_{P_j}[g(s_i,\xi)]$. The following assumptions are imposed throughout the paper.

Assumption 1. *For each* $i=1,2,\ldots,k$, $\mu_{i1}\geq\mu_{i2}\geq\cdots\geq\mu_{im}$. *Moreover,* $\mu_{11}<\mu_{21}\leq\ldots\leq\mu_{k1}$.

Assumption 2. *For each* $i=1,2,\ldots,k$ *and* $j=1,2,\ldots,m$, $\sigma^2_{ij}<\infty$.

In the RSB problem (1), our objective is to identify for each alternative its corresponding worst-case probability scenario, which is irrelevant to how the probability scenarios are ordered in the ambiguity set. Without loss of generality, we allow the means μ_{ij}'s to be of certain configuration presented in Assumption 1; otherwise, we can sort the means and relabel the systems in the desired order. Clearly, under Assumption

1, system $(i, 1)$ yields the worst-case mean performance of alternative i, $i = 1, \ldots, k$, and alternative 1 is the unique best alternative in (1).

Assumption 2 states that for each s_i, the random performance $g(s_i, \xi)$ is of finite variance under each P_j included in the ambiguity set. Considering $\mathrm{Var}_{P_0}[g(s_i, \xi)] < \infty$ in many practical situations, it is reasonable to choose the probability scenario P yielding the finite $\mathrm{Var}_P[g(s_i, \xi)]$ for all i as a candidate representative for P_0 and then include it into the ambiguity set. Besides, assuming a finite variance of the random performance of each system is common in SB literature.

2.1. Indifference-Zone Formulation

We adopt the IZ formulation to design RSB procedures. Under the IZ formulation, the sought procedures are expected to provide a lower bound for both the probability of correct selection (CS) and the probability of good selection (GS); see, e.g., Ni et al. (2017) for their definitions in the SB setting. Since the RSB problem is of a minimax structure different from SB problems, the CS and GS events need to first be carefully redefined.

Let δ be a pre-specified IZ parameter which is the smallest difference that the decision-maker deems worth detecting. If $\mu_{21} - \mu_{11} > \delta$, alternative 1 is better than the others by at least δ, measured by their worst-case mean performance over $\mathcal{P}$, due to Assumption 1. We define the CS event as the event where alternative 1 is selected. If $\mu_{21} - \mu_{11} \leq \delta$, some "good" alternatives exist, and their worst-case mean performances are within δ of alternative 1; decision-makers feel indifferent between those good alternatives and alternative 1. We define the GS event as the event where one of the good alternatives is selected. Hence, selecting alternative i is a good selection if $\mu_{i1} - \mu_{11} \leq \delta$.

Subtlety exists in the definitions of CS and GS, and it is worth some remarks. Take CS for an example. In the presence of the ambiguity set $\mathcal{P}$, it may be tempting to define CS as selecting system $(1, 1)$, which refers to a pair of the best alternative and its corresponding worst-case probability scenario. However, what matters to a decision-maker is to select the best alternative rather than identifying which input distribution yields the worst-case mean performance of the alternatives. This is because the selected alternative will be implemented later and the ambiguity set is merely used to evaluate the alternatives.

In SB literature, most IZ procedures are designed for the situation when $\mu_{21} - \mu_{11} > \delta$, and thus it is conventional to say a procedure is statistically valid if the achieved probability of correct selection (PCS) is no smaller than a pre-specified value $1 - \alpha$. Borrowing the notation from SB literature, we use PCS to denote a measure of statistical validity, but in the extended way. Particularly, we define PCS as the probability of CS if $\mu_{21} - \mu_{11} > \delta$ and the probability of GS otherwise. Then, this paper seeks RSB procedures with *statistical validity* in the following form: given a pre-specified $\alpha \in (0, 1)$

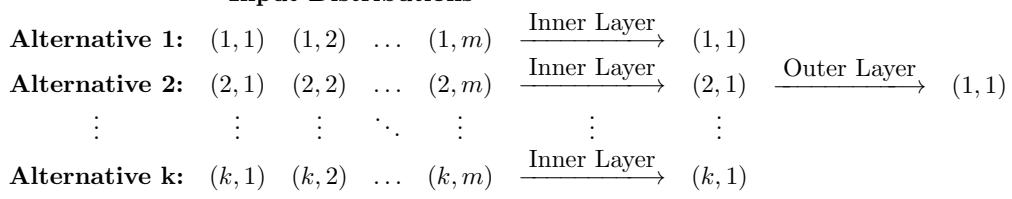
$$\mathbb{P}\{\mu_{i^*1} - \mu_{11} \leq \delta\} \geq 1 - \alpha, \tag{2}$$

where i^* is the index of the selected alternative upon termination of a procedure.

2.2. Double-Layer Structure

In view of the minimax formulation of the RSB problems (1), we propose a double-layer structure for designing RSB procedures. An inner-layer procedure aims to select system $(i, 1)$, which produces the worst-case mean performance for alternative i with at least a pre-chosen inner-layer PCS, for each $i = 1, 2, \ldots, k$. An outer-layer procedure, on the other hand, aims to select system $(1, 1)$ from the inner-layer selected systems with at least pre-chosen outer-layer PCS. After the double-layer selection process, we expect system $(1, 1)$ to be selected with at least the probability of $1 - \alpha$. To that end, the PCS in each layer must be judiciously chosen so that the overall PCS is no less than $1 - \alpha$ as in (2). The detailed discussion is deferred to Section 3.2. Figure 1 illustrates the double-layer structure of RSB procedures.

Figure 1: Two-layer Structure of RSB Procedures



Note. Systems in the same column do not necessarily have the same input distribution. They are ordered to satisfy Assumption 1. The methodology developed in this paper even allows the alternatives to have different ambiguity sets.

3. Two-Stage RSB Procedure

In this section, we develop a two-stage RSB procedure with the statistical validity in the form of (2). In the first stage, we take a small number of samples from each system to calculate the sample variance of each pairwise difference. Based on them, we calculate the sample sizes needed for the second stage, and select the best alternative based on the sample means obtained after the second-stage sampling.

The required sample size here resembles formally that of a typical two-stage procedure for the SB problem. However, the IZ parameter needs a special treatment to accommodate the double-layer structure. We wish to obtain an IZ for the RSB problem that has a given IZ parameter value. We do this by using two separate IZ parameter values, one corresponding to the inner-layer and the other to the outer-layer selection process, but calculated in a combined way so that the required IZ parameter value is obtained for the overall RSB problem.

The procedure is based on pairwise comparisons between the systems, each of which has a nonzero probability to yield an incorrect result due to simulation noise. We need to specify how much probability of making an error in each pairwise comparison is allowed, hereafter referred to as the *error allowance* and denoted by β, in order to achieve the overall PCS. It is nontrivial to specify error allowances due to the double-layer structure.

Assuming that the inner-layer and outer-layer IZ parameters, denoted by δ_I and δ_O, respectively, as well

as the error allowance are known, we now present the two-stage RSB procedure (Procedure T). Specification of (δ_I, δ_O) and β will be addressed in Section 3.1 and Section 3.2, respectively.

Procedure 1 (Procedure T: Two-Stage RSB Procedure).

0. *Setup.* Specify the inner-layer and outer-layer IZ parameters (δ_I, δ_O), the error allowance β, and the first-stage sample size $n_0 \geq 2$. Set $h = t_{1-\beta,n_0-1}$ to be the $100(1-\beta)\%$ quantile of the Student's t distribution with $n_0 - 1$ degrees of freedom.

1. *First-stage sampling.* Take n_0 independent replications $X_{ij,1}, \ldots, X_{ij,n_0}$ of each system (i, j). Compute the first-stage sample mean of each system and the sample variance of the difference between each pair of systems as follows

$$\bar{X}_{ij}(n_0) = \frac{1}{n_0}\sum_{r=1}^{n_0} X_{ij,r}, \quad 1 \leq i \leq k,\ 1 \leq j \leq m,$$

$$S^2_{ij,i'j'} = \frac{1}{n_0 - 1}\sum_{r=1}^{n_0}\left[X_{ij,r} - X_{i'j',r} - (\bar{X}_{ij}(n_0) - \bar{X}_{i'j'}(n_0))\right]^2, \quad 1 \leq i, i' \leq k,\ 1 \leq j, j' \leq m.$$

2. *Second-stage sampling.* Compute the total sample size $N = \max\limits_{(i,j),(i',j')}\left\{n_0, N_{ij,i'j'}\right\}$, where

$$N_{ij,i'j'} = \max\left\{\left\lceil\frac{h^2 S^2_{ij,i'j'}}{\delta_I^2}\right\rceil, \left\lceil\frac{h^2 S^2_{ij,i'j'}}{\delta_O^2}\right\rceil\right\}, \quad 1 \leq i, i' \leq k,\ 1 \leq j, j' \leq m,$$

with $\lceil x \rceil$ denoting the smallest integer no less than x. Take $N - n_0$ additional independent replications of each system.

3. *Selection.* Compute the overall sample mean of each system based on the N replications

$$\bar{X}_{ij}(N) = \frac{1}{N}\sum_{r=1}^{N} X_{ij,r}, \quad 1 \leq i \leq k,\ 1 \leq j \leq m.$$

Return $i^* = \arg\min\limits_{1\leq i\leq k}\max\limits_{1\leq j\leq m}\bar{X}_{ij}(N)$ as the best alternative. □

Similarly to two-stage procedures for the SB problem, the two-stage RSB procedure is easy to implement and variance reduction techniques, such as common random numbers, can readily be applied to increase the efficiency of the algorithm, where the efficiency is defined in terms of the total sample size required.

3.1. Inner-Layer and Outer-Layer IZ Parameters

The inner layer is used to estimate the worst-case mean performance of each alternative, while the outer layer is to compare the alternatives based on the estimates in the inner layer. If system (i, j_i^*) is correctly selected in the inner layer to represent the worst system of alternative i, then by the definition of δ_I and

Assumption 1,

$$0 < \mu_{i1} - \mu_{ij_i^*} \leq \delta_I, \quad i = 1, \ldots, k. \tag{3}$$

Suppose that alternative i is δ away from alternative 1, i.e., $\mu_{i1} - \mu_{11} > \delta$. Then, alternative i being eliminated by alternative 1 is necessary for the CS event. To that end, we need to ensure $\mu_{ij_i^*} - \mu_{1j_1^*} > \delta_O$, when comparing $\mu_{1j_1^*}$ and $\mu_{ij_i^*}$ in the outer-layer. Notice that

$$\mu_{ij_i^*} - \mu_{1j_1^*} = (\mu_{ij_i^*} - \mu_{i1}) + (\mu_{i1} - \mu_{11}) + (\mu_{11} - \mu_{1j_1^*}) > (-\delta_I) + \delta + 0 = \delta - \delta_I,$$

thanks to (3). Hence, it suffices to take $\delta_O = \delta - \delta_I$.

In order to determine the values of δ_I and δ_O for a given δ, we minimize the total sample size of the procedure over the choice of (δ_I, δ_O). Specifically, we solve the following optimization problem for each pair of systems

$$\begin{aligned} \underset{\delta_I, \delta_O > 0}{\text{minimize}} \quad & \max\left\{ \frac{h^2 S^2_{ij,i'j'}}{\delta_I^2}, \frac{h^2 S^2_{ij,i'j'}}{\delta_O^2} \right\} \\ \text{subject to} \quad & \delta_I + \delta_O = \delta \end{aligned} \tag{4}$$

It is straightforward to solve this problem and the optimal solution is $\delta_I = \delta_O = \delta/2$.

Remark 3. Notice that $\left\lceil h^2 S^2_{ij,i'j'}/\delta_I^2 \right\rceil$ (resp., $\left\lceil h^2 S^2_{ij,i'j'}/\delta_O^2 \right\rceil$) represent the sample size required by the comparison between system (i, j) and system (i', j') in the inner-layer (resp., outer-layer) selection process. Since the inner-layer and outer-layer selection processes are conducted simultaneously after all the samples are generated, we should treat both layers equally and thus assign equal computational budget to them, leading to the optimal choice $\delta_I = \delta_O = \delta/2$.

3.2. Error Allocation

Besides the IZ parameter, another critical parameter that determines the efficiency of the two stage RSB procedure is the error allowance β for each necessary pairwise comparison. It must be chosen judiciously in order that the statistical validity (2) be achieved.

First, we notice that $\mathbb{P}\{\text{CS}\} \geq \mathbb{P}\{\text{system } (1,1) \text{ is selected}\}$, since selecting system $(1, 1)$ is obviously a CS event. Then, the Bonferroni inequality can bound the right-hand-side from below, allowing us to give a PCS guarantee via choosing an appropriate β. Nevertheless, it is worthwhile to point out that due to the double-layer structure, CS can be obtained even if system $(1, 1)$ is eliminated, as long as the selected worst system of alternative 1 is better than any other alternative's selected worst system. This contributes the over-coverage of the realized PCS; see Section 5.

Specifically, system $(1, 1)$ being eliminated by system $(1, j)$ amounts to $\bar{X}_{11}(N) < \bar{X}_{1j}(N)$, $j = 2, \ldots, m$, whereas system $(1, 1)$ being eliminated by system (i, j) amounts to $\bar{X}_{11}(N) > \bar{X}_{ij}(N) =$

$\max_{1\le l\le m} \bar{X}_{il}(N), i = 2,\ldots,k, j = 1,\ldots,m$. Then, it follows that

$$\begin{aligned}
\mathbb{P}\{\text{ICS}\} &\le \mathbb{P}\{\text{system } (1,1) \text{ is not selected}\} \\
&= \mathbb{P}\left\{\bigcup_{j=2}^{m}\left\{\bar{X}_{11}(N) < \bar{X}_{1j}(N)\right\} \bigcup \bigcup_{i=2}^{k}\bigcup_{j=1}^{m}\left\{\bar{X}_{11}(N) > \bar{X}_{ij}(N) = \max_{1\le l\le m}\bar{X}_{il}(N)\right\}\right\} \\
&\le \sum_{j=2}^{m}\mathbb{P}\left\{\bar{X}_{11}(N) < \bar{X}_{1j}(N)\right\} + \sum_{i=2}^{k}\sum_{j=1}^{m}\mathbb{P}\left\{\bar{X}_{11}(N) > \bar{X}_{ij}(N)\right\}, \qquad (5)
\end{aligned}$$

where ICS is short for incorrect selection and the last inequality follows the Bonferroni inequality. Therefore, we can achieve a target PCS $1-\alpha$ by ensuring each of the $km-1$ terms in the summation (5) bounded by $\beta = \alpha/(km-1)$ from above. We name this method of error allocation the *multiplicative rule*.

Nevertheless, the multiplicative rule can easily become over-conservative even if k and m are both moderate. For instance, if $k = m = 10$, then the error allowance under the multiplicative rule is equivalent to that for the SB problem with 100 alternatives. Observe that the over-conservativeness of the multiplicative rule stems from the fact that it uses the event of not selecting system $(1,1)$ to represent the ICS event itself. By doing so, we implicitly treat all the $km-1$ pairwise comparisons as equally important. In fact, we do not need to ensure correct selection of the worst system for each alternative in the inner-layer selection process, except for alternative 1. The bulk of the $km-1$ pairwise comparisons associated with the multiplicative rule turn out to be unnecessary. To see this, notice that

$$\mathbb{P}\{\text{ICS}\} = \mathbb{P}\left\{\bigcup_{i=2}^{k}\left\{\max_{1\le j\le m}\bar{X}_{1j}(N) > \max_{1\le j\le m}\bar{X}_{ij}(N)\right\}\right\} \le \mathbb{P}\{A\cup B\}, \qquad (6)$$

where $A = \bigcup_{i=2}^{k}\left\{\max_j \bar{X}_{1j}(N) > \max_j \bar{X}_{ij}(N)\right\}$ and $B = \bigcup_{j=2}^{m}\left\{\bar{X}_{11}(N) < \bar{X}_{1j}(N)\right\}$, and that

$$A\cup B = (A\cap B^c)\cup B = \bigcup_{i=2}^{k}\left\{\bar{X}_{11}(N) > \max_{1\le j\le m}\bar{X}_{ij}(N)\right\} \cup B, \qquad (7)$$

since $\max_j \bar{X}_{1j}(N) = \bar{X}_{11}(N)$ on B^c. Hence, by (6) and (7),

$$\begin{aligned}
\mathbb{P}\{\text{ICS}\} &\le \mathbb{P}\left\{\bigcup_{i=2}^{k}\left\{\bar{X}_{11}(N) > \max_{1\le j\le m}\bar{X}_{ij}(N)\right\} \bigcup \bigcup_{j=2}^{m}\left\{\bar{X}_{11}(N) < \bar{X}_{1j}(N)\right\}\right\} \\
&\le \mathbb{P}\left\{\bigcup_{i=2}^{k}\left\{\bar{X}_{11}(N) > \bar{X}_{i1}(N)\right\} \bigcup \bigcup_{j=2}^{m}\left\{\bar{X}_{11}(N) < \bar{X}_{1j}(N)\right\}\right\} \\
&\le \sum_{i=2}^{k}\mathbb{P}\left\{\bar{X}_{11}(N) > \bar{X}_{i1}(N)\right\} + \sum_{j=2}^{m}\mathbb{P}\left\{\bar{X}_{11}(N) < \bar{X}_{1j}(N)\right\}. \qquad (8)
\end{aligned}$$

The inequality above implies that there are $k+m-2$ "critical" pairwise comparisons in the RSB problem.

To achieve a target PCS $1-\alpha$, we can simply make each of the $k+m-2$ terms in the summation (8) be no greater than $\beta = \alpha/(k+m-2)$. We name this method of error allocation the *additive rule*. Since its total sample size is increasing in β, the two-stage RSB procedure with the additive rule is significantly more efficient than the one with the multiplicative rule.

