Optimal Sampling Laws for Stochastically Constrained Simulation Optimization on Finite Sets

Susan R. Hunter, Raghu Pasupathy
The Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA 24061, USA
{srhunter@vt.edu, pasupath@vt.edu}

Consider the context of selecting an optimal system from amongst a finite set of competing systems, based on a "stochastic" objective function and subject to multiple "stochastic" constraints. In this context, we characterize the asymptotically optimal sample allocation that maximizes the rate at which the probability of false selection tends to zero. Since the optimal allocation is the result of a concave maximization problem, its solution is particularly easy to obtain in contexts where the underlying distributions are known or can be assumed, e.g., normal, Bernoulli. We provide a consistent estimator for the optimal allocation, and a corresponding sequential algorithm that is fit for implementation. Various numerical examples demonstrate where and to what extent the proposed allocation differs from competing algorithms.

1. Introduction

The simulation-optimization (SO) problem is a nonlinear optimization problem where the objective and constraint functions, defined on a set of candidate solutions or "systems," are observable only through consistent estimators. The consistent estimators can be defined implicitly, e.g., through a stochastic simulation model. Since the functions involved in SO can be specified implicitly, the formulation affords virtually any level of complexity. Due to this generality, the SO problem has received much attention from both researchers and practitioners in the last decade. Variations of the SO problem are readily applicable in such diverse contexts as vehicular transportation networks, quality control, telecommunication systems, and health care. See Andradóttir [2006], Spall [2003], Fu [2002], Barton and Meckesheimer [2006], and Ólafsson and Kim [2002] for overviews and entry points into this literature, and Henderson and Pasupathy [2011] for a collection of contributed SO problems.

SO's large number of variations stem primarily from differences in the nature of the feasible set and constraints. Among SO's variations, the unconstrained SO problem on finite sets has arguably seen the most development. Appearing broadly as ranking and selection [Kim and Nelson, 2006], the currently available solution methods are reliable and have stable digital implementations. In contrast, the constrained version of the problem — SO on finite sets having "stochastic" constraints

— has seen far less development, despite its usefulness in the context of multiple performance measures.

To explore the constrained SO variation in more detail, consider the following setting. Suppose there exist multiple performance measures defined on a finite set of systems, one of which is primary and called the objective function, while the others are secondary and called the constraint functions. Suppose further that the objective and constraint function values are estimable for any given system using a stochastic simulation, and that the quality of the objective and constraint function estimators is dependent on the simulation effort expended. The constrained SO problem is then to identify that system having the best objective function value, from amongst those systems whose constraint values cross a pre-specified threshold, using only the simulation output. The efficiency of a solution to this problem, which we will define in rigorous terms later in the paper, is measured in terms of the total simulation effort expended.

The broad objective of our work is to rigorously characterize the nature of optimal sampling plans when solving the constrained SO problem on finite sets. As we demonstrate, such characterization is extremely useful in that it facilitates the construction of asymptotically optimal algorithms. The specific questions we ask along the way are twofold.

Q.1 Let an algorithm for solving the constrained SO problem estimate the objective and constraint functions by allocating a portion of an available simulation budget to each competing system. Suppose further that this algorithm returns to the user that system having the best estimated objective function, amongst the estimated-feasible systems. As the simulation budget increases, the probability that such an algorithm returns any system other than the truly best system decays to zero. Can the asymptotic behavior of this probability of false selection be characterized? Specifically, can its rate of decay be deduced as a function of the sampling proportions allocated to the various systems?

Q.2 Given a satisfactory answer to Q.1, can a method be devised to identify the sampling proportion that maximizes the rate of decay of the probability of false selection?

This work answers both of the above questions in the affirmative. Relying on large-deviation principles and generalizing prior work in the context of unconstrained systems [Glynn and Juneja, 2004], we fully characterize the probabilistic decay behavior of the false selection event as a function of the budget allocations. We then use this characterization to formulate a mathematical program whose solution is the allocation that maximizes the rate of probabilistic decay. Since the constructed mathematical program is a concave maximization problem, identifying the asymptotically optimal solution is easy, at least in contexts where the underlying distributional family of the simulation estimator is known or assumed.

1.1 This Work in Context

Prior research on selecting the best system in the unconstrained context falls broadly under one of three categories:

- traditional ranking and selection (R&S) procedures [see, e.g., Kim and Nelson, 2006, for an overview], which typically require a normality assumption and provide finite-time probabilistic guarantees on the probability of false selection,

- the Optimal Computing Budget Allocation (OCBA) framework [see, e.g., Chen et al., 2000], which, under the assumption of normality, provides an approximately optimal sample allocation, and

- the large-deviations (LD) approach [see, e.g., Glynn and Juneja, 2004], which provides an asymptotically optimal sample allocation in the context of general light-tailed distributions.

Corresponding research in the constrained context is taking an analogous route. For example, as illustrated in table 1, recent work by Andradóttir and Kim [2010] provides finite-time guarantees on the probability of false selection in the context of "stochastically" constrained SO and parallels traditional R&S work. Similarly, recent work by Lee et al. [2011] in the context of "stochastically" constrained SO parallels the previous OCBA work in the unconstrained context. Our work, which appears in the bottom left-hand cell of table 1, provides the complete generalization of previous large deviations work in ordinal optimization by Glynn and Juneja [2004] and in feasibility determination by Szechtman and Yücesan [2008].

1.2 Problem Statement

Consider a finite set $i = 1, \ldots, r$ of systems, each with an unknown objective value $h_i \in \mathbb{R}$ and unknown constraint values $g_{ij} \in \mathbb{R}$. Given constants $\gamma_j \in \mathbb{R}$, we wish to select the system with the lowest objective value h_i, subject to the constraints $g_{ij} \geq \gamma_j$, $j = 1, 2, \ldots, s$. That is, we consider

$$\text{Problem } P: \quad \arg\min_{i=1,\ldots,r} \; h_i \quad \text{s.t.} \quad g_{ij} \geq \gamma_j, \text{ for all } j = 1, 2, \ldots, s;$$

where h_i and g_{ij} are expectations, estimates of h_i and g_{ij} are observed together through simulation as sample means, and a unique solution to Problem P is assumed to exist.

Table 1: Research in the area of simulation optimization on finite sets can be categorized by the nature of the result, the required distributional assumption, and the presence of objective function or constraints.

Result Time	Required Dist'n	Optimization: only objective(s)	Feasibility: only constraint(s)	Constrained Optimization: objective(s) & constraint(s)
Finite	Normal	Ranking & Selection [e.g., Kim and Nelson, 2006]	Batur and Kim [2010]	Andradóttir and Kim [2010]
Infinite	Normal	OCBA [e.g., Chen et al., 2000]	[application of general solution][1]	OCBA-CO [Lee et al., 2011]
	General	Glynn and Juneja [2004]	Szechtman and Yücesan [2008]	?

[1] Problems lying in the infinite-time, normal row are also solved as applications of the solutions in the infinite-time, general row.

Let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_r)$ be a vector denoting the proportion of the total sampling budget given to each system, so that $\sum_{i=1}^{r} \alpha_i = 1$ and $\alpha_i \geq 0$ for all $i = 1, 2, \ldots, r$. Furthermore, let the system having the smallest estimated objective value amongst the estimated-feasible systems be selected as the estimated solution to Problem P. Then we ask, what vector of proportions $\boldsymbol{\alpha}$ maximizes the rate of decay of the probability that this procedure returns a suboptimal solution to Problem P?

1.3 Organization

In Section 2 we discuss the contributions of this work. Notation and assumptions for the paper are described in Section 3. In Section 4 we derive an expression for the rate function of the probability of false selection. In Section 5, we present a general sampling framework and a conceptual algorithm to solve for the optimal allocation. A consistent estimator and an implementable sequential algorithm for the optimal allocation is provided in Section 6. Section 7 contains numerical illustrations for the normal case and a comparison with OCBA-CO [Lee et al., 2011]. Section 8 contains concluding remarks.

2. Contributions

This paper addresses the question of identifying the "best" amongst a finite set of systems in the presence of multiple "stochastic" performance measures, one of which is used as an objective function and the rest as constraints. This question has been identified as a crucial generalization of the problem of unconstrained simulation optimization on finite sets [Glynn and Juneja, 2004]. The following are our specific contributions.

C.1 We present the first complete characterization of the optimal sampling plan for constrained SO on finite sets when the performance measures can be observed as simulation output. Relying on a large-deviations framework, we derive the probability law for erroneously obtaining a suboptimal solution as a function of the sampling plan. We then demonstrate that the optimal sampling plan can be identified as the solution to a strictly concave maximization problem.

C.2 We present a consistent estimator and a corresponding algorithm toward estimating the optimal sampling plan. The algorithm is easy to implement in contexts where the underlying distributions governing the performance measures are known or assumed, e.g., the underlying distributions are normal or Bernoulli. The normal context is particularly relevant since a substantial portion of the corresponding literature in the unconstrained context makes a normality assumption. In the absence of such distributional knowledge or assumption, the proposed framework inspires an approximate algorithm derived through an approximation of the rate function using Taylor's Theorem [Rudin, 1976, p. 110].

C.3 For the specific context involving performance measures constructed using normal random variables, we use numerical examples to demonstrate where and to what extent our only competitor in the normal context (OCBA-CO) is suboptimal. There currently appear to be no competitors to the proposed framework for more general contexts.

3. Preliminaries

In this section, we define notation, conventions, and key assumptions used in the paper.

3.1 Notation and Conventions

For notational convenience, we use $i \leq r$ and $j \leq s$ as shorthand for $i = 1, \ldots, r$ and $j = 1, \ldots, s$. Also, we refer to the feasible system with the lowest objective value as system 1. We partition the set of r systems into the following four mutually exclusive and collectively exhaustive subsets.