Nevertheless, the additive rule is not applicable for the sequential RSB procedure developed in Section 4 and the multiplicative rule will be used there. This is because the additive rule assumes implicitly that the worst system of each alternative is always retained during the inner-layer selection process, while it may be eliminated in early iterations of the sequential procedure; see Remark 5 for more discussion.

3.3. Statistical Validity

We show that the two-stage RSB procedure equipped with the additive rule of error allocation is statistically valid. The case of the multiplicative rule can be proved similarly.

Theorem 1. *Suppose that* $\{X_{ij} : i = 1, 2, \ldots, k, j = 1, 2, \ldots, m\}$ *are jointly normally distributed. Set the error allowance* $\beta = \alpha/(k+m-2)$*. Then, the two-stage RSB procedure is statistically valid, i.e.,* $\mathbb{P}\{\mu_{i^*1} - \mu_{11} \leq \delta\} \geq 1-\alpha$.

The two-stage RSB procedure selects the best alternative based on the means of typically large samples, which can be viewed as approximately normally distributed. To simplify theoretical analysis, we assume that the observations of the systems are jointly normally distributed. Then, Theorem 1 states that the two-stage RSB procedure has finite-sample statistical validity. We relax the normality assumption to allow the simulation outputs to have a general distribution for the sequential RSB procedure in next section at the expense of the finite-sample statistical validity. The sequential RSB procedure is statistically valid only asymptotically as the target PCS goes to 1.

4. Sequential RSB Procedure

Sequential procedures for the SB problem typically require smaller sample sizes than two-stage procedures, because the former allow inferior systems to be eliminated dynamically during iterations (Kim and Nelson 2001). If switching between simulations of different systems does not incur substantial computational overhead, then the overall efficiency of sequential procedures is usually much higher (Hong and Nelson 2005).

Before presenting our sequential RSB procedure, we remark that there is a plain-vanilla sequential procedure for the RSB problem. One can simply apply a sequential SB procedure separately in each layer. Specifically, for each alternative, a sequential SB procedure is applied to select its worst system; it is then applied again to the collection of "worst systems" to select the best among them. However, this procedure has a major drawback: outer-layer eliminations occur only *after* the worst system of each alternative is identified in the inner layer. This incurs excessive samples in the inner layer selection process, since it is unnecessary to identify the worst system for alternatives that are unlikely to be the best. By contrast, our sequential RSB procedures facilitates simultaneous elimination of all the surviving systems of an alternative

when the alternative appears to be inferior with high likelihood. Our sequential RSB procedure is iterative with the following structure.

(i) Take an initial number of samples to estimate the mean of each system and the variance of each pairwise difference.

(ii) Perform the inner-layer selection: for each surviving alternative, eliminate systems that are unlikely to produce the worst-case mean performance.

(iii) Perform the outer-layer selection: eliminate inferior alternatives based on the estimated worst-case mean performance of each surviving alternative.

(iv) If there is only one surviving alternative or all the surviving alternatives are close enough (determined by the IZ parameter) to each other, then stop; otherwise, take one additional sample from each surviving system, update the statistics, and return to step (ii).

For the inner-layer selection in step (ii), we apply the IZ-free sequential SB procedure in Fan et al. (2016), hereafter referred to as the FHN procedure. This procedure does not require an IZ parameter and thus we can set the outer-layer IZ parameter to be the same as the overall IZ parameter. The reason for choosing the FHN procedure instead of other sequential SB procedures that based on the IZ formulation is because it is hard to construct an analytically tractable optimization problem, similar to (4), for the sequential RSB procedure that relates the decomposition of the overall IZ parameter to the efficiency of the procedure. See Section 4.1 for details.

For the outer-layer selection in step (iii), a pairwise comparison between two surviving alternatives is done by constructing a confidence interval that bounds the difference between their worst-case mean performances. The confidence level of this interval depends on the error allowance β. If the confidence interval does not contain zero, then the two competing alternative are differentiated and the inferior one is eliminated (i.e. all the surviving systems of that alternative are eliminated simultaneously). See Section 4.2 for details.

4.1. Inner-Layer: Eliminating Systems

The objective of the inner-layer selection process is to perform sequential screening to eliminate systems that are unlikely to produce the worst-case mean performance of each alternative.

We apply the FHN procedure to the systems of each alternative. In this procedure, the (normalized) partial-sum difference process between two systems is approximated by a Brownian motion with drift. We can then differentiate the two systems by checking if the drift of the Brownian motion is nonzero. This is done by monitoring if the Brownian motion exits a well-designed continuation region, whose boundaries are formed by $\pm g_c(t)$ for $t \geq 0$, where $g_c(t) = \sqrt{[c + \log(t+1)](t+1)}$ for some carefully chosen constant c that depends on the target PCS $1 - \alpha$.

More specifically, consider alternative i and let $\bar{X}_{ij}(n)$ denote the sample mean based on the first n independent replications of system (i, j). Define $t_{ij,ij'}(n) = n\sigma^{-2}_{ij,ij'}$ and $Z_{ij,ij'}(n) = t_{ij,ij'}(n)[\bar{X}_{ij}(n) -$

$\bar{X}_{ij'}(n)]$, where $\sigma^2_{ij,ij'} = \mathrm{Var}[X_{ij} - X_{ij'}]$, for any $1 \le j \ne j' \le m$. Then, $Z_{ij,ij'}(n)$ can be approximated in distribution by a Brownian motion possibly with a nonzero drift. For any pairwise comparison between system (i, j) and system (i, j') with $j \ne j'$, we keep taking samples from them (i.e., increasing n) until $|Z_{ij,ij'}(t_{ij,ij'}(n))| \ge g_c(t_{ij,ij'}(n))$, at which point the one with a smaller estimated mean performance is eliminated by the other since we are seeking the system that has the largest mean performance. Once eliminated, a system will not be considered in any subsequent comparisons.

4.2. Outer-Layer: Eliminating Alternatives

The sequential RSB procedure allows *simultaneous* elimination of all the surviving systems of an alternative. This is achieved in the outer-layer selection process by comparing the estimated worst-case mean performances of the surviving alternatives. Notice that pairwise comparisons in the outer-layer are done for alternatives, instead of systems, and the comparisons are based on random sets of surviving systems of the two alternatives. As a result, sequential procedures for the SB problem are not applicable here.

Consider alternatives i and i' that have survived after n samples of the relevant systems. To design an elimination rule between them, we construct a dynamic confidence interval $(L_{ii'}(n), U_{ii'}(n))$ for $\mu_{i1} - \mu_{i'1}$, the difference between their worst-case mean performances, i.e.,

$$\mathbb{P}\{\mu_{i1} - \mu_{i'1} \in (L_{ii'}(n), U_{ii'}(n)), \text{ for all } n < \infty\} \ge 1 - \epsilon,$$

for a given confidence level $1 - \epsilon$. Hence, if $L_{ii'}(n) > 0$ (resp., $U_{ii'}(n) < 0$), then $\mu_{i1} > \mu_{i'1}$ (resp., $\mu_{i1} < \mu_{i'1}$) with statistical significance and we eliminate alternative i (resp., i'); otherwise, we continue sampling. In Proposition 1, we present an asymptotically valid approach for constructing such a confidence interval.

Proposition 1. *For $i = 1, 2, \ldots, k$, let $\mathcal{S}_i(n)$ denote the set of surviving systems of alternative i after taking n samples of the relevant systems and the subsequent inner-layer elimination. For $\beta \in (0, 1)$, let $g_c(t) = \sqrt{[c + \log(t + 1)](t + 1)}$ with $c = -2\log(2\beta)$. For any two alternatives i and i', define an interval $(L_{ii'}(n), U_{ii'}(n))$ as follows*

$$\begin{aligned} L_{ii'}(n) &= \max_{(i,j)\in\mathcal{S}_i(n)} \bar{X}_{ij}(n) - \max_{(i',j)\in\mathcal{S}_{i'}(n)} \bar{X}_{i'j}(n) - C_i(n) - D_{ii'}(n), \\ U_{ii'}(n) &= \max_{(i,j)\in\mathcal{S}_i(n)} \bar{X}_{ij}(n) - \max_{(i',j)\in\mathcal{S}_{i'}(n)} \bar{X}_{i'j}(n) + C_{i'}(n) + D_{ii'}(n), \end{aligned} \tag{9}$$

where

$$C_i(n) = \max_{(i,j),(i,j')\in\mathcal{S}_i(n)} \frac{g_c(t_{ij,ij'}(n))}{t_{ij,ij'}(n)} \quad \textit{and} \quad D_{ii'}(n) = \max_{(i,j)\in\mathcal{S}_i(n),(i',j')\in\mathcal{S}_{i'}(n)} \frac{g_c(t_{ij,i'j'}(n))}{t_{ij,i'j'}(n)}.$$

If $(i, 1) \in \mathcal{S}_i(n)$ and $(i', 1) \in \mathcal{S}_{i'}(n)$ for all $n \ge 1$, then

$$\limsup_{\beta\to 0} \frac{1}{2\beta} \mathbb{P}\left\{\mu_{i1} - \mu_{i'1} \notin (L_{ii'}(n), U_{ii'}(n)) \text{ for some } n \ge 1\right\} \le 1.$$

Under the IZ formulation, the sequential RSB procedure stops if either of the following conditions holds:

(i) all but one alternatives are eliminated; (ii) all the surviving alternatives are sufficiently close to each other. The latter condition amounts to $C_i(n) + D_{ii'}(n) \leq \delta$ for any pair of surviving alternatives i and i' in the light of (9). The above stopping criterion ensures that the unique best alternative or a good alternative is ultimately selected with certain statistical guarantee.

4.3. The Procedure

We now present the sequential RSB procedure (Procedure S).

Procedure 2 (Procedure S: Sequential RSB Procedure).

0. *Setup.* Specify the error allowance $\beta = \alpha/(km-1)$ and the first-stage sample size $n_0 \geq 2$. Set $c = -2\log(2\beta)$.

1. *Initialization.* Set $n = n_0$. Set $\mathcal{S} = \{1, 2, \ldots, k\}$ to be the set of surviving alternatives. Set $\mathcal{S}_i = \{(i, j) : j = 1, 2, \ldots, m\}$ to be the set of surviving systems of alternative i, $i = 1, \ldots, k$. Take n independent replications $X_{ij,1}, \ldots, X_{ij,n}$ of each system (i, j).

2. *Updating.* Compute the sample mean of each surviving system and the sample variance of the difference between each pair of surviving systems as follows

$$\bar{X}_{ij}(n) = \frac{1}{n}\sum_{r=1}^{n} X_{ij,r}, \quad i \in \mathcal{S},\ (i,j) \in \mathcal{S}_i,$$

$$S^2_{ij,i'j'}(n) = \frac{1}{n-1}\sum_{r=1}^{n}\left[X_{ij,r} - X_{i'j',r} - (\bar{X}_{ij}(n) - \bar{X}_{i'j'}(n))\right]^2, \quad i, i' \in \mathcal{S},\ (i,j) \in \mathcal{S}_i,\ (i',j') \in \mathcal{S}_{i'}.$$

3. *Elimination.* For each $(i, j) \in \mathcal{S}_i$, $(i', j') \in \mathcal{S}_{i'}$ with $i, i' \in \mathcal{S}$ and $i \neq i'$ or $j \neq j'$, compute

$$\tau_{ij,i'j'}(n) = \frac{n}{S^2_{ij,i'j'}(n)} \quad \text{and} \quad Z_{ij,i'j'}(n) = \tau_{ij,i'j'}(n)[\bar{X}_{ij}(n) - \bar{X}_{i'j'}(n)].$$

   3.1 *Inner-layer.* For each $i \in \mathcal{S}$, assign

$$\mathcal{S}_i \leftarrow \mathcal{S}_i \setminus \{(i,j) \in \mathcal{S}_i : Z_{ij,ij'}(n) \leq -g_c(\tau_{ij,ij'}(n)) \text{ for some } (i,j') \in \mathcal{S}_i\}.$$

   3.2 *Outer-layer.* For each $i \in \mathcal{S}$, compute

$$C_i(n) = \max_{(i,j),(i,j') \in \mathcal{S}_i} \frac{g_c(\tau_{ij,ij'}(n))}{\tau_{ij,ij'}(n)};$$

   for any other $i' \in \mathcal{S}$, compute

$$\tau^*_{ii'}(n) = \min_{(i,j) \in \mathcal{S}_i, (i',j') \in \mathcal{S}_{i'}} \tau_{ij,i'j'}(n) \quad \text{and} \quad W_{ii'}(n) = \max_{(i,j) \in \mathcal{S}_i} \bar{X}_{ij}(n) - \max_{(i',j) \in \mathcal{S}_{i'}} \bar{X}_{i'j}(n).$$

Assign

$$\mathcal{S} \leftarrow \mathcal{S} \setminus \{i \in \mathcal{S} : \tau^*_{ii'}(n)[W_{ii'}(n) - C_i(n)] > g_c(\tau^*_{ii'}(n)) \text{ for some } i' \in \mathcal{S}\}.$$

4. *Stopping.* If either $|\mathcal{S}| = 1$ or

$$\tau^*_{ii'}(n)[\delta - C_i(n)] \geq g_c(\tau^*_{ii'}(n)) \text{ and } \tau^*_{ii'}(n)[\delta - C_{i'}(n)] \geq g_c(\tau^*_{ii'}(n)), \quad \text{for all } i, i' \in \mathcal{S},$$

then stop and select $i^* = \arg\min_{i \in \mathcal{S}} \max_{(i,j) \in \mathcal{S}_i} \bar{X}_{ij}(n)$ as the best alternative. Otherwise, take one additional replication of each $(i, j) \in \mathcal{S}_i$ with $i \in \mathcal{S}$, assign $n \leftarrow n + 1$, and return to step 2.

4.4. Asymptotic Statistical Validity

The FHN procedure for the SB problem allows the samples of the competing alternatives to have a general distribution at the expense of the finite-sample statistical validity. We show that the sequential RSB procedure equipped with the multiplicative rule of error allocation is statistically valid in an asymptotic regime in which the targeted PCS level goes to 1 (i.e., $1 - \alpha \to 1$). This regime is adopted by Fan et al. (2016) and dates back to Perng (1969) and Dudewicz (1969).

Theorem 2. *Suppose that $\{X_{ij} : i = 1, 2, \ldots, k, j = 1, 2, \ldots, m\}$ are generally distributed, and that the moment generating function of $\{X_{ij} : i = 1, 2, \ldots, k, j = 1, 2, \ldots, m\}$ is finite in a neighborhood of the origin of $\mathbb{R}^{k \times m}$. Let $n_0(\alpha)$ denote the initial sample size of the sequential RSB procedure as a function of α. Set the error allowance $\beta = \alpha/(km - 1)$. If $n_0(\alpha) \to \infty$ as $\alpha \to 0$, then the sequential RSB procedure is statistically valid asymptotically as $\alpha \to 0$, i.e., $\limsup_{\alpha \to 0} \mathbb{P}\{\mu_{i^*1} - \mu_{11} > \delta\}/\alpha \leq 1$.*

Remark 4. Theorem 2 indicates that with generally distributed simulation outputs, the sequential RSB procedure is statistically valid for all $\alpha > 0$ small enough. The assumption on n_0 is imposed to make sure that the distribution of the sample means converges to the normal distribution as $\alpha \to 0$. This assumption ensures the strong consistency of sequentially updated variance estimators, thereby facilitating asymptotic analysis of the procedure. Nevertheless, numerical experiments show that even with a moderate n_0, the procedure can still deliver a pre-specified PCS.

Remark 5. The additive rule of error allocation does not apply here. In fact, inequality (8), which underpins the additive rule, implicitly assumes that the worst system of each alternative is in contention in the inner-layer selection process. This is certainly the case for the two-stage RSB procedure. By contrast, the sequential RSB procedure has multiple rounds of inner/outer-layer selection, and the worst system of an alternative may be eliminated in an early round. Notice that the dynamic confidence interval (9) hinges on the condition that the worst system of each alternative is never eliminated in the inner-layer. Indeed, if the system having the largest mean performance of an alternative is eliminated in the inner-layer, the remaining systems of the same alternative will yield a smaller estimate of the worst-case mean performance, which makes this alternative less likely be eliminated in the outer-layer. Therefore, in order that the sequential

RSB procedure be statistically valid, the errors associated with all the pairwise comparisons among the systems of each alternative must be controlled, which makes the multiplicative rule necessary.

5. Computational Efficiency

In this section we focus on the procedures' efficiency for a small α through a set of numerical experiments that generalize standard tests for SB procedures.

Suppose that there are $k \times m$ systems, where system (i, j) refers to the pair of alternative i and jth probability scenario in the ambiguity set. Let X_{ij} denote the random performance of system (i, j), for $i = 1, 2, \ldots, k, j = 1, 2, \ldots, m$ and suppose that $(X_{ij} : i = 1, 2, \ldots, k,\ j = 1, 2, \ldots, m)$ are mutually independent normal random variables, $X_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2_{ij})$. Under Assumption 1, the two RSB procedures aim to select alternative 1 upon termination in an attempt to solve $\min_i \max_j \mathbb{E}[X_{ij}]$. In particular, we consider two different configurations of the means that generalize the *slippage configuration* (SC) and the *monotone decreasing means* (MDM) configuration for SB procedures. For SC we use

$$[\mu_{ij}]_{k\times m} = \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0.5 & 0.5 & \ldots & 0.5 \\ \vdots & \vdots & \ddots & \vdots \\ 0.5 & 0.5 & \ldots & 0.5 \end{pmatrix}, \tag{10}$$

and for MDM we use

$$[\mu_{ij}]_{k\times m} = \Big(0.5(i-1) - 0.2(j-1)\Big)_{1\le i\le k, 1\le j\le m} \tag{11}$$

Notice that with the configuration (10), the outer-layer selection process deals with the largest means of each row, i.e., $(0, 0.5, 0.5, \ldots, 0.5)$, which is a SC for the SB problem with IZ formulation. With the configuration (11), the means of each row are monotonically decreasing so the inner-layer selection process for each row corresponds to a SB problem with IZ formulation and a MDM configuration; the outer-layer selection process, on the other hand, deals with $(0, 0.5, \ldots, 0.5(k-1))$, also a monotone configuration of means.

Independently of the means, we further consider three configurations of the variances:

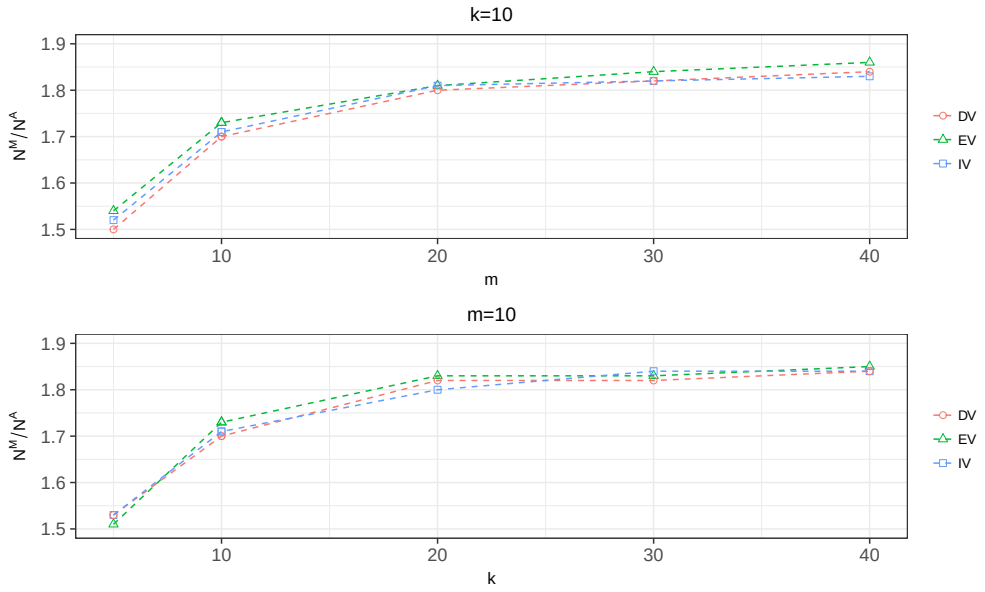
(i) Equal-variance (EV) configuration: $\sigma^2_{ij} = 1$ for all i, j.

(ii) Increasing-variance (IV) configuration: $\sigma^2_{ij} = (1 + 0.1(i-1))(1 + 0.1(j-1))$ for all i, j.

(iii) Decreasing-variance (DV) configuration: $\sigma^2_{ij} = (1 + 0.1(i-1))^{-1}(1 + 0.1(j-1))^{-1}$ for all i, j.

In all the experiments below we set the initial sample size $n_0 = 10$ and the target PCS $1 - \alpha = 0.95$. For each experiment specification (i.e., IZ parameter, values of k and m, configuration of the means and the variances, rule of error allocation, RSB procedure), we repeat the experiment 1000 times independently. We find that the realized PCS is 1.00 for all the cases. It is well known in SB literature that selection procedures that rely on the Bonferroni inequality usually over-deliver PCS (Frazier 2014). Thus, we only report the average sample size of each procedure in this section.

Figure 2: Average Sample Sizes of Procedure T with Different Error Allocation Rules



Note. Top: m varies with $k = 10$; bottom: k varies with $m = 10$. The vertical axis represents the ratio of the average sample size of Procedure T under the multiplicative rule to that under the additive rule.

5.1. Comparison Between Multiplicative Rule and Additive Rule

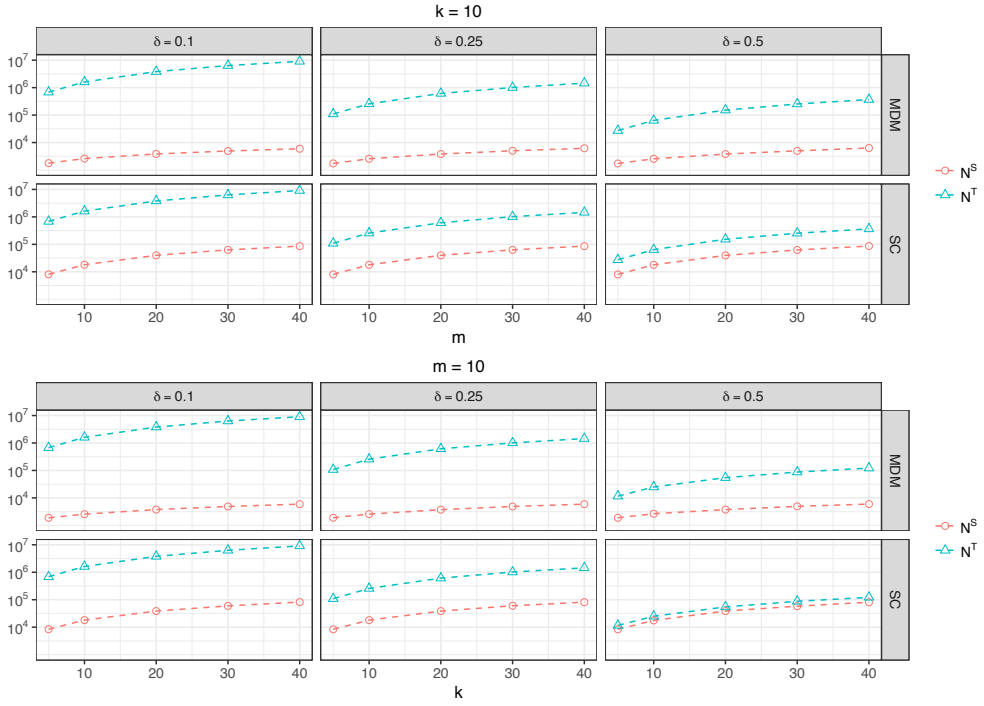
We set the IZ parameter $\delta = 0.5$ and decompose it equally into the inner-layer and outer-layer IZ parameters. Notice that the average sample size of Procedure T does not depend on the means. We study the impact of error allocation on Procedure T's efficiency by varying the problem scale (i.e., k and m) and configuration of the variances. The numerical results are presented in Figure 2.

First, as expected the multiplicative rule requires more samples than the additive rule, regardless of problem scale or configuration of the means/variances. In all the experiments, N^{M} is about $50\% \sim 90\%$ greater than N^{A}, where N^{M} and N^{A} denote the average sample size under the multiplicative and additive rules, respectively.

Second, for a given problem scale, the ratio $N^{\mathrm{M}}/N^{\mathrm{A}}$ is almost independent of configuration of the variances. This is because with an equal split of δ into the inner-layer and the outer-layer, the average sample size is $N \approx 4h^2\delta^{-2} \max_{i,j} \sigma_{ij}^2$, where $h = t_{1-\beta,n_0-1}$ is the $100(1-\beta)\%$ quantile of the student distribution with $n_0 - 1$ degrees of freedom. Hence, for any given configuration of the variances, $N^{\mathrm{M}}/N^{\mathrm{A}} \approx h_{\mathrm{M}}^2/h_{\mathrm{A}}^2$. So essentially the ratio is determined by the value of β, which depends on the rule of error allocation and problem scale for a given target PCS.

Third, for the same reason, given a configuration of the variances the ratio $N^{\mathrm{M}}/N^{\mathrm{A}}$ increases as k or m increases. However, the rate at which this ratio increases is low relative to the increase in the values of k and m. This is caused by the fact that the standard normal quantile grows very slowly towards the tail of the distribution.

Figure 3: Average Sample Sizes of Procedure T and Procedure S Under the EV Configuration



Note. Top: m varies with $k = 10$; Bottom: k varies with $m = 10$. The vertical axis is on a logarithmic scale.

5.2. Comparison Between Procedure T and Procedure S

We have argued in Remark 5 that the additive rule does not apply to Procedure S. We now compare Procedure S under the multiplicative rule with Procedure T under the additive rule, and show that the former is much more efficient.

The two RSB procedures are implemented under different combinations of problem scale, IZ parameter, and configuration of the means/variances. Again, we choose δ from 0.5, 0.25 and 0.1, but only present the numerical results for EV configuration (seen in Figure 3) because the results for the other two configurations of the variances are very similar.

First, Procedure S requires a dramatically fewer samples than Procedure T, seen from the fact that N^{S} is much smaller than N^{T}, where N^{S} and N^{T} denote the average sample sizes of Procedure S and Procedure T, respectively. The difference can be orders of magnitude for large-scale problems with a small IZ parameter; see, e.g., the subplots in the first column of Figure 3. This suggests that there are an enormous number of early eliminations of systems/alternatives in Procedure S.

Second, interestingly, N^{S} is not sensitive to δ when δ is small relative to the difference between the best and the second-best alternatives, i.e., $\delta \leq \mu_{21} - \mu_{11}$. For example, the dashed lines with circles in the three subplots in the first row of Figure 3 are almost identical to each other. Recall that Procedure S terminates if either of the following two conditions is met: (i) there is only one alternative left, and (ii) all the surviving alternatives differ from each other in terms of their worst-case mean performance by no more than δ. The

low sensitivity of N^{S} with respect to δ for small δ suggests that if $\delta \leq \mu_{21} - \mu_{11}$, the procedure terminates primarily because the first stopping condition is met, and thus the procedure mostly selects the unique best alternative rather than a good alternative between several that it cannot differentiate. This feature makes Procedure S particularly attractive in practice. Knowing little of the differences between competing alternatives, a decision-maker tends to choose a small δ to make sure that the unique best alternative is identified. Hence, when using Procedure S, the decision-maker can specify an arbitrarily small δ without any increase in computational cost. By contrast, N^T is roughly proportional to δ^{-2}.

Third, unlike Procedure T, the efficiency of Procedure S does depend on the configuration of the means; see, e.g., the first two subplots in the first column of Figure 3. With everything else the same, SC requires substantially more samples than MDM does. In SC all the systems of each alternative have the same mean performance, and all alternatives but the best one have the same worst-case mean performance. By contrast, in MDM all the systems of each alternative differ from one another by a clear margin and so do the alternatives. Hence, both inner-layer and outer-layer eliminations in Procedure S with SC occur significantly less than with MDM.

Last, it turns out that as the problem scale (i.e., km) increases or as δ decreases, $N^{\mathrm{T}}/N^{\mathrm{S}}$ in general increases. Hence, the advantage of Procedure S in terms of efficiency is more significant for problems with a larger scale or for a smaller IZ parameter; see the numerical results in Section C. Moreover, such an increase in the ratio is more substantial for MDM than for SC. This is a consequence of of the last finding – it is more difficult for Procedure S to conduct eliminations with SC than with MDM.

5.3. Comparison Between Procedure S and Plain-Vanilla Sequential RSB Procedure

We have argued at the beginning of Section 4 that a major advantage of Procedure S relative to a plain-vanilla sequential RSB procedure is that it does not need to identify the worst system of each alternative, and can perform simultaneous elimination of all the surviving systems of an alternative if the alternative is unlikely to be the best, which can save a substantial number of samples.

Notice that virtually any sequential SB procedure can be used to form a plain-vanilla sequential RSB procedure. To ensure a fair comparison against Procedure S whose inner-layer selection is performed by the FHN procedure, we use Procedure 3 in Fan et al. (2016), which is a truncated version of the FHN procedure that has an additional stopping criterion to guarantee termination of the procedure in finite time in the case of two or more alternatives having the same mean performance. We shall call this plain-vanilla sequential RSB procedure Procedure V; see the detailed specification in Section D of Appendix.

Both Procedure S and Procedure V are implemented under different combinations of problem scale, IZ parameter, and configuration of the means/variances. Notice that the difference between the worst-case mean performance of the best and the second-best alternatives is 0.5 in both configurations of the means (10) and (11). We choose δ from 0.5, 0.25, and 0.1. The numerical experiment shows that Procedure S requires significantly fewer samples than Procedure V in general. One exception, however, is that if the configuration of the means is SC and $\delta = \mu_{21} - \mu_{11} = 0.5$, then the computational costs of Procedure S and Procedure V are almost identical. In this case, the alternatives are hard to differentiate in early iterations of Procedure S, so simultaneous elimination of the surviving systems of an alternative that is unlikely to be

the best rarely happens. The numerical results and more discussion are presented in Section D of Appendix.

6. A Multiserver Queue with Abandonment

The theory in this paper is developed under the premise that the ambiguity set is known and fixed. In practice, however, an ambiguity set is typically constructed based on available data, and thus it may vary significantly. Therefore, it is important to study the potential impact of data variation on usefulness of the RSB approach for simulation-based decision-making. Nevertheless, to characterize theoretically such impact is beyond the scope of this paper. Instead, we present an extensive numerical investigation in the context of queueing simulation.

Consider a $G/G/s+G$ model, i.e., a queueing system that has s identical servers and allows a customer to abandon the system before receiving any service if her waiting time in the queue is deemed to exceed her patience. The interarrival times, service times, and patience times are independent and generally distributed. Suppose that both the interarrival time and the patience time have a known distribution, but the distribution of the service time P_0 is unknown. Instead, a finite sample from P_0 is available for constructing an ambiguity set $\mathcal{P}$.

For customer i, $i = 1, \ldots, n$, let I_i and W_i denote her indicator for abandonment and the waiting time, respectively. Let ξ denote the service time. We measure the quality of service by both the probability of abandonment and the mean waiting time of those customers who do not abandon the system. In particular, consider the following cost function

$$f(s,\xi) = c_A U\left(n^{-1}N_A\right) + c_W(n-N_A)^{-1}\sum_{\{i:I_i=0\}} W_i + c_S s, \tag{12}$$

where $N_A = \sum_{i=1}^{n} I_i$ is the (random) number of customers who abandon the system, $U(\cdot)$ is a utility function, and c_A, c_W, and c_S are all positive constants. Let s be the decision variable and assume that it takes values from $\{1, 2, \ldots, k\}$. To minimize the worst-case mean cost over the ambiguity set, we solve $\min_{1\le s\le k}\max_{P\in\mathcal{P}} \mathbb{E}[f(s,\xi)]$, which becomes a RSB problem of the form (1).

Recent empirical studies on service times in various service industries, including telephone call centers (Brown et al. 2005) and health care (Strum et al. 2000), show that the lognormal distribution often fits historical data well. Hence, we assume that P_0 is the lognormal distribution with mean 1; equivalently, $\log(\xi)$ is normally distributed with mean $-\sigma^2/2$ and variance σ^2. The ambiguity set $\mathcal{P}$ is constructed as follows. Upon observing a sample from P_0, we use the maximum likelihood estimation (MLE) to fit three distribution families (lognormal, gamma, and Weibull) to the sample. Then, we conduct the Kolmogorov-Smirnov (K-S) test to each fitted distribution, and include in $\mathcal{P}$ those that are not rejected by the test at significance level 0.05.

Other related parameters are specified as follows. We assume that the interarrival time and the patience time are both exponentially distributed with mean 0.1 and 5, respectively. We set the largest number of servers $k = 10$, the length of each sample path $n = 10,000$, the utility function $U(p) = \log(1/(1-p))$, the constants $c_A = 4$, $c_W = 2$, $c_S = 1$, the first-stage sample size $n_0 = 10$, and the target PCS $1-\alpha = 0.95$.

We vary $\sigma \in \{1, 2, 3\}$, corresponding to different extents of variability of the service time. (The coefficient of variation of ξ is 1.31, 7.32, 90.01, respectively, in the three cases.) We set the sample size ℓ to be either 50 or 500 to imitate the scenarios of having a small and large amount of data to construct the ambiguity set, respectively.

We now assess usefulness of the RSB approach for a decision maker who does not know the input distribution but has input data to work with. The benchmark approach is a typical one in practice: fit a group of distribution families to the data and use the "best fitted" distribution as if it were the true distribution, discarding the others some of which may be plausible as well. We call this approach the best-fitting (BF) approach. We simply define the best fitted distribution as the one having the smallest K-S statistic among all the fitted distributions. Notice that the RSB approach is reduced to the BF approach if the ambiguity set contains only the best fitted distribution.

We first generate 10,000 samples of the cost function $f(s, \xi)$ for each $s = 1, \ldots, k$, with ξ following the true distribution P_0. For each s, we estimate based on the samples the mean and the p-quantile of $f(s, \xi)$ under P_0, denoted by M and Q^p, respectively. They are used as performance measures for evaluating $s_{\rm RSB}$ and $s_{\rm BF}$, the decisions obtained by the RSB and the BF approach, respectively. We also compare the RSB approach with clairvoyance, i.e., the decision maker knows the true distribution and makes the decision $s_{\rm Tr}$ that minimizes $\mathbb{E}[f(s, \xi)]$ under P_0. Let $M_{\rm Tr}$ and $Q^p_{\rm Tr}$ denote the corresponding performance measures. Albeit impractical, clairvoyance provides a lower-bound on the mean cost that other approaches could possibly achieve.

Then, we conduct 1,000 macro-replications of the following experiment.

(i) Generate a sample of service times from P_0.

(ii) Construct an ambiguity set $\mathcal{P}$ based on the sample.