$1 := \arg\min_i \{h_i : g_{ij} \geq \gamma_j \text{ for all } j \leq s\}$ is the unique best feasible system;

$\Gamma := \{i : g_{ij} \geq \gamma_j \text{ for all } j \leq s,\ i \neq 1\}$ is the set of suboptimal feasible systems;

$\mathcal{S}_b := \{i : h_1 \geq h_i \text{ and } g_{ij} < \gamma_j \text{ for at least one } j \leq s\}$ is the set of infeasible systems that have *better* (lower) objective values than system 1; and

$\mathcal{S}_w := \{i : h_1 < h_i \text{ and } g_{ij} < \gamma_j \text{ for at least one } j \leq s\}$ is the set of infeasible systems that have *worse* (higher) objective values than system 1.

The partitioning of the suboptimal systems into the sets Γ, $\mathcal{S}_b$ and $\mathcal{S}_w$ implies that to be falsely selected as the best feasible system, systems in Γ must "pretend" to be optimal, systems in $\mathcal{S}_b$ must "pretend" to be feasible, and systems in $\mathcal{S}_w$ must "pretend" to be both optimal and feasible. As will become evident, this partitioning is strategic and facilitates analyzing the behavior of the false selection probability.

We use the following notation to distinguish between constraints on which the system is classified as feasible or infeasible.

$\mathcal{C}_F^i := \{j : g_{ij} \geq \gamma_j\}$ is the set of constraints satisfied by system i; and

$\mathcal{C}_I^i := \{j : g_{ij} < \gamma_j\}$ is the set of constraints not satisfied by system i.

We interpret the minimum over the empty set as infinity [see, e.g., Dembo and Zeitouni, 1998, p. 127], and we likewise interpret the union over the empty set as an event having probability zero. We interpret the intersection over the empty set as the certain event, that is, an event having probability one. Also, we say that a sequence of sets $\mathcal{A}_m$ converges to the set $\mathcal{A}$, denoted $\mathcal{A}_m \to \mathcal{A}$, if for large enough m the symmetric difference $(\mathcal{A}_m \cap \mathcal{A}^c) \cup (\mathcal{A} \cap \mathcal{A}_m^c) = \emptyset$.

To aid readability, we have adopted the following notational convention throughout the paper: lower-case letters denote fixed values; upper-case letters denote random variables; upper-case Greek or script letters denote fixed sets; estimated (random) quantities are accompanied by a "hat," e.g., $\hat{H}_1$ estimates the fixed value h_1; optimal values have an asterisk, e.g., x^*.

3.2 Assumptions

To estimate the unknown quantities h_i and g_{ij}, we assume we may obtain replicates of the output random variables $(H_i, G_{i1}, \ldots, G_{is})$ from each system. We make the following further assumptions.

Assumption 1. *The output random variables $(H_i, G_{i1}, \ldots, G_{is})$ are mutually independent for all $i \leq r$, and for any particular system i, the output random variables $H_i, G_{i1}, \ldots, G_{is}$ are mutually independent.*

While it is possible to relax Assumption 1, we have chosen not to do so in the interest of minimizing distraction from the main thrust of the paper.

To ensure that each system is distinguishable from the quantity on which its potential false evaluation as the "best" system depends, and to ensure that the sets of systems may be correctly estimated with probability one (wp1), we make the following assumption.

Assumption 2. *No system has the same objective value as system 1, and no system lies exactly on any constraint, that is,* $h_1 \neq h_i$ *for all* $i \leq r, i \neq 1$ *and* $g_{ij} \neq \gamma_j$ *for all* $i \leq r, j \leq s$.

Assumptions of this type also appear in Glynn and Juneja [2004] and Szechtman and Yücesan [2008].

We use observations of the output random variables to form estimators $\hat{H}_i = (\alpha_i n)^{-1} \sum_{k=1}^{\alpha_i n} H_{ik}$ and $\hat{G}_{ij} = (\alpha_i n)^{-1} \sum_{k=1}^{\alpha_i n} G_{ijk}$ of h_i and g_{ij}, respectively, where $\alpha_i > 0$ denotes the proportion of the total sample n which is allocated to system i. Let

$$\Lambda_i^{\hat{H}}(\theta) = \log E[e^{\theta \hat{H}_i}] \text{ and } \Lambda_{ij}^{\hat{G}}(\theta) = \log E[e^{\theta \hat{G}_{ij}}]$$

be the cumulant generating functions of $\hat{H}_i$ and $\hat{G}_{ij}$, respectively. Let the effective domain of a function $f(\cdot)$ be denoted by $\mathcal{D}_f = \{x : f(x) < \infty\}$, and its interior by $\mathcal{D}_f^\circ$. As is usual in LD contexts, we make the following assumption.

Assumption 3. *For each system* i *and for each constraint* j *of system* i,

(1) the limits $\Lambda_i^H(\theta) = \lim_{n\to\infty} \frac{1}{\alpha_i n} \Lambda_i^{\hat{H}}(\alpha_i n \theta)$ *and* $\Lambda_{ij}^G(\theta) = \lim_{n\to\infty} \frac{1}{\alpha_i n} \Lambda_{ij}^{\hat{G}}(\alpha_i n \theta)$ *exist as extended real numbers for all* θ;

(2) the origin belongs to the interior of $\mathcal{D}_{\Lambda_i^H}$ *and* $\mathcal{D}_{\Lambda_{ij}^G}$, *that is,* $0 \in \mathcal{D}^\circ_{\Lambda_i^H}$ *and* $0 \in \mathcal{D}^\circ_{\Lambda_{ij}^G}$;

(3) $\Lambda_i^H(\theta)$ *and* $\Lambda_{ij}^G(\theta)$ *are strictly convex and* C^∞ *on* $\mathcal{D}^\circ_{\Lambda_i^H}$ *and* $\mathcal{D}^\circ_{\Lambda_{ij}^G}$, *respectively;*

(4) $\Lambda_i^H(\theta)$ *and* $\Lambda_{ij}^G(\theta)$ *are steep, that is, for any sequence* $\{\theta_n\} \in \mathcal{D}_{\Lambda_i^H}$ *that converges to a boundary point of* $\mathcal{D}_{\Lambda_i^H}$, $\lim_{n\to\infty} |\Lambda_i^{H\prime}(\theta_n)| = \infty$, *and likewise, for* $\{\theta_n\} \in \mathcal{D}_{\Lambda_{ij}^G}$ *converging to a boundary point of* $\mathcal{D}_{\Lambda_{ij}^G}$, $\lim_{n\to\infty} |\Lambda_{ij}^{G\prime}(\theta_n)| = \infty$.

Assumption 3 implies that $\hat{H}_i \to h_i$ wp1 and $\hat{G}_{ij} \to g_{ij}$ wp1 [see Bucklew, 2003, Remark 3.2.1]. Furthermore, Assumption 3 ensures that $\hat{H}_i$ and $\hat{G}_{ij}$ satisfy the large deviations principle [Dembo and Zeitouni, 1998, p. 44] with good rate functions $I_i(x) = \sup_{\theta\in\mathbb{R}}\{\theta x - \Lambda_i^H(\theta)\}$ and $J_{ij}(y) = \sup_{\theta\in\mathbb{R}}\{\theta y - \Lambda_{ij}^G(\theta)\}$. Assumption 3(3) is stronger than what is needed for the Gärtner-Ellis theorem to hold. However, we require $\Lambda_i^H(\theta)$ and $\Lambda_{ij}^G(\theta)$ to be strictly convex and C^∞ on $\mathcal{D}^\circ_{\Lambda_i^H}$ and $\mathcal{D}^\circ_{\Lambda_{ij}^G}$, respectively, so that $I_i(x)$ and $J_{ij}(y)$ are strictly convex and C^∞ for $x \in \mathcal{F}_i^{H\circ} = \text{int}\{\Lambda_i^{H\prime}(\theta) : \theta \in \mathcal{D}^\circ_{\Lambda_i^H}\}$ and $y \in \mathcal{F}_{ij}^{G\circ} = \text{int}\{\Lambda_{ij}^{G\prime}(\theta) : \theta \in \mathcal{D}^\circ_{\Lambda_{ij}^G}\}$, respectively.

Let $h_\ell = \arg\min_i\{h_i\}$ and let $h_u = \arg\max_i\{h_i\}$. We further assume,

Assumption 4. *(1) the interval* $[h_\ell, h_u] \subset \cap_{i=1}^r \mathcal{F}_i^{H\circ}$, *and (2)* $\gamma_j \in \cap_{i=1}^r \mathcal{F}_{ij}^{G\circ}$ *for all* $j \leq s$.

As in Glynn and Juneja [2004], Assumption 4(1) ensures that $\hat{H}_i$ may take any value in the interval $[h_\ell, h_u]$ and that $P(\hat{H}_1 \geq \hat{H}_i) > 0$ for $2 \leq i \leq r$. Assumption 4(2) ensures there is a nonzero probability that each system will be deemed feasible or infeasible on any of its constraints. Specifically, it ensures there is a nonzero probability that an infeasible system will be estimated feasible and that system 1 will be estimated infeasible. Thus $P(\cap_{j \in \mathcal{C}_I^i} \hat{G}_{ij} \geq \gamma_j) > 0$ for $i \in \mathcal{S}_b \cup \mathcal{S}_w$ and $P(\hat{G}_{1j} < \gamma_j) > 0$ for all $j \leq s$.