(iii) Run a RSB procedure on $\mathcal{P}$ and $\{\hat{P}_0\}$ to obtain solutions $s_{\rm RSB}$ and $s_{\rm BF}$, respectively.

(iv) Retrieve the performance measures of $s_{\rm RSB}$ (resp., $s_{\rm BF}$) computed earlier under P_0.

(v) Compute the relative difference between $s_{\rm Tr}$ and $s_{\rm RSB}$ and that between $s_{\rm BF}$ and $s_{\rm RSB}$ in terms of the performance measures.

Hence, there are 1,000 realizations of each relative difference above. Their average values over the macro-replications are reported in Table 1 and Table 2. We also estimate the probability that the MLE-fitted lognormal distribution is rejected by the K-S test. It turns out to be 0 based on the 1,000 macro-replications for each case ($\sigma = 1, 2, 3$, $\ell = 50, 500$). Hence, the ambiguity sets constructed in our experiments always include the true distribution family; see also Table 3 for their average size.

We compare the RSB approach with the clairvoyance based on Table 1. First, as expected, the RSB approach yields a higher mean cost than clairvoyance. The right tail quantiles are also higher for the RSB approach. The gap between the performance measures of the two approaches reflects the decision maker's lack of information about the true distribution. Second, with everything else the same, the gap increases as σ increases. This is because a larger σ means a larger stochastic variability in the service times, in which

Table 1: Relative Differences (in %) Between Clairvoyance and RSB

σ	Sample Size	Relative Difference			
		$\frac{M_{\mathrm{Tr}}}{M_{\mathrm{RSB}}}-1$	$\frac{Q^{0.7}_{\mathrm{Tr}}}{Q^{0.7}_{\mathrm{RSB}}}-1$	$\frac{Q^{0.8}_{\mathrm{Tr}}}{Q^{0.8}_{\mathrm{RSB}}}-1$	$\frac{Q^{0.9}_{\mathrm{Tr}}}{Q^{0.9}_{\mathrm{RSB}}}-1$
1	50	$-1.72_{\pm 0.13}$	$-1.71_{\pm 0.13}$	$-1.70_{\pm 0.14}$	$-1.69_{\pm 0.14}$
	500	$-0.40_{\pm 0.05}$	$-0.39_{\pm 0.05}$	$-0.38_{\pm 0.05}$	$-0.38_{\pm 0.05}$
2	50	$-3.55_{\pm 0.32}$	$-3.77_{\pm 0.34}$	$-3.88_{\pm 0.35}$	$-4.07_{\pm 0.38}$
	500	$-0.71_{\pm 0.07}$	$-0.88_{\pm 0.05}$	$-1.00_{\pm 0.08}$	$-1.17_{\pm 0.10}$
3	50	$-7.35_{\pm 0.48}$	$-7.03_{\pm 0.51}$	$-6.75_{\pm 0.54}$	$-6.30_{\pm 0.59}$
	500	$-1.36_{\pm 0.11}$	$-1.21_{\pm 0.12}$	$-1.09_{\pm 0.14}$	$-0.93_{\pm 0.17}$

Note. "$\pm$" indicates 95% confidence interval.

Table 2: Relative Differences (in %) Between BF and RSB

σ	Sample Size	Relative Difference			
		$\frac{M_{\mathrm{BF}}}{M_{\mathrm{RSB}}}-1$	$\frac{Q^{0.7}_{\mathrm{BF}}}{Q^{0.7}_{\mathrm{RSB}}}-1$	$\frac{Q^{0.8}_{\mathrm{BF}}}{Q^{0.8}_{\mathrm{RSB}}}-1$	$\frac{Q^{0.9}_{\mathrm{BF}}}{Q^{0.9}_{\mathrm{RSB}}}-1$
1	50	$0.12_{\pm 0.07}$	$0.14_{\pm 0.08}$	$0.15_{\pm 0.08}$	$0.17_{\pm 0.08}$
	500	$0.00_{\pm 0.04}$	$0.00_{\pm 0.04}$	$0.00_{\pm 0.04}$	$0.00_{\pm 0.04}$
2	50	$1.64_{\pm 0.32}$	$1.76_{\pm 0.33}$	$1.84_{\pm 0.34}$	$1.96_{\pm 0.35}$
	500	$0.02_{\pm 0.04}$	$0.01_{\pm 0.04}$	$0.01_{\pm 0.04}$	$0.01_{\pm 0.05}$
3	50	$5.32_{\pm 0.92}$	$5.66_{\pm 0.93}$	$5.90_{\pm 0.95}$	$6.23_{\pm 0.98}$
	500	$0.07_{\pm 0.12}$	$0.08_{\pm 0.12}$	$0.09_{\pm 0.12}$	$0.11_{\pm 0.12}$

case the decision maker is more uncertain about the true distribution, thereby paying a larger penalty for the deeper uncertainty. Third, in the same vein, the gap decreases as the sample size ℓ increases, since the input uncertainty can be greatly reduced by a large amount of data. Indeed, the relative differences in absolute value are about 1% or even lower when $\ell = 500$.

Assuming clairvoyance is obviously not practical. It is more interesting for practitioners to compare the RSB approach with the BF approach, which is almost common practice for simulation-based decision-making. From Table 2, we first find that the RSB solution outperforms, or performs at least as good as, the BF solution for all the performance measures. Relying on worst-case analysis, the RSB approach is conservative by design and should protect the decision-maker against extreme cases, producing reliable performance even if the true distribution is not in her favor. Therefore, the RSB solution performing better for the right tail quantiles is expected. It is somewhat surprising, however, that the RSB solution performs better for the mean cost as well. This suggests that in the presence of input uncertainty, the potential risk that the BF approach ends up with a misspecified input distribution can be so significant that it overwhelms the price that the decision maker needs to pay for being conservative. Second, the advantage of the RSB approach over the BF approach is larger when the uncertainty about the true distribution is deeper, which means

either the true distribution has a larger stochastic variability (larger σ), or the sample size ℓ is smaller. In particular, with $\sigma = 3$ and $\ell = 50$, the RSB approach outperforms the BF approach by a significant margin ($5\% \sim 6\%$). Third, notice that with $\ell = 500$, the two approaches deliver nearly identical performance. This is because with a large ℓ, the input uncertainty is marginal and the ambiguity set mostly consists of only the best fitted distribution, since the possibility that the fitted gamma or Weibull distribution is not rejected by the K-S test is nearly zero in this situation.

Table 3: Relative Differences Between BF and RSB Conditional on Model Misspecification

σ	ℓ	$\mathbb{E}(\lvert\mathcal{P}\rvert)$	$\mathbb{P}(\text{misspec.})$ (in %)	Rel. Diff. (in %)			
				$\frac{M_{\text{BF}}}{M_{\text{RSB}}} - 1$	$\frac{Q^{0.7}_{\text{BF}}}{Q^{0.7}_{\text{RSB}}} - 1$	$\frac{Q^{0.8}_{\text{BF}}}{Q^{0.8}_{\text{RSB}}} - 1$	$\frac{Q^{0.9}_{\text{BF}}}{Q^{0.9}_{\text{RSB}}} - 1$
1	50	$2.93_{\pm 0.02}$	$17.90_{\pm 2.38}$	$0.39_{\pm 0.26}$	$0.45_{\pm 0.26}$	$0.48_{\pm 0.26}$	$0.53_{\pm 0.26}$
2	50	$2.56_{\pm 0.03}$	$17.40_{\pm 2.35}$	$9.05_{\pm 1.38}$	$9.70_{\pm 1.35}$	$10.14_{\pm 1.34}$	$10.84_{\pm 1.33}$
3	50	$2.27_{\pm 0.03}$	$17.30_{\pm 2.34}$	$31.28_{\pm 3.09}$	$33.22_{\pm 2.88}$	$34.56_{\pm 2.78}$	$36.41_{\pm 2.64}$

$\mathbb{E}(\lvert\mathcal{P}\rvert)$*: the mean size of the ambiguity set.*

We now further discuss the cause of the difference in performance between the RSB approach and the BF approach. When ℓ is not large, the best fitted distribution is likely to be gamma or Weibull instead of lognormal, which we refer to as *model misspecification*. We estimate the probability of model misspecification and compute the relative differences between the two approaches conditional on model misspecification. The results are shown in Table 3. (The estimated probability of model misspecification is 0 for $\sigma = 1, 2, 3$ if $\ell = 500$, so we do not include them in the table.)

Table 3 shows that first, the probability of model misspecification is fairly large ($17\% \sim 18\%$ on average), if ℓ is not large. Second, if the best fitted distribution is misspecified, the consequence for the BF approach can be severe, resulting in a cost that can be over 30% higher than the cost of the RSB solution. Notice that the ambiguity set contains $2 \sim 3$ plausible probability distributions on average. This suggests that in the case of model misspecification, both the true and the incorrectly chosen distributions inform the decision produced by the RSB approach. Third, in the case of model misspecification, the relative difference in performance between the two approaches grows dramatically (from less than 1% to over 30%) as σ increases. Since the estimated probability of model misspecification is roughly the same for different values of σ, the characteristics of the queueing system under the wrongly chosen input model is clearly the cause of the significant degradation in performance of the BF approach as σ grows. Therefore, the RSB approach has a substantial advantage over the BF approach for protecting the decision-maker against model misspecification, especially when the input uncertainty is deep.

7. An Appointment-Scheduling Problem

Appointment-scheduling problems are ubiquitous in the healthcare industry; see Gupta and Denton (2008) for a comprehensive survey. One challenge in addressing these problems in practice is that the appointment duration is often random and its distribution is hard to estimate due to lack of the data. In the light of the

distributional uncertainty, robust optimization has recently emerged as a popular framework for solving this class of problems (Kong et al. 2013, Mak et al. 2015, Qi 2017). In this section, we apply our RSB approach to study an appointment-scheduling problem with real data and compare it with existing approaches.

We consider the appointment-scheduling problem in Mak et al. (2015). There are n operations to be performed by n different surgeons in an operating room during a time interval $[0, T]$. By operation i, we mean the operation performed by surgeon i, $i = 1, \ldots, n$. Operation i requires a random duration d_i to complete, which follows distribution P_i. Let $\mathcal{P}_i$ denote the ambiguity set for P_i. We assume that P_i's are mutually independent, so the ambiguity set for the joint distribution $\mathbf{P} = (P_1, \ldots, P_n)$ can be expressed as the Cartesian product of $(\mathcal{P}_1, \ldots, \mathcal{P}_n)$, i.e., $\mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$.

Let ψ be a permutation of $\{1, 2, \ldots, n\}$ that indicates the sequence of the operations performed in the operating room. Let t_i denote the time allowance of operation i. The planner makes two decisions: the sequence of the operations ψ and the time allowance of each operation $\mathbf{t} = (t_1, \ldots, t_n)$.

Suppose that all the appointments are *scheduled* to be completed by time T, so the feasible region of $\mathbf{t}$ is $\mathcal{T} = \{\mathbf{t} \in \mathbb{R}^n_+ : \sum_{i=1}^n t_i \leq T\}$. If an operation does not start as planned because of delay of completion of the previous operation, a waiting cost is incurred at the rate of c_W; if the last operation is completed after time T, an overtime cost is incurred at the rate of c_O. Let W_i denote the waiting time of operation ψ_i, $i = 1, \ldots, n$, and W_{n+1} denote the overtime. Then, $W_1 = 0$ and $W_i = \max\{0, W_{i-1} + d_{\psi_{i-1}} - t_{\psi_{i-1}}\}$, $i = 2, \ldots, n+1$. Hence, letting $\mathbf{d} = (d_1, \ldots, d_n)$, the total waiting and overtime cost as a function of $(\psi, \mathbf{d}, \mathbf{t})$ is $f(\psi, \mathbf{d}, \mathbf{t}) = c_W \sum_{i=1}^n W_i + c_O W_{n+1}$. To minimize the worst-case mean cost over the ambiguity set, the planner solves

$$\min_{\psi} \min_{\mathbf{t} \in \mathcal{T}} \max_{\mathbf{P} \in \mathcal{P}} \mathbb{E}[f(\psi, \mathbf{d}, \mathbf{t})]. \tag{13}$$

In Mak et al. (2015) each ambiguity set $\mathcal{P}_i$ is the set of distributions having the same first two moments as those of P_i, which are assumed to be known. Then, problem (13) for a given sequence ψ can be reformulated as a second-order conic program that has an analytical solution. Specifically, the optimal time allowance is of the form $\tilde{t}^*_{\psi_i} = \mu_{\psi_i} + \eta_{\psi_i} \sigma_{\psi_i}$, where μ_i and σ_i are the mean and standard deviation of P_i, and η_i is a constant that depends on $\{i, \psi, (\mu_1, \sigma_1), \ldots, (\mu_n, \sigma_n)\}$ and can be computed analytically. Theorem 3 in their paper further shows that the optimal sequence $\tilde{\psi}^*$ follows the increasing order of variances (OV), i.e., $\sigma_{\tilde{\psi}^*_1} \leq \cdots \leq \sigma_{\tilde{\psi}^*_n}$.

However, their approach does not apply in our setting where we formulate $\mathcal{P}_i$ as a finite set of plausible probability distributions. In this study, we use the same rule of determining the time allowances as Mak et al. (2015) for a given sequence, and focus on the optimal sequencing problem. The scheduling problem (13) is then reduced to

$$\min_{\psi} \max_{\mathbf{P} \in \mathcal{P}} \mathbb{E}[f(\psi, \mathbf{d}, \tilde{\mathbf{t}}^*_{\psi})], \tag{14}$$

which becomes a RSB problem of the form (1), where $\tilde{\mathbf{t}}^*_{\psi} = (\tilde{t}^*_{\psi_1}, \ldots, \tilde{t}^*_{\psi_n})$.

We now solve (14) with the ambiguity set $\mathcal{P}$ constructed based on real data from a hospital in Anhui Province, China. This hospital is a major healthcare facility in the province. [1] We use the data on durations

[1] This hospital has around 2,800 beds. According to Becker's Hospital Review, the largest hospital in the United States has around 2,400 beds as of 2015.

of cesarean sections in the hospital that were performed in 2014. We fix $n = 4$ as a representative scenario in the hospital, so the number of competing alternatives is $4! = 24$.

We consider two datasets: one is relatively large, denoted by $\mathcal{D}^L$, and consists of 4 obstetricians (OBs) having 138, 97, 84, and 68 cases in the record, respectively; the other is relatively small, denoted by $\mathcal{D}^S$, and consists of another 4 OBs having 66, 60, 55, and 54 cases, respectively.

Given a dataset $\mathcal{D}$ (either $\mathcal{D}^L$ or $\mathcal{D}^S$), let $\mathcal{D}_i$ denote the observations of d_i for OB i, $i = 1, \ldots, 4$. Since the distribution that generates the real data is unknown, we assume that the "true" distribution of d_i is the empirical distribution based on $\mathcal{D}_i$. Let μ_i and σ_i^2 denote the mean and variance of this distribution. We stress here that the RSB approach as well as other approaches introduced later do not have access to (μ_i, σ_i^2). Instead, they only have access to a random sample $\mathcal{F}_i$ from $\mathcal{D}_i$.

For comparison, we also apply the following approaches to find a proper sequence of OBs.

- *Best-Fitting* (BF): For $i = 1, \ldots, n$, let $\hat{P}_i$ be the best fitted distribution for $\mathcal{F}_i$. Solve (14) with $\mathcal{P}_i = \{\hat{P}_i\}$ to obtain a sequence ψ_{BF}.
- *Empirical* (Em): For each $i = 1, \ldots, n$, let $\hat{P}_i$ be the empirical distribution based on $\mathcal{F}_i$. Solve (14) with $\mathcal{P}_i = \{\hat{P}_i\}$ to obtain a sequence ψ_{Em}.
- *OV*: Sort the OBs in the increasing order of $\hat{\sigma}_i^2$ to obtain a sequence ψ_{OV}.

We conduct 1,000 macro-replications of the following experiment.

(i) Randomly select a fraction $\gamma \in (0, 1)$ of $\mathcal{D}_i$, denoted by $\mathcal{F}_i$.

(ii) Construct an ambiguity set $\mathcal{P}_i$ by using MLE to fit six widely used distributions (exponential, gamma, Weibull, lognormal, Pareto, and triangular) to $\mathcal{F}_i$, and retaining the fitted distributions that are not rejected by the K-S test at significance level 0.05.

(iii) Compute the mean and variance of $\mathcal{F}_i$, denoted by $(\hat{\mu}_i, \hat{\sigma}_i^2)$. Apply Theorem 2 of Mak et al. (2015) to compute the time allowances $\tilde{\mathbf{t}}_\psi^*$ with $(\hat{\mu}_i, \hat{\sigma}_i^2)$, $i = 1, \ldots, 4$ for each sequence ψ.

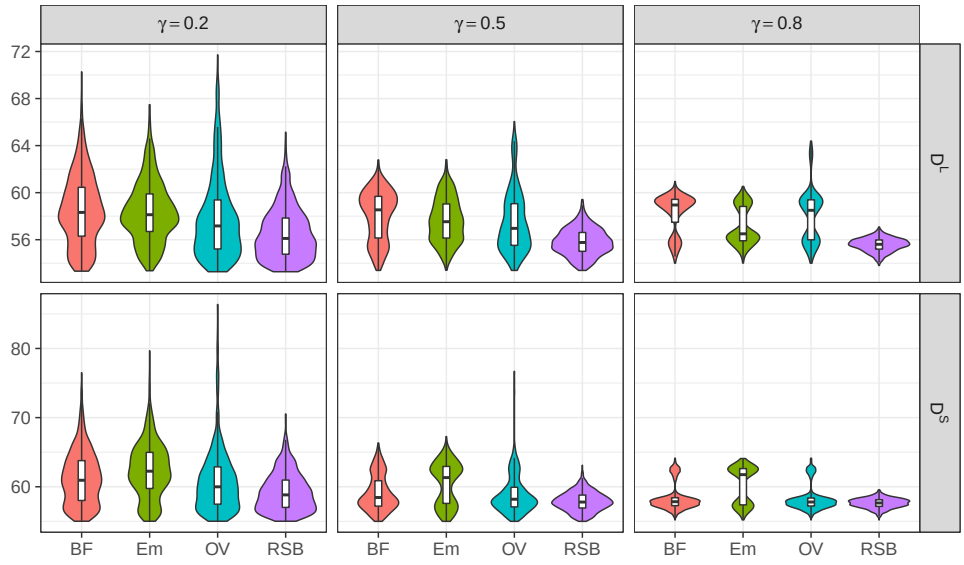
(iv) Run the four competing approaches to solve (14).

(v) For each $\psi = \psi_{\mathrm{RSB}}, \psi_{\mathrm{BF}}, \psi_{\mathrm{Em}}, \psi_{\mathrm{OV}}$, generate 10^7 samples of the cost function $f(\psi, \mathbf{d}, \tilde{\mathbf{t}}_\psi^*)$ under the "true" distribution of d_i. Compute M and Q^p, $p = 0.7, 0.8, 0.9$, based on the samples.

The other parameters involved in our experiment are specified as follows: $c_W = 1.0$, $c_O = 0.5$, $T = \sum_{i=1}^4 \mu_i$, the first-stage sample size $n_0 = 10$, the IZ parameter $\delta = 1.0$, and the target PCS $1 - \alpha = 0.95$. Moreover, we vary the fraction $\gamma \in \{0.2, 0.5, 0.8\}$ to imitate the scenarios of having a small, medium, and large amount of available data to construct the ambiguity set, respectively.

Remark 6. We find that $|\mathcal{P}_i|$ is typically 3 or 4, so the size of the ambiguity set $m = |\mathcal{P}| = \prod_{i=1}^4 |\mathcal{P}_i|$ is fairly large, which typically ranges from 100 to 200 depending on $\mathcal{D}$ and γ.

Figure 4: Distribution of Realizations of M



Note. The violin plots depict the distributions of the presented data constructed via kernel density estimation; the box plots indicate the median, 25% quantile, and 75% quantile.

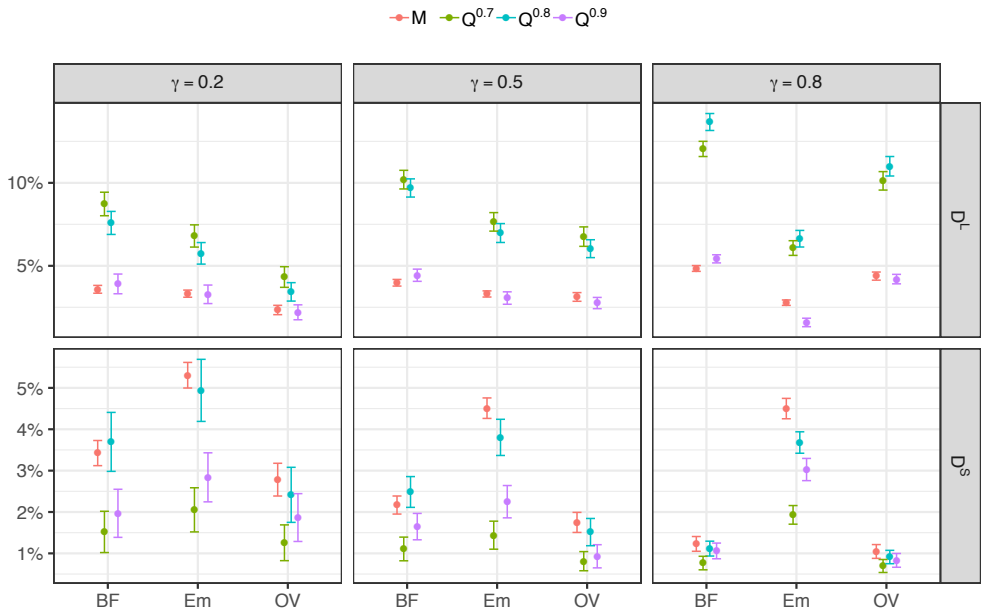
We obtain 1,000 realizations of M and Q^p, one from each macro-replication. Figure 4 illustrates the distribution of M for different values of $(\psi, \gamma, \mathcal{D})$. We omit the figures for $Q^{0.7}$, $Q^{0.8}$, and $Q^{0.9}$ since they are qualitatively similar to Figure 4. In addition, we compute the relative difference in M or Q^p between a competing approach and the RSB approach: $\frac{M_A}{M_{\text{RSB}}} - 1$ and $\frac{Q^p_A}{Q^p_{\text{RSB}}} - 1$, where the subscript A denotes the competing approach, i.e., BF, Em, or OV, for $p = 0.7, 0.8, 0.9$. Their average over the macro-replications is shown in Figure 5.

We have several findings. First, the RSB approach consistently outperforms the other three approaches by a clear margin, producing the lowest mean cost and the lowest right tail quantiles.

Second, Figure 4 also shows as γ or equivalently the sample size increases, the range of the realizations of the mean cost for each approach obviously decreases. This is because the larger sample size is, the more information it provides about the "true" distribution. In addition, as input data varies, the mean cost of the RSB solution is the most stable, indicated by the range of its realizations being considerably narrower than the others. Notice that the sample size for constructing an ambiguity set in this example is not large, ranging from 11 to 110 depending on $\mathcal{D}$ and γ, and therefore the input uncertainty is substantial. In particular, the sample $\mathcal{F}_i$ may be far from being representative of the "true" distribution, i.e., the whole dataset $\mathcal{D}_i$. In this case, both the BF approach and the Empirical approach suffer from using a misspecified distribution for decision-making, while the RSB approach can effectively protect the planner against such a risk.

Third, the OV approach is well known to be a good heuristic rule in appointment-scheduling literature. It is essentially built upon worst-case analysis as well. Indeed, Mak et al. (2015) proves that the OV sequence is optimal for the robust optimization problem (13), provided that the ambiguity set $\mathcal{P} = \{(P_1, \ldots, P_n) :$

Figure 5: Relative Differences Between Competing Methods and RSB



Note. M means $\frac{M_A}{M_{\text{RSB}}} - 1$ and Q^p means $\frac{Q^p_A}{Q^p_{\text{RSB}}} - 1$, for $A = \text{BF}, \text{Em}, \text{OV}$, and $p = 0.7, 0.8, 0.9$. The dots denote the average over the 1,000 macro-replications, whereas the error bars denote the 95% confidence intervals.

$\mathbb{E}(P_i) = \hat{\mu}_i, \text{Var}(P_i) = \hat{\sigma}_i^2, i = 1, \ldots, n\}$. Therefore, the OV approach also provides protection against input uncertainty, outperforming both the BF approach and the Empirical approach in all cases expect $(\mathcal{D}, \gamma) = (\mathcal{D}^L, 0.8)$; see Figure 5. Nevertheless, its performance is significantly worse than the RSB approach. (The relative difference in $Q^{0.7}$ and $Q^{0.8}$ can be over 10%.) This is because (i) the OV approach uses only the first two moments of the data, whereas the RSB approach uses more information by fitting various distribution families to the data; and (ii) the moments are estimated from data and the estimation error can be substantial if the sample size is not large, whereas fitting distributions to data appears to be more resilient to such error.

In Kong et al. (2016) several artificial examples of the scheduling problem are constructed for which the OV sequence is not optimal. They assume that the input distribution is known and find that the performance gap between the OV sequence and the optimal sequence is marginal. Our experiments, however, demonstrate that such performance gap may be sizable due to input uncertainty. In the light of the fact that lack of data is a common challenge for appointment-scheduling in health care (Macario 2010), our findings are practically meaningful, as they may encourage hospital administrators to adopt conservative formulations such as the RSB approach that do not assume mean and variance are correctly estimated from tens of data points, and to follow more established statistical procedures to accept or reject representative distributions.

8. Concluding Remarks

We propose to use worst-case analysis to address the SB problem with input uncertainty. Two selection procedures are developed to solve the resulting RSB problem and are shown to be statistically valid in the finite-sample and asymptotic regimes, respectively. The two-stage RSB procedure is simple but rather conservative, requiring an excessive number of samples. The sequential RSB procedure, on the other hand, is not as easy to implement, but requires a dramatically smaller sample size even than the plain-vanilla sequential RSB procedure.

As shown by both the queueing example in Section 6 and the scheduling example in Section 7, not only can the RSB approach generate decisions that are reliable in extreme cases, but also perform better than the best-fitting approach on average. This makes the RSB approach a useful tool when the simulation model suffers from deep input uncertainty. In practice, given limited input data one may consider to apply both the RSB approach and the best-fitting (BF) approach to solve the SB problem in the presence deep input uncertainty, and use the latter as a numerical check to ensure that the former works as intended. Another possible way of checking the robustness of the RSB approach in practice is to numerically evaluate and compare the mean performance of the RSB decision under several plausible input distributions included in the ambiguity set.

This paper focuses on the uncertainty in specifying the parametric family of the input distribution, which motivates the critical assumption that the ambiguity set consists of finitely many distributions. This form of ambiguity set entails the double-layer structure of the two selection procedures. However, there are a variety of other ways to construct an ambiguity set, e.g., via moment constraints (Delage and Ye 2010) or via statistical divergence (Ben-Tal et al. 2013). Conceivably, changing the form of the ambiguity set would dramatically alter the structure of the RSB problem, and new selection procedures would need to be developed.

Parameter uncertainty is an equally important issue. Ideally, it should be addressed in conjunction with the uncertainty about the parametric family. One possible approach is to use likelihood ratio to characterize parameter uncertainty. Then, the worst-case mean performance over the uncertain parameters within the same family can be estimated by simulation under a "nominal" distribution only; see Hu and Hong (2015) for details. Then, the ambiguity set that characterizes the uncertainty about both the parametric family and its parameters can be reduced to a finite set. We leave the exploration of these questions to future research.

A. Proof for the Two-Stage RSB Procedure

Proof of Theorem 1. We first notice that if $\mu_{k1} - \mu_{11} \leq \delta$, then by Assumption 1 $\mu_{i1} - \mu_{11} \leq \delta$ for all $i = 1, \ldots, k$, which implies that $\mathbb{P}(\mu_{i^*1} - \mu_{11} \leq \delta) = 1$ and the theorem trivially holds. Hence, without loss of generality we assume that there exists $l = 1, \ldots, k-1$ for which $\mu_{l1} - \mu_{11} \leq \delta$ and $\mu_{l+1,1} - \mu_{11} > \delta$. Then, a good selection (i.e., $\{\mu_{i^*1} - \mu_{11} \leq \delta\}$) occurs if any alternative i, $i = l+1, \ldots, k$ is not selected. It

follows that

$$
\begin{aligned}
& \mathbb{P}\{\mu_{i^*1} - \mu_{11} \leq \delta\} \\
\geq\ & \mathbb{P}\left\{ \bigcap_{i=l+1}^{k} \left\{ \max_{1\leq j\leq m} \bar{X}_{ij}(N) > \max_{1\leq j\leq m} \bar{X}_{1j}(N) \right\} \right\} \\
\geq\ & \mathbb{P}\left\{ \bigcap_{i=l+1}^{k} \left\{ \max_{1\leq j\leq m} \bar{X}_{ij}(N) > \max_{1\leq j\leq m} \bar{X}_{1j}(N) \right\} \bigcap \left\{ \max_{1\leq j\leq m} \bar{X}_{1j}(N) - \bar{X}_{11}(N) < \delta_I \right\} \right\} \\
\geq\ & \mathbb{P}\left\{ \bigcap_{i=l+1}^{k} \left\{ \bar{X}_{i1}(N) > \max_{1\leq j\leq m} \bar{X}_{1j}(N) \right\} \bigcap \left\{ \max_{1\leq j\leq m} \bar{X}_{1j}(N) - \bar{X}_{11}(N) < \delta_I \right\} \right\} \\
\geq\ & \mathbb{P}\left\{ \bigcap_{i=l+1}^{k} \left\{ \bar{X}_{i1}(N) > \bar{X}_{11}(N) + \delta_I \right\} \bigcap \bigcap_{j=2}^{m} \left\{ \bar{X}_{1j}(N) - \bar{X}_{11}(N) < \delta_I \right\} \right\} \\
\geq\ & 1 - \sum_{i=l+1}^{k} \mathbb{P}\left\{ \bar{X}_{i1}(N) \leq \bar{X}_{11}(N) + \delta_I \right\} - \sum_{j=2}^{m} \mathbb{P}\left\{ \bar{X}_{1j}(N) \geq \bar{X}_{11}(N) + \delta_I \right\}, \qquad (15)
\end{aligned}
$$

where the last step is due to the Bonferroni inequality. For each $i = l+1, \ldots, k$,

$$
\begin{aligned}
\mathbb{P}\{\bar{X}_{i1}(N) \leq \bar{X}_{11}(N) + \delta_I\} &= \mathbb{P}\left\{ \bar{X}_{i1}(N) - \bar{X}_{11}(N) - (\mu_{i1} - \mu_{11}) \leq -\mu_{i1} + \mu_{11} + \delta_I \right\} \\
&\leq \mathbb{P}\left\{ \bar{X}_{i1}(N) - \bar{X}_{11}(N) - (\mu_{i1} - \mu_{11}) \leq -\delta_O \right\} \\
&= \mathbb{P}\left\{ \frac{\bar{X}_{i1}(N) - \bar{X}_{11}(N) - (\mu_{i1} - \mu_{11})}{\sqrt{S_{11,i1}^2/N}} \leq -\frac{\delta_O}{\sqrt{S_{11,i1}^2/N}} \right\} \\
&\leq \mathbb{P}\left\{ \frac{\bar{X}_{i1}(N) - \bar{X}_{11}(N) - (\mu_{i1} - \mu_{11})}{\sqrt{S_{11,i1}^2/N}} \leq -h \right\}, \qquad (16)
\end{aligned}
$$

where the first inequality holds because $\delta_I + \delta_O = \delta$ and $\mu_{i1} - \mu_{11} > \delta$ for each $i = l+1, \ldots, k$ under Assumption 1, and the second inequality holds because $N \geq h^2 S_{11,i1}^2 / \delta_O^2$.

For notational simplicity, we suppress its dependence on i and set $Y_r = X_{i1,r} - X_{11,r} - (\mu_{i1} - \mu_{11})$ for $r = 1, \ldots, N$, $\sigma_Y^2 = \mathrm{Var}[Y_r]$, and $S_Y^2 = S_{11,i1}^2$. Applying (16) and following a similar derivation in Stein (1945),

$$
\mathbb{P}\{\bar{X}_{i1}(N) \leq \bar{X}_{11}(N) + \delta_I\} \leq \mathbb{P}\left\{ \frac{\sum_{r=1}^{N} Y_r}{\sqrt{N S_Y^2}} \leq -h \right\} = \mathbb{P}\{Z \leq -h\} = \beta, \qquad (17)
$$

where Z has the distribution of Student's t with $n_0 - 1$ degrees of freedom, and the second inequality holds due to the definition of h. Likewise, we can show that

$$
\mathbb{P}\left\{ \bar{X}_{1j}(N) \geq \bar{X}_{11}(N) + \delta_I \right\} \leq \beta, \qquad (18)
$$

for each $j = 2, \ldots, m$. Combining (15), (17), and (18), we have

$$\begin{aligned}\mathbb{P}\{\mu_{i^*1} - \mu_{11} > \delta\} &\leq \sum_{i=2}^{k} \mathbb{P}\left\{\bar{X}_{i1}(N) \leq \bar{X}_{11}(N) + \delta_I\right\} + \sum_{j=2}^{m} \mathbb{P}\left\{\bar{X}_{1j}(N) \geq \bar{X}_{11}(N) + \delta_I\right\} \\ &\leq (k+m-2)\beta = \alpha,\end{aligned}$$

which completes the proof. □

B. Proofs for the Sequential RSB Procedure

We prove Proposition 1 and Theorem 2 in this section. To facilitate their proofs, we first present a result that characterizes the first-exit probability that a random walk exits from the region $(-g_c(t), g_c(t))$.