4. Rate Function of Probability of False Selection

The false selection (FS) event is the event that the actual best feasible system, system 1, is not the estimated best feasible system. More specifically, FS is the event that system 1 is incorrectly estimated infeasible on *any* of its constraints, or that system 1 is estimated feasible on all of its constraints but another system, also estimated feasible on all of its constraints, has the best estimated-objective value. Let $\bar{\Gamma}$ be the set of estimated-feasible systems, excluding system 1, that is, $\bar{\Gamma} = \{i : \hat{G}_{ij} \geq \gamma_j \text{ for all } j \leq s, i \neq 1\}$. Then formally, the probability of false selection is

$$\begin{aligned} P\{FS\} &= P\{(\cup_j \hat{G}_{1j} < \gamma_j) \cup ((\cap_j \hat{G}_{1j} \geq \gamma_j) \cap (\hat{H}_1 \geq \min_{i \in \bar{\Gamma}} \hat{H}_i))\} \\ &= P\{\cup_j \hat{G}_{1j} < \gamma_j\} + P\{(\cap_j \hat{G}_{1j} \geq \gamma_j) \cap (\cup_{i \in \bar{\Gamma}} \hat{H}_1 \geq \hat{H}_i)\} \\ &= P\{FS_1\} + P\{FS_2\}. \end{aligned} \tag{1}$$

In the following Theorems 1 and 2, we individually derive the rate functions for $P\{FS_1\}$ and $P\{FS_2\}$ appearing in equation (1).

First let us consider the rate function for $P\{FS_1\}$, the probability that system 1 is declared infeasible on any of its constraints. Theorem 1 establishes the asymptotic behavior of $P\{FS_1\}$ as the rate function corresponding to the constraint on system 1 that is most likely to be declared unsatisfied.

Theorem 1. *The rate function for $P\{FS_1\}$ is given by*

$$-\lim_{n\to\infty} \frac{1}{n} \log P\{FS_1\} = \min_j \alpha_1 J_{1j}(\gamma_j).$$

Proof. We find the following upper and lower bounds for $P\{FS_1\}$,

$$\max_j P\{\hat{G}_{1j} < \gamma_j\} \leq P\{\cup_j \hat{G}_{1j} < \gamma_j\} \leq \; s \max_j P\{\hat{G}_{1j} < \gamma_j\}.$$

To find the rate function for $\max_j P\{\hat{G}_{1j} < \gamma_j\}$, we apply Proposition 5 (see Appendix) to find

$$\lim_{n\to\infty} \frac{1}{n} \log \max_j P\{\hat{G}_{1j} < \gamma_j\} = \max_j \lim_{n\to\infty} \frac{1}{n} \log P\{\hat{G}_{1j} < \gamma_j\}.$$

By the Gärtner-Ellis Theorem and Assumption 1,

$$\lim_{n\to\infty} \frac{1}{n} \log P\{FS_1\} = \max_j \lim_{n\to\infty} \frac{1}{n} \log P\{\hat{G}_{1j} < \gamma_j\} = -\min_j \alpha_j J_{1j}(\gamma_j). \qquad \square$$

Theorem 1 implies that the rate function for $P\{FS_1\}$ is determined by the constraint that is most likely to qualify system 1 as infeasible. Under our assumptions and with logic similar that given in the proof of Theorem 1, it can be shown that for any system i with constraint j, the rate function for the probability that system i is estimated infeasible on constraint j is

$$\lim_{n\to\infty} \frac{1}{n} \log P\{\hat{G}_{ij} < \gamma_j\} = -\alpha_i J_{ij}(\gamma_j).$$

We now consider $P\{FS_2\}$. Since the probability that system 1 is estimated feasible tends to one and under the independence assumption (Assumption 1), we have

$$\lim_{n\to\infty} \frac{1}{n} \log P\{(\cap_j \hat{G}_{1j} \geq \gamma_j) \cap (\cup_{i\in\bar{\Gamma}} \hat{H}_1 \geq \hat{H}_i)\} = \lim_{n\to\infty} \frac{1}{n} \log P\{\cup_{i\in\bar{\Gamma}} \hat{H}_1 \geq \hat{H}_i\}. \tag{2}$$

Therefore the rate function of $P\{FS_2\}$ is governed by the rate at which the probability that system 1 is "beaten" by another estimated-feasible system tends to zero. Since the equality in equation (2) always holds, in the remainder of the paper we omit the explicit statement of the event that system 1 is estimated feasible. Since the estimated set of feasible systems $\bar{\Gamma}$ may contain worse feasible systems ($i \in \Gamma$), better infeasible systems ($i \in \mathcal{S}_b$), and worse infeasible systems ($i \in \mathcal{S}_w$), we strategically consider the rate functions for the probability that system 1 is beaten by a system in $\Gamma, \mathcal{S}_b$, or $\mathcal{S}_w$ separately. Theorem 2 states that the rate function of $P\{FS_2\}$ is determined by the slowest-converging probability that system 1 will be "beaten" by an estimated-feasible system from $\Gamma, \mathcal{S}_b$, or $\mathcal{S}_w$.

Theorem 2. *The rate function for $P\{FS_2\}$ is given by the minimum rate function of the probability that system 1 is beaten by an estimated-feasible system that is (i) feasible and worse, (ii) infeasible and better, or (iii) infeasible and worse. That is,*

$$-\lim_{n\to\infty} \frac{1}{n} \log P\{FS_2\} = \min\Bigg(\overbrace{\min_{i\in\Gamma}(\inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x))),}^{\text{system 1 beaten by feasible and worse system}} \quad \overbrace{\min_{i\in\mathcal{S}_b} \alpha_i \sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j),}^{\text{system 1 beaten by infeasible and better system}}$$
$$\underbrace{\min_{i\in\mathcal{S}_w}(\inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x)) + \alpha_i \sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j))}_{\text{system 1 beaten by infeasible and worse system}} \Bigg).$$

Proof. See Appendix. □

Like the intuition behind Theorem 1, that the rate function of $P\{FS_1\}$ is determined by the constraint most likely to disqualify system 1, in Theorem 2, the rate function of $P\{FS_2\}$ is determined by the system most likely to "beat" system 1. However systems in $\Gamma, \mathcal{S}_b$, and $\mathcal{S}_w$ must overcome different obstacles to be declared the best feasible system. Since systems in Γ are truly feasible, they must overcome one obstacle: optimality. The rate function for systems in Γ is thus identical to the unconstrained optimization case presented in Glynn and Juneja [2004] and is determined by the system in Γ best at "pretending" to be optimal. Systems in $\mathcal{S}_b$ are truly better than system 1, but are infeasible. They also have one obstacle to overcome to be selected as best: feasibility. The rate function for systems in $\mathcal{S}_b$ is thus determined by the system in $\mathcal{S}_b$ which is best at "pretending" to be feasible. Since an infeasible system in $\mathcal{S}_b$ must falsely be declared feasible on *all* of its infeasible constraints, the rate functions for the infeasible constraints simply add up inside the overall rate function for each system in $\mathcal{S}_b$. Systems in $\mathcal{S}_w$ are worse and infeasible, so two obstacles must be overcome: optimality and feasibility. The rate function for systems in $\mathcal{S}_w$ is thus determined by the system that is best at "pretending" to be optimal and feasible, and there are two terms added in the rate function corresponding to optimality and feasibility.

We will now combine the results for $P\{FS_1\}$ and $P\{FS_2\}$ to derive the rate function for $P\{FS\}$. Recalling from (1) that $P\{FS\} = P\{FS_1\} + P\{FS_2\}$, the overall rate function for the probability of false selection is governed by the minimum of the rate functions for $P\{FS_1\}$ and $P\{FS_2\}$.

Theorem 3. *The rate function for the probability of false selection, that is, the probability that we return to the user a system other than system 1 is given by*

$$-\lim_{n\to\infty}\frac{1}{n}\log P\{FS\} = \min\Bigg(\overbrace{\min_j \alpha_1 J_{1j}(\gamma_j),}^{\substack{\textit{system 1}\\ \textit{estimated infeasible}}}$$

$$\overbrace{\underbrace{\min_{i\in\Gamma}(\inf_x(\alpha_1 I_1(x)+\alpha_i I_i(x))),}_{\substack{\textit{system 1 beaten by}\\ \textit{feasible and worse system}}}\ \underbrace{\min_{i\in\mathcal{S}_b}\alpha_i\sum_{j\in\mathcal{C}_I^i}J_{ij}(\gamma_j),}_{\substack{\textit{system 1 beaten}\\ \textit{by infeasible and}\\ \textit{better system}}}\ \underbrace{\min_{i\in\mathcal{S}_w}(\inf_x(\alpha_1 I_1(x)+\alpha_i I_i(x))+\alpha_i\sum_{j\in\mathcal{C}_I^i}J_{ij}(\gamma_j))}_{\substack{\textit{system 1 beaten by}\\ \textit{infeasible and worse system}}}}^{\textit{system 1 estimated feasible}}\Bigg).$$

Theorem 3 asserts that the overall rate function of the probability of false selection is determined by the most likely of the following four events: (i) system 1 is incorrectly declared infeasible on one of its constraints; (ii) a feasible and worse system is correctly declared feasible, but incorrectly declared best; (iii) an infeasible and better system is correctly declared better, but incorrectly declared feasible; (iv) an infeasible and worse system is incorrectly declared feasible and best. This result is intuitive since we expect an unlikely event to happen in the most likely way.