Lemma 1. *Let $\{Y_n : i = n = 1, 2, \ldots\}$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean 0 and variance $\sigma^2 < \infty$. Let $\bar{Y}(n)$ and $S^2(n)$ denote the sample mean and the sample variance of $(Y_i : i = 1, \ldots, n)$, respectively. Let $g_c(t) = \sqrt{[c + \log(t+1)](t+1)}$ with $c = -2\log(2\beta)$ for some $\beta \in (0, 1)$. Assume that the moment generating function of Y_1 is finite in a neighborhood of zero.*

(i) Define $t(n) = n/\sigma^2$ and $N = \min\{n \geq n_0 : t(n)|\bar{Y}(n)| \geq g_c(t(n))\}$. If $n_0 \to \infty$ as $\beta \to 0$, then,

$$\limsup_{\beta \to 0} \frac{1}{\beta} \mathbb{P}\left\{t(N)\bar{Y}(N) \leq -g_c(t(N)), N < \infty\right\} \leq 1.$$

(ii) Define $\tau(n) = n/S^2(n)$ and $N' = \min\{n \geq n_0 : \tau(n)|\bar{Y}(n)| \geq g_c(\tau(n))\}$. If $n_0 \to \infty$ as $\beta \to 0$, then,

$$\limsup_{\beta \to 0} \frac{1}{\beta} \mathbb{P}\left\{\tau(N')\bar{Y}(N') \leq -g_c(\tau(N')), N' < \infty\right\} \leq 1.$$

Proof of Lemma 1. We provide a proof sketch here. The complete proof can be found in the proof for Theorem 2 in Fan et al. (2016). Let $(B(t) : t \geq 0)$ be a standard Brownian motion. By virtue of the functional central limit theorem (Whitt 2002, p.102), it can be shown that

$$\mathbb{P}\left\{t(N)\bar{Y}(N) \leq -g_c(t(N)), N < \infty\right\} \leq \mathbb{P}\left\{B(T_c) \leq -g_c(T_c), T_c < \infty\right\},$$

where $T_c = \inf\{t \geq 0 : |B(t)| \geq g_c(t)\}$. Moreover, Example 6 in Jennen and Lerche (1981) shows that

$$\mathbb{P}\left\{B(T_c) \leq -g_c(T_c), T_c < \infty\right\} = \frac{1}{2}e^{-c/2} = \beta,$$

and thus (i) follows immediately. On the other hand, note that

$$\limsup_{\beta \to 0} \frac{1}{\beta} \mathbb{P}\left\{\tau(N')\bar{Y}(N') \leq -g_c(\tau(N')), N' < \infty\right\} \leq \limsup_{\beta \to 0} \frac{1}{\beta} \mathbb{P}\left\{t(N)\bar{Y}(N) \leq -g_c(t(N)), N < \infty\right\},$$

which proves (ii). □

B.1. Proof of Proposition 1

Proof of Proposition 1. If $(i,1) \in \mathcal{S}_i(n)$, then

$$\begin{aligned} U_{ii'}(n) &= \max_{(i,j)\in\mathcal{S}_i(n)} \bar{X}_{ij}(n) - \max_{(i',j)\in\mathcal{S}_{i'}(n)} \bar{X}_{i'j}(n) + C_{i'}(n) + D_{ii'}(n) \\ &\geq \bar{X}_{i1}(n) - \max_{(i',j)\in\mathcal{S}_{i'}(n)} \bar{X}_{i'j}(n) + C_{i'}(n) + \frac{g_c(t_{i1,i'1}(n))}{t_{i1,i'1}(n)}. \end{aligned} \tag{19}$$

Let $(i', j^*_{i'})$ denote the system having the largest mean performance among $\mathcal{S}_{i'}(n)$. If $(i',1) \in \mathcal{S}_{i'}(n)$, then system $(i',1)$ has not been eliminated by $(i', j^*_{i'})$. According to the mechanism of the inner-layer elimination and its related discussion in Section 4.1,

$$\max_{(i',j)\in\mathcal{S}_{i'}(n)} \bar{X}_{i'j}(n) - \bar{X}_{i'1}(n) = \bar{X}_{i'j^*_{i'}}(n) - \bar{X}_{i'1}(n) < \frac{g_c(t_{i'1,i'j^*_{i'}}(n))}{t_{i'1,i'j^*_{i'}}(n)} \leq C_{i'}(n). \tag{20}$$

Plugging (20) into (19) yields

$$U_{ii'}(n) \geq \bar{X}_{i1}(n) - \bar{X}_{i'1}(n) + \frac{g_c(t_{i1,i'1}(n))}{t_{i1,i'1}(n)}. \tag{21}$$

Likewise, we can show that

$$L_{ii'}(n) \leq \bar{X}_{i1}(n) - \bar{X}_{i'1}(n) - \frac{g_c(t_{i1,i'1}(n))}{t_{i1,i'1}(n)}. \tag{22}$$

By (21) and (22), if $(i,1) \in \mathcal{S}_i(n)$ and $(i',1) \in \mathcal{S}_{i'}(n)$ for all $n \geq 1$, then

$$\begin{aligned} & \mathbb{P}\left\{\mu_{i1} - \mu_{i'1} \notin (L_{ii'}(n), U_{ii'}(n)) \text{ for some } n \geq 1\right\} \\ \leq\ & \mathbb{P}\left\{\mu_{i1} - \mu_{i'1} \notin \left(\bar{X}_{i1}(n) - \bar{X}_{i'1}(n) - \frac{g_c(t_{i1,i'1}(n))}{t_{i1,i'1}(n)}, \bar{X}_{i1}(n) - \bar{X}_{i'1}(n) + \frac{g_c(t_{i1,i'1}(n))}{t_{i1,i'1}(n)}\right) \text{ for some } n \geq 1\right\} \\ =\ & \mathbb{P}\left\{t_{i1,i'1}(n)|\bar{X}_{i1}(n) - \bar{X}_{i'1}(n) - (\mu_{i1} - \mu_{i'1})| \geq g_c(t_{i1,i'1}(n)) \text{ for some } n \geq 1\right\} \\ =\ & \mathbb{P}\left\{N_{i1,i'1} < \infty\right\}, \end{aligned} \tag{23}$$

where $N_{i1,i'1} = \min\{n \geq 1 : t_{i1,i'1}(n)|\bar{X}_{i1}(n) - \bar{X}_{i'1}(n) - (\mu_{i1} - \mu_{i'1})| \geq g_c(t_{i1,i'1}(n))\}$. It follows from Lemma 1(i) that, letting $\bar{Y}_{ii'}(n) = \bar{X}_{i1}(n) - \bar{X}_{i'1}(n) - (\mu_{i1} - \mu_{i'1})$,

$$\limsup_{\beta\to 0} \frac{1}{\beta}\mathbb{P}\left\{t_{i1,i'1}(n)\bar{Y}_{ii'}(n) \leq -g_c(t_{i1,i'1}(n)),\ N_{i1,i'1} < \infty\right\} \leq 1.$$

By the symmetry of the random walk paths,

$$\limsup_{\beta\to 0} \frac{1}{\beta}\mathbb{P}\left\{t_{i1,i'1}(n)\bar{Y}_{ii'}(n) \geq g_c(t_{i1,i'1}(n)),\ N_{i1,i'1} < \infty\right\} \leq 1.$$

Hence, in the light of (23),

$$\begin{aligned}
& \limsup_{\beta\to 0}\frac{1}{\beta}\mathbb{P}\left\{\mu_{i1}-\mu_{i'1}\notin(L_{ii'}(n),U_{ii'}(n)) \text{ for some } n\geq 1\right\}\\
\leq\ & \limsup_{\beta\to 0}\frac{1}{\beta}\mathbb{P}\left\{N_{i1,i'1}<\infty\right\}\\
\leq\ & \limsup_{\beta\to 0}\frac{1}{\beta}\mathbb{P}\left\{t_{i1,i'1}(N_{i1,i'1})\bar{Y}_{ii'}(N_{i1,i'1})\leq -g_c(t_{i1,i'1}(N_{i1,i'1})),\ N_{i1,i'1}<\infty\right\}\\
& +\limsup_{\beta\to 0}\frac{1}{\beta}\mathbb{P}\left\{t_{i1,i'1}(N_{i1,i'1})\bar{Y}_{ii'}(N_{i1,i'1})\geq g_c(t_{i1,i'1}(N_{i1,i'1})),\ N_{i1,i'1}<\infty\right\}\\
\leq\ & 2,
\end{aligned}$$

which completes the proof. □

B.2. Proof of Theorem 2

In order to prove Theorem 2, we characterize various scenarios that can lead to an incorrect selection (ICS) event that alternative 1 is not ultimately selected. One such scenario is that in step 3.2, i.e., outer-layer elimination, alternative 1 may be eliminated because the approximate dynamic confidence interval for $\mu_{11}-\mu_{i1}$, which is constructed in the spirit of Proposition 1, is entirely to the right of the origin. In this scenario, we say "alternative 1 is *eliminated* by alternative i".

The other possible scenario for ICS is that in step 4, i.e., stopping, the stopping criterion is met with both alternative 1 and alternative i having survived, but alternative 1 has a larger worst-case sample mean than alternative i. From now on, when we say "alternative 1 is *killed* by alternative i", we mean that *either* of the above two scenarios occurs.

Lemma 2. *Assume that* $(1,1)\in\mathcal{S}_1(n)$ *and* $(i,1)\in\mathcal{S}_i(n)$ *for all* $n\geq 1$. *Then,*

$$\limsup_{\beta\to 0}\frac{1}{\beta}\mathbb{P}\{\textit{alternative 1 is eliminated by alternative } i\}\leq 1.$$

Proof of Lemma 2. Let M denote the sample size n when the stopping criterion is met. Define

$$\tilde{N}_{11,i1}=\min\{n\geq n_0:\tau^*_{1i}(n)[W_{1i}(n)-C_1(n)]\geq g_c(\tau^*_{1i}(n)) \text{ or } \tau^*_{1i}(n)[W_{1i}(n)+C_i(n)]\leq -g_c(\tau^*_{1i}(n))\},$$

where $\tau^*_{1i}(n)=\min_{(1,j)\in\mathcal{S}_1(n),(i,j')\in\mathcal{S}_i(n)}\tau_{1j,ij'}(n), W_{1i}(n)=\max\limits_{(1,j)\in\mathcal{S}_1(n)}\bar{X}_{1j}(n)-\max\limits_{(i,j)\in\mathcal{S}_i(n)}\bar{X}_{ij}(n)$, and $C_i(n)=\max_{(i,j),(i,j')\in\mathcal{S}_i(n)}g_c(\tau_{ij,ij'}(n))/\tau_{ij,ij'}(n)$. Then, alternative 1 is eliminated by alternative i if and only if

$$\begin{aligned}
& \{\tau^*_{1i}(\tilde{N}_{11,i1})[W_{1i}(\tilde{N}_{11,i1})-C_1(\tilde{N}_{11,i1})]\geq g_c(\tau^*_{1i}(\tilde{N}_{11,i1})),\ \tilde{N}_{11,i1}\leq M<\infty\}\\
\subseteq\ & \{\tau^*_{1i}(\tilde{N}_{11,i1})[W_{1i}(\tilde{N}_{11,i1})-C_1(\tilde{N}_{11,i1})]\geq g_c(\tau^*_{1i}(\tilde{N}_{11,i1})),\ \tilde{N}_{11,i1}<\infty\}. \qquad (24)
\end{aligned}$$

Since $(1,1)\in\mathcal{S}_1(n)$, following the argument for deriving (20) we can show that $\max_{(1,j)\in\mathcal{S}_1(n)}\bar{X}_{1j}(n)-$

$\bar{X}_{11} < C_1(n)$. Hence,

$$W_{1i}(n) - C_1(n) = \max_{(1,j)\in\mathcal{S}_1(n)} \bar{X}_{1j}(n) - \max_{(i,j)\in\mathcal{S}_i(n)} \bar{X}_{ij}(n) - C_1(n) < \bar{X}_{11}(n) - \bar{X}_{i1}(n) \tag{25}$$

It is easy to see that $g_c(t)/t$ is decreasing in $t > 0$, and thus

$$\frac{g_c\tau^*_{1i}(n)}{\tau^*_{1i}(n)} = \max_{(1,j)\in\mathcal{S}_1(n),(i,j')\in\mathcal{S}_i(n)} \frac{g_c(\tau_{1j,ij'}(n))}{\tau_{1j,ij'}(n)} \geq \frac{g_c(\tau_{11,i1}(n))}{\tau_{11,i1}(n)}, \tag{26}$$

where the inequality holds because $(1,1) \in \mathcal{S}_1(n)$ and $(i,1) \in \mathcal{S}_i(n)$. It follows from (24), (25) and (26) that

$$0 \leq [W_{1i}(n) - C_1(n)] - \frac{g_c(\tau^*_{1i}(n))}{\tau^*_{1i}(n)} < [\bar{X}_{11}(n) - \bar{X}_{i1}(n)] - \frac{g_c(\tau_{11,i1}(n))}{\tau_{11,i1}(n)}.$$

Therefore,

$$\begin{aligned}
&\overset{(24)}{\subseteq} \left\{\tau_{11,i1}(\tilde{N}_{11,i1})[\bar{X}_{11}(\tilde{N}_{11,i1}) - \bar{X}_{i1}(\tilde{N}_{11,i1})] \geq g_c(\tau_{11,i1}(\tilde{N}_{11,i1})),\ \tilde{N}_{11,i1} < \infty\right\} \\
&\subseteq \left\{\tau_{11,i1}(\tilde{N}_{11,i1})[\bar{X}_{11}(\tilde{N}_{11,i1}) - \bar{X}_{i1}(\tilde{N}_{11,i1}) - (\mu_{11} - \mu_{i1})] \geq g_c(\tau_{11,i1}(\tilde{N}_{11,i1})),\ \tilde{N}_{11,i1} < \infty\right\}.
\end{aligned} \tag{27}$$

Define $N_{11,i1} = \min\{n \geq n_0 : \left|\tau_{11,i1}(n)[\bar{X}_{11}(n) - \bar{X}_{i1}(n) - (\mu_{11} - \mu_{i1})]\right| \geq g_c(\tau_{11,i1}(n))\}$. By (24) and (27), in order to prove Lemma 2 it suffices to show

$$\limsup_{\beta\to 0} \frac{1}{\beta}\mathbb{P}\{\tau_{11,i1}(\tilde{N}_{11,i1})[\bar{X}_{11}(\tilde{N}_{11,i1}) - \bar{X}_{i1}(\tilde{N}_{11,i1}) - (\mu_{11} - \mu_{i1})] \geq g_c(\tilde{N}_{11,i1})),\ \tilde{N}_{11,i1} < \infty\} \leq 1. \tag{28}$$

Following the proof of Theorem 2 of Fan et al. (2016) and the functional central limit theorem, it can be shown that the left-hand-side of the inequality (28) is upper bounded by its counterpart for the standard Brownian motion, i.e.,

$$\begin{aligned}
&\limsup_{\beta\to 0} \frac{1}{\beta}\mathbb{P}\{\tau_{11,i1}(\tilde{N}_{11,i1})[\bar{X}_{11}(\tilde{N}_{11,i1}) - \bar{X}_{i1}(\tilde{N}_{11,i1}) - (\mu_{11} - \mu_{i1})] \geq g_c(\tilde{N}_{11,i1})),\ \tilde{N}_{11,i1} < \infty\} \\
\leq\ &\limsup_{\beta\to 0} \mathbb{P}\{B(\tilde{T}_{11,i1}) \geq g_c(\tilde{T}_{11,i1}),\ \tilde{T}_{11,i1} < \infty\},
\end{aligned} \tag{29}$$

where $\tilde{T}_{11,i1}$ is the random time that can be seen as the limit of $\tilde{N}_{11,i1}$ as $n \to \infty$. Its explicit form can be written by applying the functional central limit theorem, but we omit it since it is quite involved. Moreover, (27) implies that $T \leq \tilde{T}_{11,i1}$ and $|B(\tilde{T}_{11,i1})| \geq g_c(\tilde{T}_{11,i1})$, where $T = \inf\{t > 0 : |B(t)| \geq g_c(t)\}$. By the symmetry of standard Brownian motion $B(\cdot)$, we have

$$\mathbb{P}\{B(\tilde{T}_{11,i1}) \geq g_c(\tilde{T}_{11,i1}),\ \tilde{T}_{11,i1} < \infty\} \leq \mathbb{P}\{B(T) \geq g_c(T),\ T < \infty\}. \tag{30}$$

Combining (29) and (30),

$$\begin{aligned}
& \limsup_{\beta\to 0}\frac{1}{\beta}\mathbb{P}\{\tau_{11,i1}(\tilde{N}_{11,i1})[\bar{X}_{11}(\tilde{N}_{11,i1})-\bar{X}_{i1}(\tilde{N}_{11,i1})-(\mu_{11}-\mu_{i1})]\geq g_c(\tilde{N}_{11,i1}),\ \tilde{N}_{11,i1}<\infty\}\\
\leq\ & \limsup_{\beta\to 0}\frac{1}{\beta}\mathbb{P}\{B(T)\geq g_c(T),\ T<\infty\}\\
=\ & \limsup_{\beta\to 0}\frac{1}{\beta}\cdot\frac{1}{2}e^{-c/2}=1,
\end{aligned}$$

where the equality follows from Example 6 in Jennen and Lerche (1981). Therefore, (28) is true and the proof is complete. □

Lemma 3. *Assume that* $(1,1)\in\mathcal{S}_1(n)$ *and* $(i,1)\in\mathcal{S}_i(n)$ *for all* $n\geq 1$. *If* $\mu_{i1}-\mu_{11}\geq\delta$, *then*

$$\limsup_{\beta\to 0}\frac{1}{\beta}\mathbb{P}\{\textit{alternative 1 is killed by alternative } i\}\leq 1.$$

Proof of Lemma 3. We follow the notation in the proof of Lemma 2. Notice that alternative 1 is killed by alternative i either immediately after the stopping criterion is met, i.e.,

$$\{W_{1i}(M)>0,\ M<\tilde{N}_{11,i1}\wedge\infty\}\bigcap\{\tau_{1i}^*(M)[\delta-C_1(M)\vee C_i(M)]\geq g_c(\tau_{1i}^*(M))\}, \tag{31}$$

or before the stopping criterion is met, i.e.,

$$\{\tau_{1i}^*(\tilde{N}_{11,i1})[W_{1i}(\tilde{N}_{11,i1})-C_1(\tilde{N}_{11,i1})]\geq g_c(\tau_{1i}^*(\tilde{N}_{11,i1})),\ \tilde{N}_{11,i1}\leq M<\infty\}. \tag{32}$$