5. Optimal Allocation Strategy

In this section, we derive an optimal allocation strategy that asymptotically minimizes the probability of false selection. From Theorem 3, an asymptotically optimal allocation strategy will result from maximizing the rate at which $P\{FS\}$ tends to zero as a function of $\boldsymbol{\alpha}$. Thus we wish to allocate the α_i's to solve the following optimization problem:

$$\begin{aligned} \max \quad & \min\Bigg(\min_j \alpha_1 J_{1j}(\gamma_j),\ \min_{i\in\Gamma}\left(\inf_x(\alpha_1 I_1(x)+\alpha_i I_i(x))\right),\ \min_{i\in\mathcal{S}_b}\alpha_i \sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j), \\ & \min_{i\in\mathcal{S}_w}\Bigg(\inf_x(\alpha_1 I_1(x)+\alpha_i I_i(x)) + \alpha_i \sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j)\Bigg)\Bigg) \qquad (3) \\ s.t. \quad & \sum_{i=1}^{r}\alpha_i = 1, \quad \alpha_i \geq 0. \end{aligned}$$

By Glynn and Juneja [2006], $\inf_x(\alpha_1 I_1(x)+\alpha_i I_i(x))$ is a concave, strictly increasing, C^∞ function of α_1 and α_i. Let $x(\alpha_1,\alpha_i) = \arg\inf_x(\alpha_1 I_1(x)+\alpha_i I_i(x))$. As Glynn and Juneja [2006] demonstrate, for $\alpha_1 > 0$ and $\alpha_i > 0$, $x(\alpha_1,\alpha_i)$ is a C^∞ function of α_1 and α_i. Likewise, the linear functions $\alpha_1 J_{1j}(\gamma_j)$ and $\alpha_i \sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j)$ and the sum $\inf_x(\alpha_1 I_1(x)+\alpha_i I_i(x))+\alpha_i\sum_{j\in\mathcal{C}_I^i} J_i(\gamma_j)$ are also concave, strictly increasing C^∞ functions of α_1 and α_i. Since the minimum of concave, strictly increasing functions is also concave and strictly increasing, the problem in (3) is a concave maximization problem. Equivalently, we may rewrite the problem in (3) as the following Problem Q.

$$\begin{aligned} \text{Problem } Q: \qquad \max \quad z \quad s.t.& \\ \alpha_1 J_{1j}(\gamma_j) &\geq z,\ j\in\mathcal{C}_F^1 \\ \alpha_1 I_1(x(\alpha_1,\alpha_i)) + \alpha_i I_i(x(\alpha_1,\alpha_i)) &\geq z,\ i\in\Gamma \\ \alpha_i \sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j) &\geq z,\ i\in\mathcal{S}_b \\ \alpha_1 I_1(x(\alpha_1,\alpha_i)) + \alpha_i I_i(x(\alpha_1,\alpha_i)) + \alpha_i \sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j) &\geq z,\ i\in\mathcal{S}_w \\ \sum_{i=1}^{r}\alpha_i &= 1,\ \alpha_i\geq 0. \end{aligned}$$

Since Problem Q is a strictly concave, continuous function of $\boldsymbol{\alpha}$ on a compact set, a unique solution exists. Proposition 1 states this result, without a formal proof.

Proposition 1. *There exists a unique solution* $\boldsymbol{\alpha^*} = \{\alpha_1^*, \alpha_2^*, \ldots, \alpha_r^*\}$ *to Problem* Q*, with optimal value* z^**.*

Let us define Problem Q^* by replacing the inequality constraints corresponding to systems in $\Gamma, \mathcal{S}_b$, and $\mathcal{S}_w$ with equality constraints, and forcing each α_i to be strictly greater than zero.

$$\begin{aligned} \text{Problem } Q^*: \qquad \max \quad z \quad s.t.& \\ \alpha_1 J_{1j}(\gamma_j) \geq z,\ j \in \mathcal{C}_F^1& \\ \alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i)) = z,\ i \in \Gamma& \\ \alpha_i \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j) = z,\ i \in \mathcal{S}_b& \\ \alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i)) + \alpha_i \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j) = z,\ i \in \mathcal{S}_w& \\ \sum_{i=1}^{r} \alpha_i = 1,\ \alpha_i > 0.& \end{aligned}$$

We present the following proposition regarding the equivalence of Problem Q and Problem Q^*.

Proposition 2. *Problems* Q *and* Q^* *are equivalent, that is, Problem* Q^* *has the unique solution* $\boldsymbol{\alpha}^*$ *with optimal value* z^**.*

Proof. First, we note that for $\alpha_i = 1/r, i \leq r$, we can have $z > 0$ in Q. Therefore $\alpha_i = 0$ for $i \in \{1\} \cup \mathcal{S}_b$ is suboptimal since $z = 0$. Now consider $\alpha_i = 0$ for $i \in \Gamma \cup \mathcal{S}_w$. In this case, the constraints for $i \in \Gamma \cup \mathcal{S}_w$ reduce to $\alpha_1 \inf_x I_1(x) = \alpha_1 I_1(h_1) = 0$, and hence $z = 0$. Therefore in Problem Q, we must have $\alpha_i^* > 0$ for all $i \leq r$.

Denoting the dual variables ν and $\boldsymbol{\lambda} = (\lambda_j^1 \geq 0, \lambda_i \geq 0 : j = 1, \ldots, |\mathcal{C}_F^1|, i = 2, \ldots, r)$, we solve for the KKT conditions. We note that since $x(\alpha_1, \alpha_i)$ solves $\alpha_1 I_1^{'}(x) + \alpha_i I_i^{'}(x) = 0$, we have $\frac{\partial}{\partial \alpha_1}\big(\alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i))\big) = I_1(x(\alpha_1, \alpha_i))$ and $\frac{\partial}{\partial \alpha_i}\big(\alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i))\big) = I_i(x(\alpha_1, \alpha_i))$ [see Glynn and Juneja, 2004]. Then we have the following stationarity conditions,

$$\sum_{j=1}^{|\mathcal{C}_F^1|} \lambda_j^1 + \sum_{i=2}^{r} \lambda_i = 1 \tag{4}$$

$$\sum_{j \in \mathcal{C}_F^1} \lambda_j^1 J_{1j}(\gamma_j) + \sum_{i \in \Gamma \cup \mathcal{S}_w} \lambda_i I_1(x(\alpha_1^*, \alpha_i^*)) = \nu \tag{5}$$

$$\lambda_i I_i(x(\alpha_1^*, \alpha_i^*)) = \nu,\ i \in \Gamma \tag{6}$$

$$\lambda_i \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j) = \nu,\ i \in \mathcal{S}_b \tag{7}$$

$$\lambda_i\big[I_i(x(\alpha_1^*,\alpha_i^*)) + \sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j)\big] = \nu,\ i\in\mathcal{S}_w, \tag{8}$$

and the complementary slackness conditions,

$$\lambda_j^1\big[\alpha_1^* J_{1j}(\gamma_j) - z\big] = 0,\ j\in\mathcal{C}_F^1 \tag{9}$$

$$\lambda_i\big[\alpha_1^* I_1(x(\alpha_1^*,\alpha_i^*)) + \alpha_i^* I_i(x(\alpha_1^*,\alpha_i^*)) - z\big] = 0,\ i\in\Gamma \tag{10}$$

$$\lambda_i\Bigg[\alpha_i^* \sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j) - z\Bigg] = 0,\ i\in\mathcal{S}_b \tag{11}$$

$$\lambda_i\Bigg[\alpha_1^* I_1(x(\alpha_1^*,\alpha_i^*)) + \alpha_i^* I_i(x(\alpha_1^*,\alpha_i^*)) + \alpha_i^* \sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j) - z\Bigg] = 0,\ i\in\mathcal{S}_w. \tag{12}$$

Equation (4) implies that at least one $\lambda_i > 0$. Suppose $\lambda_i = 0$ for some $i\in\Gamma\cup\mathcal{S}_b\cup\mathcal{S}_w$. Since $\alpha_i > 0$ for all $i \le r$, the rate functions in equations (6)–(8) are strictly greater than zero, which implies $\nu = 0, \lambda_i = 0$ for all $i\in\Gamma\cup\mathcal{S}_b\cup\mathcal{S}_w$, and $\sum_{j=1}^{|\mathcal{C}_F^1|}\lambda_j^1 = 1$. Therefore at least one $\lambda_j^1 > 0$. Then in equation (5), it must be the case that for $\lambda_j^1 > 0$, the corresponding $J_{1j}(\gamma_j) = 0$. However we have a contradiction since by assumption, $J_{1j}(\gamma_j) > 0$ for all $j\in|\mathcal{C}_F^1|$. Therefore $\lambda_i > 0$ for all $i\in\Gamma\cup\mathcal{S}_b\cup\mathcal{S}_w$.

Since $\lambda_i > 0$ in equations (10)–(12), then complementary slackness implies each of these constraints is binding. Therefore we may replace the inequality constraints corresponding to $i\in\Gamma\cup\mathcal{S}_b\cup\mathcal{S}_w$ in Problem Q with equality constraints in Problem Q^*. □

The structure of the identical Problem Q^* lends intuition to the structure of the optimal allocation, as noted in the following steps: (i) Solve a relaxation of Problem Q^* without the feasibility constraint for system 1. Let this problem be called Problem $\widetilde{Q}^*$, and let $\tilde{z}^*$ be the optimal value at the optimal solution $\tilde{\boldsymbol{\alpha}}^* = (\tilde{\alpha}_1^*,\ldots,\tilde{\alpha}_r^*)$ to Problem $\widetilde{Q}^*$. (ii) Check if the feasibility constraint for system 1 is satisfied by the solution $\tilde{\boldsymbol{\alpha}}^*$. If the feasibility constraint is satisfied, $\tilde{\boldsymbol{\alpha}}^*$ is the optimal solution for Problem Q^*. Otherwise, (iii) force the feasibility constraint to be binding. The steps (i), (ii), and (iii) are equivalent to solving one of two systems of nonlinear equations, as identified by the KKT conditions of Problems Q^* and $\tilde{Q}^*$. Theorem 4 asserts this formally.

Theorem 4. *Let the set of suboptimal feasible systems* Γ *be non-empty, and define Problem* $\tilde{Q}^*$ *as Problem* Q^* *but with the inequality constraint relaxed. Let* $(\boldsymbol{\alpha}^*, z^*)$ *and* $(\tilde{\boldsymbol{\alpha}}^*, \tilde{z}^*)$ *denote the unique optimal solution and optimal value pairs for Problems* Q^* *and* $\widetilde{Q}^*$*, respectively. Consider the conditions,*

C0. $\sum_{i=1}^r \alpha_i = 1$, $\boldsymbol{\alpha} > 0$, *and*

$z = \alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i)) = \alpha_k \sum_{j \in \mathcal{C}_I^k} J_{kj}(\gamma_j)$

$= \alpha_1 I_1(x(\alpha_1, \alpha_\ell)) + \alpha_\ell \big[I_\ell(x(\alpha_1, \alpha_\ell)) + \sum_{j \in \mathcal{C}_I^\ell} J_{\ell j}(\gamma_j)\big]$, *for all* $i \in \Gamma, k \in \mathcal{S}_b,\ \ell \in \mathcal{S}_w$,

C1. $\displaystyle\sum_{i \in \Gamma} \frac{I_1(x(\alpha_1, \alpha_i))}{I_i(x(\alpha_1, \alpha_i))} + \sum_{i \in \mathcal{S}_w} \frac{I_1(x(\alpha_1, \alpha_i))}{I_i(x(\alpha_1, \alpha_i)) + \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j)} = 1,$

C2. $\min_{j \in \mathcal{C}_F^1} \alpha_1 J_{1j}(\gamma_j) = z.$

Then (i) $\tilde{\boldsymbol{\alpha}}^*$ *solves C0 and C1 and* $\min_{j \in \mathcal{C}_F^1} \tilde{\alpha}_1^* J_{1j}(\gamma_j) \geq \tilde{z}^*$ *if and only if* $\tilde{\boldsymbol{\alpha}}^* = \boldsymbol{\alpha}^*$*; and*

(ii) $\boldsymbol{\alpha}^*$ *solves C0 and C2 and* $\min_{j \in \mathcal{C}_F^1} \tilde{\alpha}_1^* J_{1j}(\gamma_j) < \tilde{z}^*$ *if and only if* $\boldsymbol{\alpha}^* \neq \tilde{\boldsymbol{\alpha}}^*$*.*

Proof. Due to the structure of Problem Q, the KKT conditions are necessary and sufficient for global optimality. From prior results, we recall that the solutions to Problems Q, Q^*, and $\widetilde{Q}^*$ exist, and that condition C0 holds for the solutions $\boldsymbol{\alpha}^*$ and $\tilde{\boldsymbol{\alpha}}^*$.

We now simplify the KKT equations for Problem Q for use in the remainder of the proof. Since we found that $\lambda_i > 0$ for all $i \in \Gamma \cup \mathcal{S}_b \cup \mathcal{S}_w$ in the proof of Proposition 2, it follows that $\nu > 0$. Dividing (5) by ν and appropriately substituting in values from (6)–(8), we find

$$\frac{\sum_{j \in \mathcal{C}_F^1} \lambda_j^1 J_{1j}(\gamma_j)}{\nu} + \sum_{i \in \Gamma} \frac{I_1(x(\alpha_1^*, \alpha_i^*))}{I_i(x(\alpha_1^*, \alpha_i^*))} + \sum_{i \in \mathcal{S}_w} \frac{I_1(x(\alpha_1^*, \alpha_i^*))}{I_i(x(\alpha_1^*, \alpha_i^*)) + \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j)} = 1. \tag{13}$$

By a similar logic to that given in the proof of Proposition 2 and the simplification provided in (13), omitting terms with λ_j^1 in equation (13) yields condition C1 as a KKT condition for Problem $\widetilde{Q}^*$. Taken together, C0 and C1 create a fully-specified system of equations that form the KKT conditions for Problem $\widetilde{Q}^*$. A solution $\boldsymbol{\alpha}$ is thus optimal to Problem $\widetilde{Q}^*$ if and only if it solves C0 and C1.

Proof of Claim (i). ($\Rightarrow$) Suppose $\tilde{\boldsymbol{\alpha}}^*$ solves C0 and C1, and $\min_{j \in \mathcal{C}_F^1} \tilde{\alpha}_1^* J_{1j}(\gamma_j) \geq \tilde{z}^*$. Let $\mathcal{D}(Q^*)$ and $\mathcal{D}(\widetilde{Q}^*)$ denote the feasible regions of Problems Q^* and $\widetilde{Q}^*$, respectively. Then $\tilde{\boldsymbol{\alpha}}^* \in \mathcal{D}(Q^*)$. Since the objective functions of Problems Q^* and $\widetilde{Q}^*$ are identical, and $\mathcal{D}(Q^*) \subset \mathcal{D}(\widetilde{Q}^*)$, we know that $z^* \leq \tilde{z}^*$. Therefore $\tilde{\boldsymbol{\alpha}}^* \in \mathcal{D}(Q^*)$ implies $\tilde{\boldsymbol{\alpha}}^*$ is the optimal solution to Problem Q^*, and by the uniqueness of the optimal solution, $\tilde{\boldsymbol{\alpha}}^* = \boldsymbol{\alpha}^*$.

($\Leftarrow$) Now suppose $\tilde{\boldsymbol{\alpha}}^* = \boldsymbol{\alpha}^*$. Since $\tilde{\boldsymbol{\alpha}}^*$ is the optimal solution to Problem $\widetilde{Q}^*$, then $\tilde{\boldsymbol{\alpha}}^*$ solves C0 and C1. Further, since $\boldsymbol{\alpha}^*$ is the optimal solution to Problem Q, $\boldsymbol{\alpha}^* = \tilde{\boldsymbol{\alpha}}^* \in \mathcal{D}(Q^*)$. Therefore $\min_{j \in \mathcal{C}_F^1} \tilde{\alpha}_1^* J_{1j}(\gamma_j) \geq \tilde{z}^*$.

Proof of Claim (ii). ($\Rightarrow$) Let us suppose that $\boldsymbol{\alpha}^*$ solves C0 and C2, and $\min_{j \in \mathcal{C}_F^1} \tilde{\alpha}_1^* J_{1j}(\gamma_j) < \tilde{z}^*$. Then $\tilde{\boldsymbol{\alpha}}^* \notin \mathcal{D}(Q^*)$, and therefore $\tilde{\boldsymbol{\alpha}}^* \neq \boldsymbol{\alpha}^*$.

($\Leftarrow$) By prior arguments, C0 holds for $\boldsymbol{\alpha}^*$ and $\tilde{\boldsymbol{\alpha}}^*$. Now suppose $\boldsymbol{\alpha}^* \neq \tilde{\boldsymbol{\alpha}}^*$, which implies $\tilde{\boldsymbol{\alpha}}^* \notin \mathcal{D}(Q^*)$. Then it must be the case that $\min_{j \in \mathcal{C}_F^1} \tilde{\alpha}_1^* J_{1j}(\gamma_j) < \tilde{z}^*$. Further, since $\tilde{\boldsymbol{\alpha}}^*$ uniquely solves C0 and C1, $\boldsymbol{\alpha}^* \neq \tilde{\boldsymbol{\alpha}}^*$ implies that C1 does not hold for $\boldsymbol{\alpha}^*$. Therefore when solving Problem Q, it must be the case that $\lambda_j^1 > 0$ for at least one $j \in \mathcal{C}_F^1$ in equation (13). By the complementary slackness conditions in equation (9), $\min_{j \in \mathcal{C}_F^1} \alpha_1^* J_{1j}(\gamma_j) = \tilde{z}^*$, and hence C2 holds for $\boldsymbol{\alpha}^*$. $\square$

Theorem 4 implies that, since a solution to Problem Q^* always exists, an optimal solution to Problem Q can be obtained as the solution to one of the two sets of nonlinear equations C0 and C1 or C0 and C2. We state the procedure implicit in Theorem 4 as Algorithm 1.

Algorithm 1 Conceptual Algorithm to Solve for $\boldsymbol{\alpha}^*$

```
Solve the nonlinear system C0, C1 to obtain α̃* and z̃*.
if min_j α̃*_1 J_1j(γ_j) ≥ z̃* then
    return α* = α̃*.
else
    Solve the nonlinear system C0, C2 to obtain α*.
    return α*.
end if
```

Theorem 4 assumes that we have at least one system in Γ. In the event that Γ is empty, conditions C0 and C1 may not form a fully-specified system of equations (e.g., Γ and $\mathcal{S}_w$ are empty), or may not have a solution. In such a case, C0 and C2 provide the optimal allocation. When the sets $\mathcal{S}_b$ and $\mathcal{S}_w$ are empty but Γ is nonempty, Theorem 4 reduces to the result presented in Glynn and Juneja [2004].

6. Consistency and Implementation

In practice, the rate functions in Algorithm 1 are unavailable and must be estimated. Therefore with a view toward implementation, we address consistency of estimators in this section. Specifically, we first show that the important sets, $\{1\}, \Gamma, \mathcal{S}_b, \mathcal{S}_w, \mathcal{C}_F^i$ and $\mathcal{C}_I^i$, can be estimated consistently, that is, they can be identified correctly as simulation effort tends to infinity. Next, we demonstrate that the optimal allocation estimator, identified by using estimated rate functions in Algorithm 1, is a consistent estimator of the true optimal allocation $\boldsymbol{\alpha}^*$. These generic consistency results inspire the sequential algorithm presented in Section 6.2, which is easily implementable at least in contexts where the distribution families underlying the rate functions are known or assumed.

6.1 Generic Consistency Results

To simplify notation, let each system be allocated m samples, where we explicitly denote the dependence of the estimators on m in this section. Suppose we have at our disposal consistent estimators $\hat{I}_i^m(x), \hat{J}_{ij}^m(y), i \leq r, j \leq s$ of the corresponding rate functions $I_i(x), J_{ij}(y), i \leq r, j \leq s$. Such consistent estimators are easy to construct when the distributional families underlying the true rate functions $I_i(x), J_{ij}(y), i \leq r, j \leq s$ are known or assumed. For example, suppose $H_{ik}, k = 1, 2, \dots, m$ are simulation observations of the objective function of the ith system, assumed to be resulting from a normal distribution with unknown mean h_i and unknown variance $\sigma_{h_i}^2$. The obvious consistent estimator for the rate function $I_i(x) = \frac{(x-h_i)^2}{2\sigma_{h_i}^2}$ is then $\hat{I}_i^m(x) = \frac{(x-\hat{H}_i)^2}{2\hat{\sigma}_i^2}$, where $\hat{H}_i$ and $\hat{\sigma}_{h_i}$ are the sample mean and sample standard deviation of $H_{ik}, k = 1, 2, \dots, m$ respectively. In the more general case where the distributional family is unknown or not assumed, the rate function may be estimated as the Legendre-Fenchel transform of the cumulant generating function estimator

$$\hat{I}_i^m(x) = \sup_\theta(\theta x - \hat{\Lambda}_i^{H,m}(\theta)), \tag{14}$$

where $\hat{\Lambda}_i^{H,m}(\theta) = \log m^{-1} \sum_{k=1}^m \exp(\theta H_{ik})$. In what follows, to preserve generality, our discussion pertains to estimators of the type displayed in (14). By arguments analogous to those in Glynn and Juneja [2004] and under our assumptions, the estimator in (14) is consistent.