Let $W_{1i}^0(n)=W_{1i}(n)-(\mu_{11}-\mu_{i1})$ and

$$\tilde{N}_{11,i1}^0=\min\{n\geq n_0:\tau_{1i}^*(n)[W_{1i}^0(n)-C_1(n)]\geq g_c(\tau_{1i}^*(n))\text{ or }\tau_{1i}^*(n)[W_{1i}^0(n)+C_i(n)]\leq -g_c(\tau_{1i}^*(n))\}.$$

Then,

$$\begin{aligned}
(31)\ =\ & \{W_{1i}^0(M)>\mu_{i1}-\mu_{11},\ M<\tilde{N}_{11,i1}\wedge\infty\}\bigcap\{\tau_{1i}^*(M)[\delta-C_1(M)\vee C_i(M)]\geq g_c(\tau_{1i}^*(M))\}\\
\subseteq\ & \{W_{1i}^0(M)>\delta,\ M<\tilde{N}_{11,i1}\wedge\infty\}\bigcap\{\tau_{1i}^*(M)[\delta-C_1(M)]\geq g_c(\tau_{1i}^*(M))\}\\
\subseteq\ & \left\{\tau_{1i}^*(M)[W_{1i}^0(M)-C_1(M)]\geq g_c(\tau_{1i}^*(M)),\ M<\tilde{N}_{11,i1}\wedge\infty\right\}\\
\subseteq\ & \left\{\tau_{1i}^*(M)[W_{1i}^0(M)-C_1(M)]\geq g_c(\tau_{1i}^*(M)),\ \tilde{N}_{11,i1}^0\leq M<\tilde{N}_{11,i1}\wedge\infty\right\},
\end{aligned} \tag{33}$$

where the last step follows from the definition of $\tilde{N}_{11,i1}^0$. Moreover, notice that

$$\begin{aligned}
\{M<\tilde{N}_{11,i1}\}\ \subseteq\ & \{\tau_{1i}^*(n)[W_{1i}(n)+C_i(n)]>-g_c(\tau_{1i}^*(n))\text{ for all }n\leq M\}\\
\subseteq\ & \{\tau_{1i}^*(n)[W_{1i}^0(n)+C_i(n)]>-g_c(\tau_{1i}^*(n))\text{ for all }n\leq M\},
\end{aligned}$$

since $W_{1i}^0(n) > W_{1i}(n)$. It then follows from (33) that

$$\begin{aligned} & \overset{(31)}{} \\ \subseteq\ & \left\{\tau_{1i}^*(N_{11,i1}^0)[W_{1i}^0(N_{11,i1}^0) - C_1(N_{11,i1}^0)] \geq g_c(\tau_{1i}^*(N_{11,i1}^0)),\ \tilde{N}_{11,i1}^0 \leq M < \tilde{N}_{11,i1} \wedge \infty\right\}. \end{aligned} \tag{34}$$

For (32), the other scenario that can lead to ICS, we notice that since $W_{1i}^0(n) > W_{1i}(n)$,

$$\begin{aligned} & \overset{(32)}{} \\ \subseteq\ & \{\tau_{1i}^*(\tilde{N}_{11,i1})[W_{1i}^0(\tilde{N}_{11,i1}) - C_1(\tilde{N}_{11,i1})] \geq g_c(\tau_{1i}^*(\tilde{N}_{11,i1})),\ \tilde{N}_{11,i1} \leq M < \infty\} \\ \subseteq\ & \{\tau_{1i}^*(\tilde{N}_{11,i1})[W_{1i}^0(\tilde{N}_{11,i1}) - C_1(\tilde{N}_{11,i1})] \geq g_c(\tau_{1i}^*(\tilde{N}_{11,i1})),\ \tilde{N}_{11,i1}^0 \leq \tilde{N}_{11,i1} \leq M < \infty\}. \end{aligned} \tag{35}$$

Moreover,

$$\begin{aligned} \{\tilde{N}_{11,i1}^0 \leq \tilde{N}_{11,i1}\} \ &\subseteq\ \{\tau_{1i}^*(n)[W_{1i}(n) + C_i(n)] > -g_c(\tau_{1i}^*(n)) \text{ for all } n < \tilde{N}_{1i}^0\} \\ &\subseteq\ \{\tau_{1i}^*(n)[W_{1i}^0(n) + C_i(n)] > -g_c(\tau_{1i}^*(n)) \text{ for all } n < \tilde{N}_{1i}^0\}. \end{aligned}$$

It then follows from (35) that

$$\begin{aligned} & \overset{(32)}{} \\ \subseteq\ & \{\tau_{1i}^*(\tilde{N}_{11,i1}^0)[W_{1i}^0(\tilde{N}_{11,i1}^0) - C_1(\tilde{N}_{11,i1}^0)] \geq g_c(\tau_{1i}^*(\tilde{N}_{11,i1}^0)),\ \tilde{N}_{11,i1}^0 \leq \tilde{N}_{11,i1} \leq M < \infty\}. \end{aligned} \tag{36}$$

By (34) and (36),

$$\begin{aligned} & \mathbb{P}\{\text{alternative 1 is killed by alternative } i\} \\ \leq\ & \mathbb{P}\{\tau_{1i}^*(\tilde{N}_{11,i1}^0)[W_{1i}^0(\tilde{N}_{11,i1}^0) - C_1(\tilde{N}_{11,i1}^0)] \geq g_c(\tau_{1i}^*(\tilde{N}_{11,i1}^0)),\ \tilde{N}_{11,i1}^0 < \infty\}. \end{aligned}$$

Hence, in order to prove Lemma 3, it suffices to show

$$\limsup_{\beta \to 0} \frac{1}{\beta} \mathbb{P}\{\tau_{1i}^*(\tilde{N}_{11,i1}^0)[W_{1i}^0(\tilde{N}_{11,i1}^0) - C_1(\tilde{N}_{11,i1}^0)] \geq g_c(\tau_{1i}^*(\tilde{N}_{11,i1}^0)),\ \tilde{N}_{11,i1}^0 < \infty\} \leq 1.$$

This can be done by adopting a proof that is essentially identical to the discussion between (24) and the end of the proof of Lemma 2. □

To establish Theorem 2, we need one additional building block. Notice that a common assumption shared by both Lemma 2 and Lemma 3 is that when alternative i is compared with another alternative, system $(i, 1)$, the worst system of alternative i, is not yet eliminated in the inner-layer elimination. This assumption essentially guarantees that the worst-case mean performance of alternative i can be accurately estimated via a dynamic confidence interval; see Proposition 1. Therefore, we need to characterize the probability that system $(i, 1)$ is eliminated by some other system (i, j).

Lemma 4. *In the inner-layer selection process of the sequential RSB procedure,*

$$\limsup_{\beta\to 0}\ \frac{1}{\beta}\mathbb{P}\{\textit{system } (i,1) \textit{ is eliminated by system } (i,j)\} \le 1,$$

for each $i = 1, 2, \ldots, k$ and each $j = 2, 3, \ldots, m$.

Proof of Lemma 4. Define $N'_{i1,ij} = \min\{n \ge n_0 : \tau_{i1,ij}(n)|\bar{X}_{i1}(n) - \bar{X}_{ij}(n)| \ge g_c(\tau_{i1,ij}(n))\}$, for each $i = 1, 2, \ldots, k$ and each $j = 2, 3, \ldots, m$. Then,

$$\begin{aligned}
&\mathbb{P}\{\text{system } (i,1) \text{ is eliminated by system } (i,j)\} \\
=\ &\mathbb{P}\left\{\bar{X}_{i1}(N_{i1,ij}) - \bar{X}_{ij}(N'_{i1,ij}) \le -\frac{g_c(\tau_{i1,ij}(N'_{i1,ij}))}{\tau_{i1,ij}(N'_{i1,ij})}, N'_{i1,ij} < \infty\right\} \\
\le\ &\mathbb{P}\left\{\bar{X}_{i1}(N_{i1,ij}) - \bar{X}_{ij}(N'_{i1,ij}) - (\mu_{i1} - \mu_{ij}) \le -\frac{g_c(\tau_{i1,ij}(N'_{i1,ij}))}{\tau_{i1,ij}(N'_{i1,ij})}, N'_{i1,ij} < \infty\right\},
\end{aligned}$$

where the inequality holds because $\mu_{i1} - \mu_{ij} > 0$. The proof is completed by applying Lemma 1(ii). □

Proof of Theorem 2. We first notice that if $\mu_{k1} - \mu_{11} \le \delta$, then by Assumption 1 $\mu_{i1} - \mu_{11} \le \delta$ for all $i = 1, \ldots, k$, which implies that $\mathbb{P}(\mu_{i^*1} - \mu_{11} \le \delta) = 1$ and the theorem trivially holds. Hence, without loss of generality we assume that there exists $l = 1, \ldots, k-1$ for which $\mu_{l1} - \mu_{11} \le \delta$ and $\mu_{l+1,1} - \mu_{11} > \delta$. Then, a good selection event (i.e., alternative i is selected for any $i = 1, \ldots, l$) occurs if $\bigcap_{i=l+1}^{k}\{\text{alternative 1 kills alternative } i\}$. We denote

$$\begin{aligned}
A &= \bigcap_{i=l+1}^{k}\{\text{alternative 1 kills alternative } i\} \\
B &= \bigcap_{i=2}^{l}\{\text{alternative 1 is not eliminated by alternative } i\} \\
C &= \bigcap_{i=1}^{k}\bigcap_{j=2}^{m}\{\text{system } (i,1) \text{ is not eliminated by system } (i,j)\}.
\end{aligned}$$

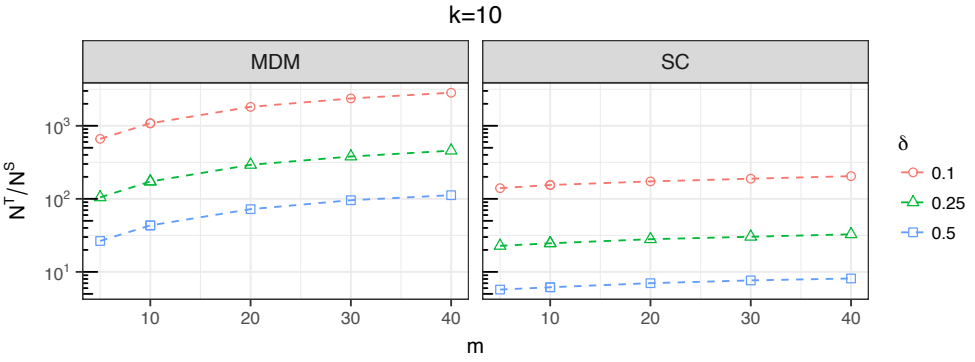
Clearly, $\mathbb{P}(A \cap B|C) \ge 1 - \mathbb{P}(A^{\mathsf{c}}|C) - \mathbb{P}(B^{\mathsf{c}}|C)$. Multiplying $\mathbb{P}(C)$ on both sides this inequality yields

$$\mathbb{P}(A \cap B \cap C) \ge \mathbb{P}(C) - \mathbb{P}(A^{\mathsf{c}} \cap C) - \mathbb{P}(B^{\mathsf{c}} \cap C) = 1 - \mathbb{P}(C^{\mathsf{c}}) - \mathbb{P}(A^{\mathsf{c}} \cap C) - \mathbb{P}(B^{\mathsf{c}} \cap C).$$

Since $\mathbb{P}\{\mu_{i^*1} - \mu_{11} \le \delta\} \ge \mathbb{P}(A) \ge \mathbb{P}(A \cap B \cap C)$, it follows that

$$\mathbb{P}\{\mu_{i^*1} - \mu_{11} > \delta\} \le \mathbb{P}(A^{\mathsf{c}} \cap C) + \mathbb{P}(B^{\mathsf{c}} \cap C) + \mathbb{P}(C^{\mathsf{c}}). \tag{37}$$

Figure 6: Average Sample Sizes of Procedure T and Procedure S Under the EV Configuration



Note. The vertical axis is on a logarithmic scale with base 10.

Notice that

$$\begin{aligned}
\mathbb{P}(A^c \cap C) &\leq \sum_{i=l+1}^{k} \mathbb{P}\{\text{alternative 1 is killed by alternative } i, (1,1) \in \mathcal{S}_1(n) \text{ and } (i,1) \in \mathcal{S}_i(n) \text{ for all n}\} \\
\mathbb{P}(B^c \cap C) &\leq \sum_{i=1}^{l} \mathbb{P}\{\text{alternative 1 is eliminated by alternative } i, (1,1) \in \mathcal{S}_1(n) \text{ and } (i,1) \in \mathcal{S}_i(n) \text{ for all n}\} \\
\mathbb{P}(C^c) &\leq \sum_{i=1}^{k}\sum_{j=2}^{m} \mathbb{P}\{\text{system } (i,1) \text{ is eliminated by system } (i,j)\}.
\end{aligned}$$

Combining Lemma 2, Lemma 3, Lemma 4, and (37), we have

$$\limsup_{\alpha\to 0} \frac{1}{\alpha}\mathbb{P}\{\mu_{i^*1} - \mu_{11} > \delta\} = \limsup_{\beta\to 0} \frac{\beta}{\alpha}\cdot\frac{1}{\beta}\mathbb{P}\{\mu_{i^*1} - \mu_{11} > \delta\} \leq \frac{1}{km-1}[(k-l) + l + k(m-1)] = 1,$$

which completes the proof. □

C. Additional Comparison Between Procedure T and Procedure S

We fix $k = 10$ and plot in Figure 3 the ratio of the average sample size of Procedure T to that of Procedure S (i.e., N^T/N^S) as a function of m. The result associated with the case of fixing $m = 10$ and varying k is almost identical and thus it is omitted.

D. Comparison Between Procedure S and Procedure V

Procedure 3 in Fan et al. (2016) requires an IZ parameter and we set it to be $\delta/2$ when the procedure is applied to both the inner-layer and outer-layer selection of Procedure V. This is inspired by the decomposition of the IZ parameter in Section 3.1 for the two-stage RSB procedure.

Procedure 3 (Procedure V).

0. *Setup.* Specify the error allowance $\beta = \alpha/(km-1)$ and the first-stage sample size $n_0 \geq 2$. Set $c = -2\log(2\beta)$ and $g_c(t) = \sqrt{[c + \log(t+1)](t+1)}$.

1. *Initialization:* Set $n = n_0$. Set $\mathcal{S} = \{1, 2, \ldots, k\}$ to be the set of surviving alternatives. Set $\mathcal{S}_i = \{(i,j) : j = 1, 2, \ldots, m\}$ to be the set of surviving systems of alternative i, $i = 1, \ldots, k$. Take n independent replications $X_{ij,1}, \ldots, X_{ij,n}$ of each system (i,j). Solve $T\delta/2 - g_c(T) = 0$ for T^*.

2. *Inner-layer Elimination.* For each $i \in \mathcal{S}$, do the following.

   2.1 *Updating.* Compute

$$\bar{X}_{ij}(n) = \frac{1}{n}\sum_{r=1}^{n} X_{ij,r}, \quad i \in \mathcal{S},\ (i,j) \in \mathcal{S}_i,$$
$$S^2_{ij,ij'}(n) = \frac{1}{n-1}\sum_{r=1}^{n}\left[X_{ij,r} - X_{ij',r} - (\bar{X}_{ij}(n) - \bar{X}_{ij'}(n))\right]^2, \quad (i,j),(i,j') \in \mathcal{S}_i.$$

   2.2 *Screening.* Compute

$$\tau_{ij,ij'}(n) = \frac{n}{S^2_{ij,ij'}(n)} \quad \text{and} \quad Z_{ij,ij'}(n) = \tau_{ij,ij'}(n)[\bar{X}_{ij}(n) - \bar{X}_{ij'}(n)], \quad (i,j),(i,j') \in \mathcal{S}_i.$$

   Assign $\mathcal{S}_i \leftarrow \mathcal{S}_i \setminus \{(i,j) \in \mathcal{S}_i : Z_{ij,ij'}(n) \leq -g_c(\tau_{ij,ij'}(n)) \text{ for some } (i,j') \in \mathcal{S}_i\}$.

   2.3 *Stopping.* If either $|\mathcal{S}_i| = 1$ or $\tau_{ij,ij'}(n) \geq T^*$ for all $(i,j),(i,j') \in \mathcal{S}_i$ with $j \neq j'$, then stop and select $j_i^* = \underset{j:(i,j)\in\mathcal{S}_i}{\arg\max}\, \bar{X}_{ij}(n)$ as the worst system. Otherwise, take one additional replication of each $(i,j) \in \mathcal{S}_i$ with $i \in \mathcal{S}$, assign $n \leftarrow n+1$, and return to step 2.1.