Let $(\hat{H}_i(m),\ \hat{G}_{i1}(m),\ \dots,\ \hat{G}_{is}(m)) = \left(\frac{1}{m}\sum_{k=1}^m H_{ik},\ \frac{1}{m}\sum_{k=1}^m G_{i1k},\ \dots,\ \frac{1}{m}\sum_{k=1}^m G_{isk}\right)$ denote the estimators of $(h_i, g_{i1}, \dots, g_{is})$. We define the following notation for estimators of all relevant sets for systems $i \leq r$.

$\hat{1}(m) := \arg\min_i\{\hat{H}_i(m) : \hat{G}_{ij}(m) \geq \gamma_j$ for all $j \leq s\}$ is the estimated best feasible system;

$\widehat{\Gamma}(m) := \{i : \hat{G}_{ij}(m) \geq \gamma_j$ for all $j \leq s, i \neq \hat{1}(m)\ \}$ is the estimated set of suboptimal feasible systems;

$\widehat{\mathcal{S}}_b(m) := \{i : \hat{H}_{\hat{1}(m)}(m) \geq \hat{H}_i(m)$ and $\hat{G}_{ij}(m) < \gamma_j$ for some $j \leq s\}$ is the estimated set of infeasible, better systems;

$\widehat{\mathcal{S}}_w(m) := \{i : \hat{H}_{\hat{1}(m)}(m) < \hat{H}_i(m)$ and $\hat{G}_{ij}(m) < \gamma_j$ for some $j \leq s\}$ is the estimated set of infeasible, worse systems;

$\widehat{\mathcal{C}}_F^i(m) := \{j : \hat{G}_{ij}(m) \geq \gamma_j\}$ is the set of constraints on which system i is estimated feasible;

$\widehat{\mathcal{C}}_I^i(m) := \{j : \hat{G}_{ij}(m) < \gamma_j\}$ is the set of constraints on which system i is estimated infeasible.

Note that $\bar{\Gamma}$ (defined in Section 4) excludes system 1 while $\widehat{\Gamma}(m)$ excludes the *estimated* system 1.

Since Assumption 3 implies $\hat{H}_i(m) \to h_i$ wp1 and $\hat{G}_{ij}(m) \to g_{ij}$ wp1 for all $i \leq r$ and $j \leq s$, and the numbers of systems and constraints are finite, all estimated sets converge to their true counterparts wp1 as $m \to \infty$. (See Section 3.1 for a rigorous definition of the convergence of sets.) Proposition 3 formally states this result.

Proposition 3. *Under Assumption 3, $\hat{1}(m) \to 1$, $\widehat{\Gamma}(m) \to \Gamma$, $\widehat{\mathcal{S}}_b(m) \to \mathcal{S}_b$, $\widehat{\mathcal{S}}_w(m) \to \mathcal{S}_w$, $\widehat{\mathcal{C}}^i_F(m) \to \mathcal{C}^i_F$, and $\widehat{\mathcal{C}}^i_I(m) \to \mathcal{C}^i_I$ wp1 as $m \to \infty$.*

Proof. See Appendix. □

Let $\hat{\boldsymbol{\alpha}}^*(m)$ denote the estimator of the optimal allocation vector $\boldsymbol{\alpha}^*$ obtained by replacing the rate functions $I_i(x), J_{ij}(x), i \leq r, j \leq s$ appearing in conditions C0, C1, and C2 with their corresponding estimators $\hat{I}^m_i(x), \hat{J}^m_{ij}(x), i \leq r, j \leq s$ obtained through sampling, and then using Algorithm 1. Since the search space $\{\boldsymbol{\alpha} : \alpha_i \geq 0, \sum_{i=1}^r \alpha_i = 1\}$ is a compact set, and the estimated (consistent) rate functions can be shown to converge uniformly over the search space, it is no surprise that $\hat{\boldsymbol{\alpha}}^*(m)$ converges to the optimal allocation vector $\boldsymbol{\alpha}^*$ as $m \to \infty$ wp1. Theorem 5 formally asserts this result, with a proof that is a direct application of results found in the stochastic root-finding literature [see, e.g., Pasupathy and Kim, 2010, Theorem 5.7].

Before we state Theorem 5, we state two additional lemmas. We omit the proof of Lemma 1 since it follows very closely along the lines of the proofs presented in Glynn and Juneja [2004].

Lemma 1. *Suppose Assumption 4 holds. Then there exists $\epsilon > 0$ such that $\hat{I}^m_i(x) \to I_i(x)$ as $m \to \infty$ uniformly in $x \in [h_\ell - \epsilon, h_u + \epsilon]$ wp1, for all $i \in \{1\} \cup \Gamma \cup \mathcal{S}_w$.*

Lemma 2. *Let the system of equations C0 and C1 be denoted $f_1(\boldsymbol{\alpha}) = 0$, and let the system of equations C0 and C2 be denoted by $f_2(\boldsymbol{\alpha}) = 0$, where f_1 and f_2 are vector-valued functions with compact support $\sum_{i=1}^r \alpha_i = 1, \boldsymbol{\alpha} \geq 0$. Let the estimators $\hat{F}^m_1(\boldsymbol{\alpha})$ and $\hat{F}^m_2(\boldsymbol{\alpha})$ be the same set of equations as $f_1(\boldsymbol{\alpha})$ and $f_2(\boldsymbol{\alpha})$, respectively, except with all unknown rate functions replaced by their corresponding estimated quantities. If Assumption 4 holds, then the functional sequences $\hat{F}^m_1(\boldsymbol{\alpha}) \to f_1(\boldsymbol{\alpha})$ and $\hat{F}^m_2(\boldsymbol{\alpha}) \to f_2(\boldsymbol{\alpha})$ uniformly in $\boldsymbol{\alpha}$ as $m \to \infty$ wp1.*

Proof. We will prove that the theorem holds in two steps. We first show that $\alpha_1 \hat{I}^m_1(\hat{x}_m(\alpha_1, \alpha_i) + \alpha_i \hat{I}^m_i(\hat{x}_m(\alpha_1, \alpha_i))$ converges uniformly in $\boldsymbol{\alpha}$ as $m \to \infty$ wp1 for all $i \in \Gamma \cup \mathcal{S}_w$, where $\hat{x}_m(\alpha_1, \alpha_i) = \arg\inf_x(\alpha_1 \hat{I}^m_1(x) + \alpha_i \hat{I}^m_i(x))$. Next we show that $\alpha_i \sum_{j \in \mathcal{C}^i_I} \hat{J}^m_{ij}(\gamma_j), i \in \mathcal{S}_b \cup \mathcal{S}_w, j \leq s$ and $\alpha_1 \hat{J}^m_{1j}(\gamma_j), j \in \mathcal{C}^1_F$ converge uniformly in $\boldsymbol{\alpha}$ as $m \to \infty$ wp1. These assertions, together with the observation that we search only in the set $\{\boldsymbol{\alpha} : \sum_{i=1}^r \alpha_i = 1, \alpha_i > 0\}$, and hence $I_i(x(\alpha_1, \alpha_i)) > \delta > 0$, which implies for large enough m, $\hat{I}^m_i(\hat{x}_m(\alpha_1, \alpha_i)) > \delta$, proves the theorem.

By Lemma 1, $\hat{I}_i^m(x) \to I_i(x)$ uniformly in x on $[h_\ell - \epsilon, h_u + \epsilon]$ wp1 for some $\epsilon > 0$. By Glynn and Juneja [2004], $\hat{x}_m(\alpha_1, \alpha_i) \to x(\alpha_1, \alpha_i)$ wp1, where $x(\alpha_1, \alpha_i) = \arg\inf_x(\alpha_1 I_1(x) + \alpha_i \hat{I}_i^m(x)) \in [h_\ell, h_u]$. Therefore for m large enough and for all feasible α_1, α_i, we have $\hat{x}_m(\alpha_1, \alpha_i) \in [h_\ell - \epsilon/2, h_u + \epsilon/2]$ wp1 for all $i \in \{1\} \cup \Gamma \cup \mathcal{S}_w$. It then follows that $\alpha_1 \hat{I}_1^m(\hat{x}_m(\alpha_1, \alpha_i) + \alpha_i \hat{I}_i^m(\hat{x}_m(\alpha_1, \alpha_i))$ converges uniformly in $\boldsymbol{\alpha}$ as $m \to \infty$ wp1, for all $i \in \Gamma \cup \mathcal{S}_w$.

Under Assumption 4, it follows from analogous arguments to those in Glynn and Juneja [2004] that $\hat{J}_{ij}^m(\gamma_j) \to J_{ij}(\gamma_j)$ as $m \to \infty$ wp1, for all $i \in \mathcal{S}_b \cup \mathcal{S}_w$ and $j \le s$. Therefore the terms $\alpha_i \sum_{j \in \mathcal{C}_I^i} \hat{J}_{ij}^m(\gamma_j)$ converge uniformly in $\boldsymbol{\alpha}$ as $m \to \infty$ wp1. Likewise, for all $j \in \mathcal{C}_F^1$, $\alpha_1 \hat{J}_{1j}^m(\gamma_j)$ converges uniformly in $\boldsymbol{\alpha}$ as $m \to \infty$ wp1. □

Theorem 5. *Let the postulates of Lemma 2 hold, and assume* Γ *is nonempty. Then the empirical estimate of the optimal allocation is consistent, that is,* $\hat{\boldsymbol{\alpha}}^*(m) \to \boldsymbol{\alpha}^*$ *as* $m \to \infty$ *wp1.*

Proof. As argued previously, $f_1(\boldsymbol{\alpha})$ and $f_2(\boldsymbol{\alpha})$ are continuous functions of $\boldsymbol{\alpha}$ on a compact set. Further, the solutions $f_1(\boldsymbol{\alpha}) = 0$ and $f_2(\boldsymbol{\alpha}) = 0$ exist. If we replace each rate function in Problem Q with estimated rate functions, these new problems remain continuous, concave maximization problems on a compact set, which attain their maximums. Therefore the systems $\hat{F}_1^m(\boldsymbol{\alpha}) = 0$ and $\hat{F}_2^m(\boldsymbol{\alpha}) = 0$ have a solution for large enough m wp1. By Lemma 2 we also have that $\hat{F}_1^m(\boldsymbol{\alpha}) \to f_1(\boldsymbol{\alpha})$ and $\hat{F}_2^m(\boldsymbol{\alpha}) \to f_2(\boldsymbol{\alpha})$ uniformly in $\boldsymbol{\alpha}$ as $m \to \infty$ wp1. We have thus satisfied all the requirements for convergence of the sample-path solution $\hat{\boldsymbol{\alpha}}^*(m)$ to its true counterpart $\boldsymbol{\alpha}^*$ as $m \to \infty$ wp1 [see Pasupathy and Kim, 2010, Theorem 5.7]. □

6.2 A Sequential Algorithm for Implementation

We conclude this section with a sequential algorithm that naturally stems from the conceptual algorithm (Algorithm 1) outlined in Section 5 and the consistent estimator that we have discussed in the previous section. Algorithm 2 formally outlines this procedure, where we let n be the total simulation budget, and n_i be the total sample expended at system i.

The essential idea in Algorithm 2 is straightforward. At the end of each iteration, the optimal allocation vector is estimated using rate function estimators constructed from samples already gathered from the various systems. Systems are chosen for sampling at the subsequent iteration by using the estimated optimal allocation vector as the sampling distribution.

We emphasize that in a context where the distributional family underlying the simulation observations are known or assumed, the rate function estimators should be estimated (in Step 3) accordingly — by simply estimating the distributional parameters appearing within the expression

Algorithm 2 Sequential Algorithm with Guaranteed Asymptotic Optimal Allocation

Require: Number of pilot samples $b_0 > 0$; number of samples between allocation vector updates $b > 0$.

1: Initialize: collect b_0 samples from each system $i \le r$.
2: Initialize: $n = rb_0, n_i = b_0$. {Initialize total simulation effort and effort for each system.}
3: Update rate function estimators $\hat{I}_i^{n_i}(x), \hat{J}_{ij}^{n_i}(x), i \le r, j \le s$.
4: Solve the system C0, C1 using the updated rate function estimators to obtain $\widehat{\tilde{\boldsymbol{\alpha}}}^*(n)$ and $\widehat{\tilde{z}}^*(n)$.
5: **if** $\min_j \widehat{\tilde{\alpha}}_1^*(n)\hat{J}_{1j}^{n_1}(\gamma_j) \ge \widehat{\tilde{z}}^*(n)$ **then**
6: $\quad\hat{\boldsymbol{\alpha}}^*(n) = \widehat{\tilde{\boldsymbol{\alpha}}}^*(n)$.
7: **else**
8: $\quad$Solve the system C0, C2 using the updated rate function estimators to obtain $\hat{\boldsymbol{\alpha}}^*(n)$.
9: **end if**
10: Collect one sample at each of the systems $X_k, k = 1, 2, \ldots, b$, where the X_k's are iid random variates having probability mass function $\hat{\boldsymbol{\alpha}}^*(n)$ and support $\{1, 2, \ldots, r\}$, and update $n_{X_k} = n_{X_k} + 1$.
11: Set $n = n + b$ and go to Step 3.

for the rate function. Also, Algorithm 2 provides flexibility on how often the optimal allocation vector is re-estimated through the algorithm parameter b. The choice of the parameter b will depend on the particular problem, and specifically, on how expensive the simulation execution is relative to solving the nonlinear systems in Steps 4 and 7. Lastly, as is clear from the algorithm listing, Algorithm 2 relies on fully sequential and simultaneous observation of the objective and constraint functions. Deviation from these assumptions, while interesting, renders the present context inapplicable.

7. Numerical Examples

To illustrate the proposed optimal allocation, we first present a simple numerical example for the case in which the underlying random variables are independent and identically distributed (iid) replicates from a normal distribution. We then compare our proposed optimal allocation to the OCBA-CO allocation presented by Lee et al. [2011].

In what follows, we have used the actual rate functions governing the simulation estimators for analysis. We have followed this route, instead of using the sequential estimator outlined in Algorithm 2, because our primary objective in this section is to understand the asymptotic allocation proposed by our theory, and to highlight its deviation from the asymptotic solution proposed by OCBA-CO. Owing to their routine nature, we have chosen not to include results from our numerical tests demonstrating that the sequential estimator in Algorithm 2 indeed converges to the optimal allocation vector identified by theory.

7.1 Illustration of Proposed Allocation on a Normal Example

Suppose H_i is distributed iid normal$(h_i, \sigma^2_{h_i})$ and G_{ij} is distributed iid normal$(g_{ij}, \sigma^2_{g_{ij}})$ for all $i \le r$, $j \le s$. The relevant rate functions for the normal case are

$$\min_j \alpha_1 J_{1j}(\gamma_j) = \min_j \frac{\alpha_1(\gamma_j - g_{1j})^2}{2\sigma^2_{g_{1j}}},\ i \in \{1\},$$
$$\alpha_1 I_1(x(\alpha_1,\alpha_i)) + \alpha_i I_i(x(\alpha_1,\alpha_i)) = \frac{(h_1 - h_i)^2}{2(\sigma^2_{h_1}/\alpha_1 + \sigma^2_{h_i}/\alpha_i)},\ i \in \Gamma,$$
$$\alpha_i \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j) = \alpha_i \sum_{j \in \mathcal{C}_I^i} \frac{(\gamma_j - g_{ij})^2}{2\sigma^2_{g_{ij}}},\ i \in \mathcal{S}_b,$$

and for $i \in \mathcal{S}_w$,

$$\alpha_1 I_1(x(\alpha_1,\alpha_i)) + \alpha_i I_i(x(\alpha_1,\alpha_i)) + \alpha_i \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j) = \frac{(h_1 - h_i)^2}{2(\sigma^2_{h_1}/\alpha_1 + \sigma^2_{h_i}/\alpha_i)} + \alpha_i \sum_{j \in \mathcal{C}_I^i} \frac{(\gamma_j - g_{ij})^2}{2\sigma^2_{g_{ij}}}.$$

Example 1. Suppose we have $r = 3$ systems and only one constraint, where the H_i's are iid normal$(h_i, \sigma^2_{h_i})$ random variables and the G_i's are iid normal$(g_i, \sigma^2_{g_i})$ random variables for all $i \le r$. Let $\gamma = 0$, and let the mean and variance of each objective and constraint random variable be as given in table 2.

Table 2: Means and variances for Example 1.

System (i)	h_i	$\sigma^2_{h_i}$	g_i	$\sigma^2_{g_i}$
1	0	1.0	$g_1 \in (0, 1.5]$	1.0
2	2.0	1.0	1.0	1.0
3	2.0	1.0	2.0	1.0

We first note that $\Gamma = \{2, 3\}$ and $\mathcal{S}_b = \mathcal{S}_w = \emptyset$. Since the basis for our allocation to systems in Γ regard their "scaled distance" from system 1, and systems 2 and 3 are equal in this respect, we intuitively expect that they will receive equal allocation. To demonstrate the effect of g_1 on the allocation to system 1, we vary g_1 in the interval $(0, 1.5]$. Solving for the optimal allocation as a function of g_1 yields the allocations displayed in figure 1 and the rate z^* displayed in figure 2.

From figure 1, we deduce that as g_1 becomes farther from $\gamma = 0$, system 1 requires a smaller portion of the sample to determine its feasibility. Beyond the point $g_1 = 1.2872$, the feasibility of system 1 is no longer binding in this example. Therefore the optimal allocation as a function of g_1 does not change for $g_1 > 1.2872$. Likewise, in figure 2, the rate of decay of $P\{FS\}$, z^*, grows as a function of g_1 until the point $g_1 = 1.2872$. For $g_1 > 1.2873$, the rate remains constant at $z^* = 0.3431$.

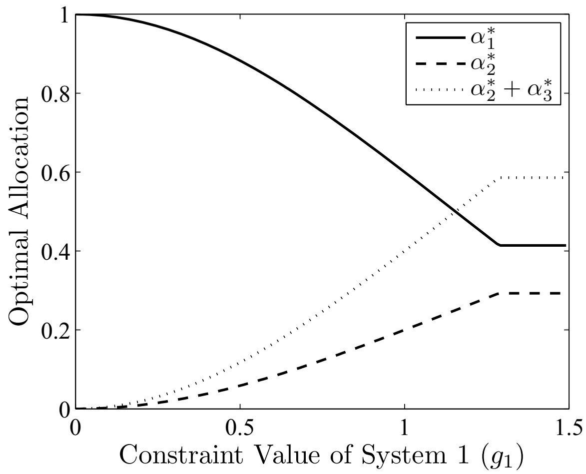

Figure 1: Graph of g_1 versus allocation for the systems in Example 1

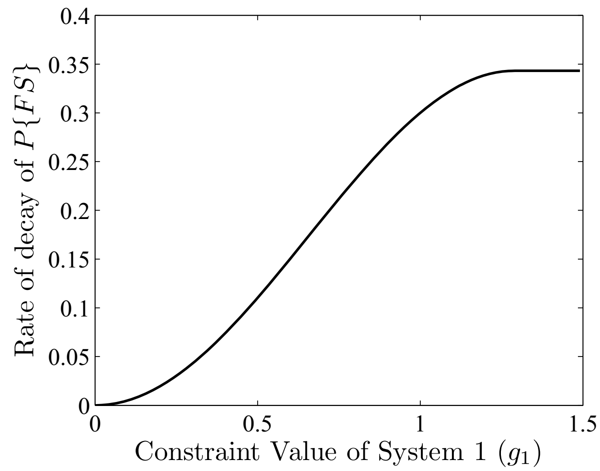

Figure 2: Graph of g_1 versus the rate of decay of $P\{FS\}$ for Example 1

7.2 Comparison with OCBA-CO

Lee et al. [2011] describe an OCBA framework for an asymptotic simulation budget allocation for constrained simulation optimization on finite sets (OCBA-CO). The work by Lee et al. [2011] provides the only other asymptotic sample allocation result for constrained simulation optimization on finite sets in the literature.