3. *Outer-layer Elimination.*

   3.1 *Updating.* Compute

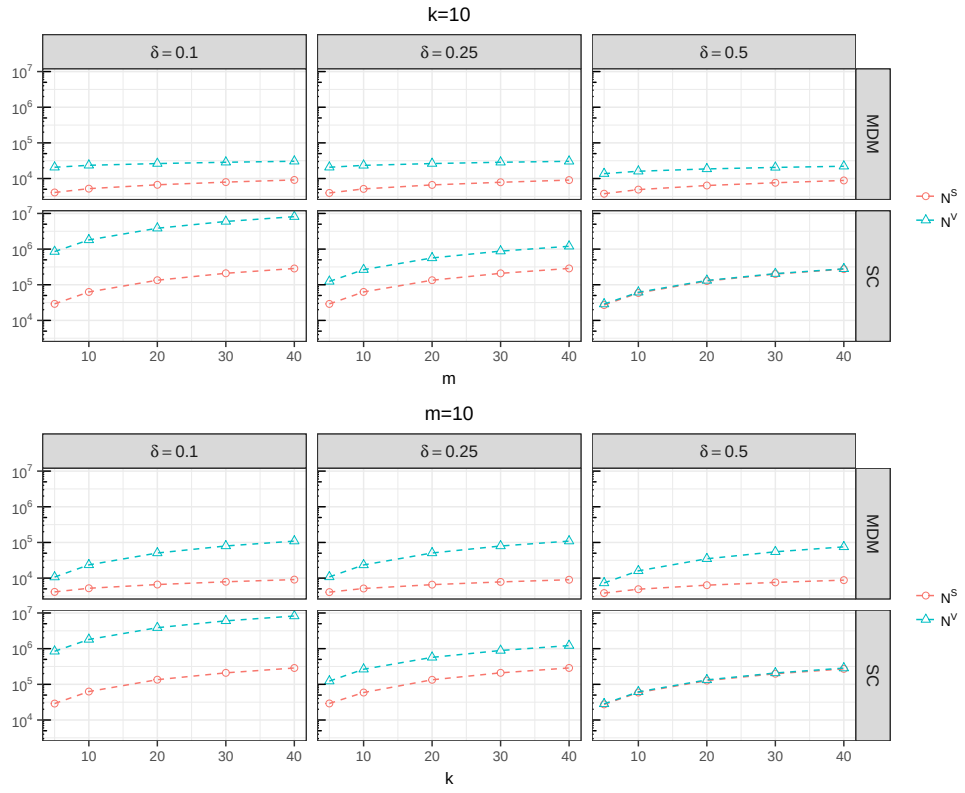
$$\bar{X}_{ij_i^*}(n) = \frac{1}{n}\sum_{r=1}^{n} X_{ij,r}, \quad i \in \mathcal{S},$$
$$S^2_{ij_i^*,i'j_{i'}^*}(n) = \frac{1}{n-1}\sum_{r=1}^{n}\left[X_{ij_i^*,r} - X_{i'j_{i'}^*,r} - (\bar{X}_{ij_i^*}(n) - \bar{X}_{i'j_{i'}^*}(n))\right]^2, \quad i,i' \in \mathcal{S}.$$

   3.2 *Screening.* For each $i, i' \in \mathcal{S}$ with $i \neq i'$, compute

$$\tau_{ij_i^*,i'j_{i'}^*}(n) = \frac{n}{S^2_{ij_i^*,i'j_{i'}^*}(n)} \quad \text{and} \quad Z_{ij_i^*,i'j_{i'}^*}(n) = \tau_{ij_i^*,i'j_{i'}^*}(n)[\bar{X}_{ij_i^*}(n) - \bar{X}_{i'j_{i'}^*}(n)], \quad i,i' \in \mathcal{S}.$$

   Assign $\mathcal{S} \leftarrow \mathcal{S} \setminus \{(i,j) \in \mathcal{S} : Z_{ij_i^*,i'j_{i'}^*}(n) \geq g_c(\tau_{ij_i^*,i'j_{i'}^*}(n)) \text{ for some } i' \in \mathcal{S}\}$.

Figure 7: Average Sample Sizes of Procedure S and Procedure V Under the EV configuration



Note. Top: m varies with $k = 10$; Bottom: k varies with $m = 10$. The vertical axis is on a logarithmic scale.

3.3 *Stopping.* If either $|\mathcal{S}| = 1$ or $\tau_{ij_i^*,i'j_{i'}^*}(n) \geq T^*$ for all $i, i' \in \mathcal{S}$ with $i \neq i'$, then stop and select $i^* = \arg\min_{i \in \mathcal{S}} \bar{X}_{ij_i^*}(n)$ as the best alternative. Otherwise, take one additional replication of each (i, j_i^*) with $i \in \mathcal{S}$, assign $n \leftarrow n + 1$, and return to step 3.1. □

The numerical results for the EV configuration are presented in Figure 7. The results for the other two configurations of the variances are very similar so we omit them.

First, as expected, Procedure S requires significantly fewer samples than Procedure V in general. In particular, under SC, if the IZ parameter δ happens to be the difference between the best and the second-best worst-case mean performances (i.e., $\delta = \mu_{21} - \mu_{11} = 0.5$), then the average sample sizes required by the two procedures are almost the same, regardless of the problem scale. This implies that in this case, simultaneous elimination of the surviving systems of an alternative that is unlikely to be the best rarely happens in Procedure S, which diminishes its advantage over Procedure V. This is because under SC, the outer-layer selection process deals with the worst-case mean performances $(0, 0.5, \ldots, 0.5)$. With $\delta = 0.5$, the alternatives are hard to differentiate in early iterations of Procedure S when the sample size is large enough.

Second, the average sample size of Procedure V is more heavily affected by the configurations of the means than Procedure S. With everything else the same, the average sample size required by Procedure V

(denoted by N^{V}) increases faster than that of Procedure S (denoted by N^{S}) under MDM than under SC. This suggests that there are a significantly larger number of early outer-layer eliminations in Procedure S under MDM than under SC.

Third, the average sample size of Procedure V is much more sensitive to δ under SC than that of Procedure S. For instance, with $k = 30$ and $m = 10$, N^{V} increases from about 8.82×10^5 to 5.96×10^6 as δ drops from 0.25 to 0.1, respectively, whereas N^{S} almost remains the value 2.10×10^5. The reason is as follows. The inner-layer selection process of Procedure V that relies on Procedure 3 in Fan et al. (2016) faces systems with equal means under SC. It does not terminate until its stopping criterion that depends on the IZ parameter δ is met. This stopping criterion is harder to meet for a smaller value of δ. Hence, a smaller δ implies that the inner-layer selection process of Procedure V needs more time to terminate, resulting in more required samples.

Last, N^{V} grows much faster than N^{S} as the problem scale k or m increases. For instance, with $\delta = 0.25$, $k = 10$ and MDM, N^{V} increases from about 2.90×10^4 to 2.87×10^5 as m increases from 5 to 40, whereas N^{S} increases from about 3.94×10^3 to 9.04×10^3. This is because, as the problem scale increases, there are more opportunities for Procedure S to eliminate alternatives early, leading to a slower growth in N^{S}.

The above numerical comparison between Procedure S and Procedure V indicates that the inferior performance of the latter stems from its non-fully sequential nature – its outer-layer selection cannot begin unless all the inner-layer eliminations are completed. Hence, an alternative having a configuration of the means that is close to the SC will dominate the inner-elimination time, even if it were otherwise a poor alternative for outer elimination. This suggests an additional comparison between Procedure S and Procedure V using a configuration of the means that somewhat combines MDM and SC as follows

$$[\mu_{ij}]_{k\times m} = \begin{pmatrix} 0 & -0.2 & -0.2 & \dots & -0.2 \\ 0.5 & 0.3 & 0.3 & \dots & 0.3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.5(k-1) & 0.5(k-1)-0.2 & 0.5(k-1)-0.2 & \dots & 0.5(k-1)-0.2 \end{pmatrix}.$$

Here, the alternatives are ordered as MDM, but the systems of each alternative are ordered as SC.

We adopt the EV configuration of the variances. The other experiment specifications remain the same. The results are given in Figure 8 and they are consistent with the findings revealed by Figure 7.

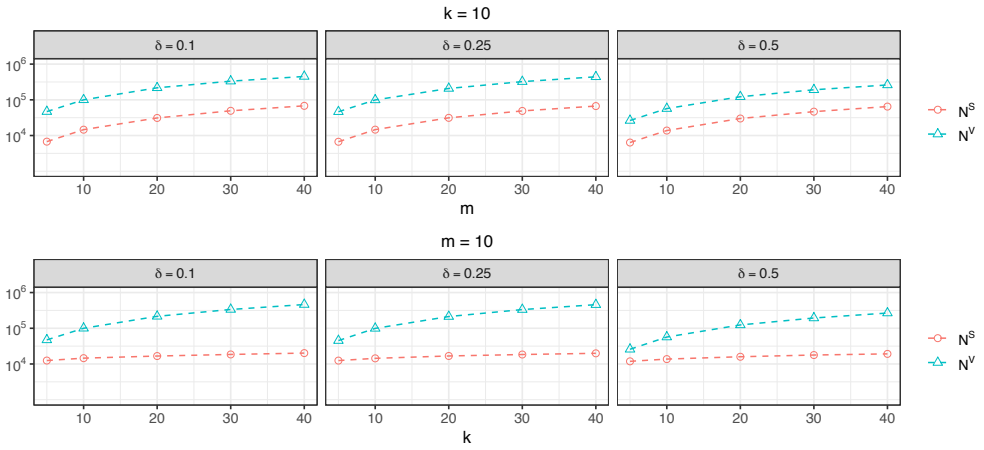
E. Realized PCS of the $G/G/s + G$ Queueing Example

This section assesses efficacy of the two proposed RSB procedures as to whether they can achieve the target PCS as promised. This complements the analysis in Section 5, since the samples in queueing simulation are not normally distributed in general, in contrast to the normal assumptions of the numerical experiments there.

With $\sigma = 2$ and $\ell = 50$, we run 1,000 macro-replications of the experiment below.

(i) Generate a sample of service times from P_0.

Figure 8: Average Sample Sizes of Procedure S and Procedure V



Note. Top: m varies with $k = 10$; Bottom: k varies with $m = 10$. The vertical axis is on a logarithmic scale.

(ii) Construct an ambiguity set $\mathcal{P}$ based on the sample.

(iii) Compute the expected cost $\mathbb{E}[f(s, \xi)]$ with 10,000 samples for each pair (s, P), $s = 1, \ldots, k$, $P \in \mathcal{P}$ so that the estimation errors are negligible and find the best alternative,

(iv) Run the two RSB procedures 1,000 times independently on $\mathcal{P}$ and estimate their respective PCS. The IZ parameter δ is set to be small enough so that the indifference zone contains only the best alternative.

In summary, there are 100 ambiguity sets constructed in total, each from one macro-replication. Hence, PCS is estimated 100 times for each RSB procedure and some statistics of these estimated probabilities are reported in Table 4. Clearly, both procedures can achieve the target PCS even by a large margin in general, despite the samples' non-normal distribution. Moreover, the two-stage RSB procedure is significantly more conservative than the sequential RSB procedure, producing a larger realized PCS. This reflects that the former requires a larger number of samples, which is consistent with the findings in Section 5.

Table 4: Realized PCS

Procedure	Statistics				
	Min	25% Quantile	Median	75% Quantile	Max
Two-stage	0.992	0.998	0.999	1.000	1.000
Sequential	0.951	0.980	0.991	0.996	1.000

Note. Target PCS: 0.95.

Acknowledgments

The authors would like to thank the associate editor and three anonymous referees for their insightful and invaluable comments that have significantly improved this paper. The preliminary work of this paper (Fan et al. 2013) was presented at the 2013 Winter Simulation Conference. The first author was supported by Natural Science Foundation of China (Grant 71701196). The second author was supported by Hong Kong Research Grants Council (GRF 16203214) and Natural Science Foundation of China (Grant 71720107003). The third author was supported by Hong Kong Research Grants Council (TRS No. T32-102/14N and GRF 16211417).

References

Asmussen, S. (2003). *Applied Probability and Queues* (2nd ed.). Springer.

Barton, R. R., B. L. Nelson, and W. Xie (2014). Quantifying input uncertainty via simulation confidence intervals. *INFORMS J. Comput. 26*(1), 74–87.

Barton, R. R. and L. W. Schruben (2001). Resampling methods for input modeling. In *Proc. 2001 Winter Simulation Conf.*, pp. 372–378.

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Stat. 25*(1), 16–39.

Ben-Tal, A., D. den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Manag. Sci. 59*(2), 341–357.

Ben-Tal, A., L. El Ghaoui, and A. Nemirovski (2009). *Robust Optimization*. Princeton University Press.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc. 100*(1), 36–50.

Cheng, R. C. H. and W. Holland (1997). Sensitivity of computer simulation experiments to errors in input data. *J. Stat. Comput. Simul. 57*, 219–241.

Chick, S. E. (2001). Input distribution selection for simulation experiments: Accounting for input uncertainty. *Oper. Res. 49*(5), 744–758.

Chick, S. E., J. Branke, and C. Schmidt (2010). Sequential sampling to myopically maximize the expected value of information. *INFORMS J. Comput. 22*(1), 71–80.

Chick, S. E. and P. Frazier (2012). Sequential sampling with economics of selection procedures. *Manag. Sci. 58*(3), 550–569.

Chick, S. E. and Y. Wu (2005). Selection procedures with frequentist expected opportunity cost bounds. *Oper. Res. 53*(5), 867–878.

Corlu, C. G. and B. Biller (2013). A subset selection procedure under input parameter uncertainty. In *Proc. 2013 Winter Simulation Conf.*, pp. 463–473.

Corlu, C. G. and B. Biller (2015). Subset selection for simulations accounting for input uncertainty. In *Proc. 2015 Winter Simulation Conf.*, pp. 437–446.

Delage, E. and Y. Ye (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res. 58*(3), 595–612.

Dudewicz, E. J. (1969). An approximation to the sample size in selection problems. *Ann. Math. Stat. 40*(2), 492–497.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quart. J. Econom. 75*(4), 643–669.

Epstein, L. G. (1999). A definition of uncertainty aversion. *Rev. Econom. Stud. 66*(3), 579–608.

Fan, W., L. J. Hong, and B. L. Nelson (2016). Indifference-zone-free selection of the best. *Oper. Res. 64*(6), 1499–1514.

Fan, W., L. J. Hong, and X. Zhang (2013). Robust selection of the best. In *Proc. 2013 Winter Simulation Conf.*, pp. 868–876.

Frazier, P. (2014). A fully sequential elimination procedure for indifference-zone ranking and selection with tight bounds on probability of correct selection. *Oper. Res. 62*(4), 926–942.

Frazier, P., W. Powell, and S. Dayanik (2009). The knowledge-gradient policy for correlated normal beliefs. *INFORMS J. Comput. 21*(4), 599–613.

Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with non-unique prior. *J. Math. Econ. 18*(2), 141–153.

Gupta, D. and B. Denton (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Trans. 40*(9), 800–819.

He, D., S. E. Chick, and C.-H. Chen (2007). Opportunity cost and OCBA selection procedures in ordinal optimization for a fixed number of alternative systems. *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev. 37*(5), 951–961.

Henderson, S. G. (2003). Input model uncertainty: Why do we care and what should we do about it? In *Proc. 2003 Winter Simulation Conf.*, pp. 90–100.

Hong, L. J. and B. L. Nelson (2005). The tradeoff between sampling and switching: New sequential procedures for indifference-zone selection. *IIE Trans. 37*(7), 623–634.

Hu, Z., J. Cao, and L. J. Hong (2012). Robust simulation of global warming policies using the dice model. *Manag. Sci. 58*(12), 2190–2206.

Hu, Z. and L. J. Hong (2015). Robust simulation of stochastic systems with input uncertainties modeled by statistical divergences. In *Proc. 2015 Winter Simulation Conf.*, pp. 643–654.

Jennen, C. and H. R. Lerche (1981). First exit densities of Brownian motion through one-sided moving boundaries. *Z. Wahrsch. Verw. Gebiete 55*(2), 133–148.

Kelton, W. D., R. P. Sadowski, and N. B. Swets (2009). *Simulation with Arena* (5th ed.). McGraw-Hill Education.

Kim, S.-H. and B. L. Nelson (2001). A fully sequential procedure for indifference-zone selection in simulation. *ACM Trans. Model. Comput. Simul. 11*(3), 251–273.

Kim, S.-H. and B. L. Nelson (2006). Selecting the best system. In *Handbooks in Operations Research and Management Science*, Volume 13, pp. 501–534. Elsevier.

Kong, Q., C.-Y. Lee, C.-P. Teo, and Z. Zheng (2013). Scheduling arrivals to stochastic service delivery system using copositive cones. *Oper. Res. 61*(3), 711–726.

Kong, Q., C.-Y. Lee, C.-P. Teo, and Z. Zheng (2016). Appointment sequencing: Why the smallest-variance-first rule may not be optimal. *Eur. J. Oper. Res. 255*(3), 809–821.

Macario, A. (2010). Is it possible to predict how long a surgery will last? *Medscape Anesthesiology*, https://www.medscape.com/viewarticle/724756.

Mak, H.-Y., Y. Rong, and J. Zhang (2015). Appointment scheduling with limited distributional information. *Manag. Sci. 59*(7), 1557–1575.

Ni, E. C., D. F. Ciocan, S. G. Henderson, and S. R. Hunter (2017). Efficient ranking and selection in parallel computing environments. *Oper. Res. 65*(3), 821–836.

Perng, S. K. (1969). A comparison of the asymptotic expected sample sizes of two sequential procedures for ranking problem. *Ann. Math. Stat. 40*(6), 2198–2202.

Qi, J. (2017). Mitigating delays and unfairness in appointment systems. *Manag. Sci. 63*(2), 566–583.

Rinott, Y. (1978). On two-stage selection procedures and related probability-inequalities. *Comm. Stat. Theor. Meth. 7*(8), 799–811.

Song, E., B. L. Nelson, and L. J. Hong (2015). Input uncertainty and indifference-zone ranking & selection. In *Proc. 2015 Winter Simulation Conf.*, pp. 414–424.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Stat. 16*, 243–258.

Strum, D., J. May, and L. Vargas (2000). Modeling the uncertainty of surgical procedure times: Comparison of the log-normal and normal models. *Anesthesiology 92*(4), 1160–1167.

Whitt, W. (2002). *Stochastic-process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer.

Xie, W., B. L. Nelson, and R. R. Barton (2014). A Bayesian framework for quantifying uncertainty in stochastic simulation. *Oper. Res. 62*(6), 1439–1452.