For suboptimal systems, Lee et al. [2011] divide the systems into a "feasibility dominance" set and an "optimality dominance" set. Formally, these sets are defined as

$\mathcal{S}_F$: the feasibility dominance set, $\mathcal{S}_F = \{i : P\{\hat{G}_i \geq \gamma\} < P\{\hat{H}_1 > \hat{H}_i\}, i \neq 1\}$,

$\mathcal{S}_O$: the optimality dominance set, $\mathcal{S}_O = \{i : P\{\hat{G}_i \geq \gamma\} \geq P\{\hat{H}_1 > \hat{H}_i\}, i \neq 1\}$.

The assumption $\alpha_1 \gg \alpha_{i \in \mathcal{S}_O}$, along with an approximation to the probability of correct selection, allows Lee et al. [2011] to write their proposed allocation as

$$\frac{\alpha_i}{\alpha_k} = \frac{\left(\dfrac{h_1 - h_k}{\sigma_{h_k}}\right)^2 \mathbb{I}_{k \in \mathcal{S}_O} + \left(\dfrac{\gamma - g_k}{\sigma_{g_k}}\right)^2 \mathbb{I}_{k \in \mathcal{S}_F}}{\left(\dfrac{h_1 - h_i}{\sigma_{h_i}}\right)^2 \mathbb{I}_{i \in \mathcal{S}_O} + \left(\dfrac{\gamma - g_i}{\sigma_{g_i}}\right)^2 \mathbb{I}_{i \in \mathcal{S}_F}} \quad \text{for all} \quad i, k = 2, \ldots, r. \tag{15}$$

As can be seen from equation (15), in OCBA-CO, only one term in each of the numerator and denominator is active at a time. This artifact of the set definitions and the assumptions used in Lee et al. [2011] can sometimes lead to severely suboptimal allocations for infeasible and worse systems. The next example we present is designed to highlight this issue and the consequent inefficiency incurred in the form of a decreased convergence rate of false selection.

Example 2. Suppose we have two systems and one constraint such that each H_i and G_i are iid normally distributed. Let the means and variances be as given in table 3, and let $\gamma = 0$.

Table 3: Means and variances for Example 2

System (i)	h_i	$\sigma^2_{h_i}$	g_i	$\sigma^2_{g_i}$
1	0	2.0	10.0	1.0
2	2.0	1.0	$g_2 \in [-1.9, 0)$	1.0

Note the following features of this example: (i) Since system 2 belongs to $\mathcal{S}_O$ for large enough n and $g_2 \in [-1.9, 0)$, the OCBA-CO allocation to system 2 does not depend on g_2; (ii) For all values of g_2, system 2 is an element of $\mathcal{S}_w$, and hence the proposed allocation *will* change as a function of g_2; (iii) System 1 is decidedly feasible ($g_1 = 10$ and $\sigma_{g_1} = 1$), and so does not require much sample for detecting its feasibility.

Solving for the optimal allocation as a function of g_2 yields the allocations displayed in figure 3 and the overall rate of decay of $P\{FS\}$ displayed in figure 4. From the proposed optimal allocation

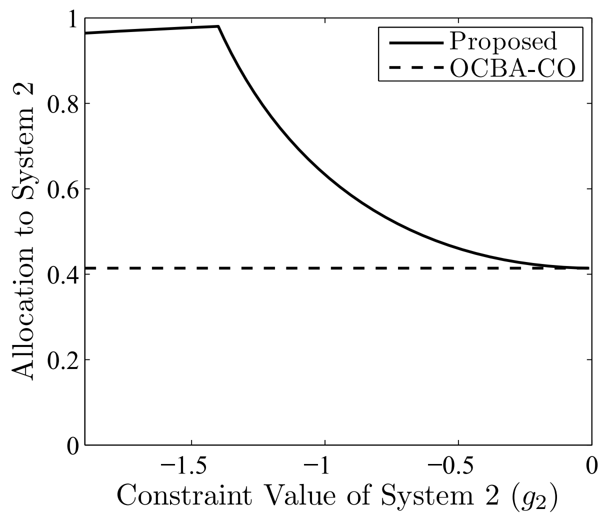

Figure 3: Graph of g_2 versus allocation for the systems in Example 2

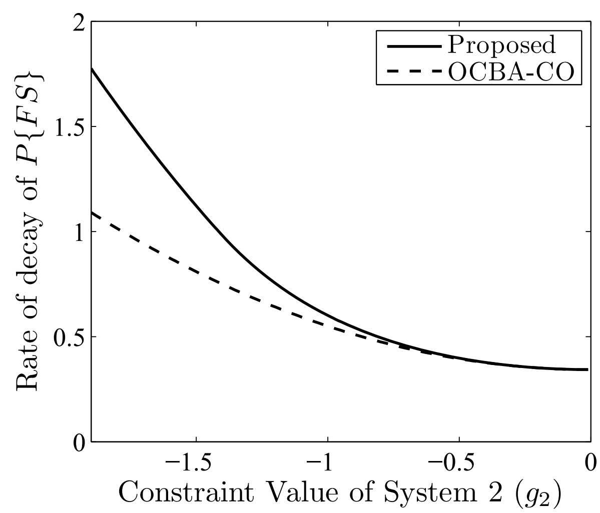

Figure 4: Graph of g_2 versus the rate of decay of $P\{FS\}$ for the systems in Example 2

in figure 3, we see that the allocation to system 2 should not remain constant as a function of g_1, as proposed by Lee et al. [2011]. In fact, for certain values of g_1, we give nearly all of our sample to system 2.

Now suppose we fix the constraint value for system g_2 and explore the allocation to system 1 as a function of σ_{h_1}. As a result of the $\alpha_1 \gg \alpha_i$ assumption, the OCBA-CO allocation to α_1 increases as a function σ_{h_1}, the variance of the objective value of system 1. The next example we present in this section is designed to show how this allocation policy can be severely suboptimal.

Example 3. Let us retain the two systems and their values from Example 2, except we will fix $g_2 = -1.6$, and vary $\sigma^2_{h_1}$ in the interval $[0.2, 4]$. Solving for the optimal allocation as a function of $\sigma^2_{h_1}$ yields the allocations displayed in figure 5 and the achieved rate of decay of $P\{FS\}$ displayed in figure 6.

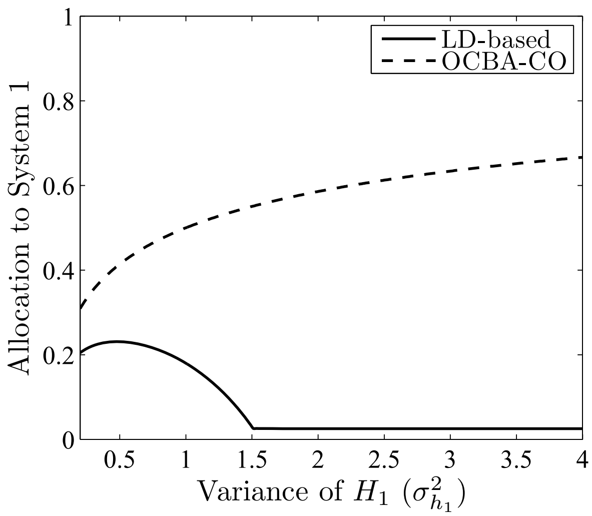

Figure 5: Graph of $\sigma^2_{h_1}$ versus allocation for the systems in Example 3

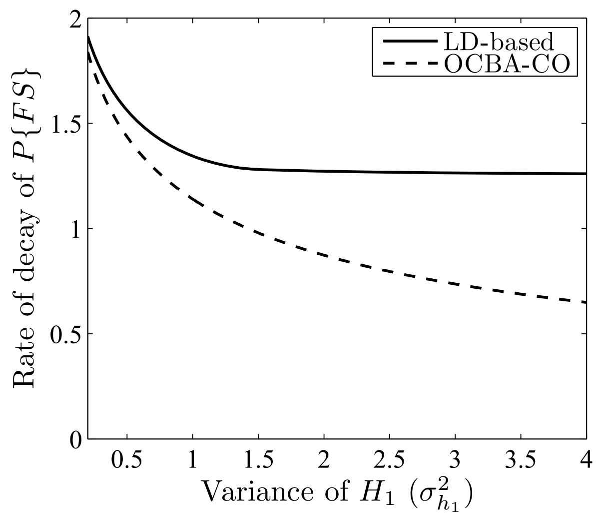

Figure 6: Graph of $\sigma^2_{h_1}$ versus the rate of decay of $P\{FS\}$ for the systems in Example 3

From figure 5, we see that the proposed allocation to system 1 increases slightly at first, and then decreases to a very low, steady allocation from approximately $\sigma^2_{h_1} = 1.5$ onwards. The steady allocation occurs because we require only a minimal sample size allocated to system 1 to determine its feasibility.

Contrasting this allocation is the OCBA-CO allocation, which constantly increases as $\sigma^2_{h_1}$ increases. The OCBA-CO allocation does not exploit the fact that we can correctly select system 1 by allocating more sample to system 2 to disqualify it more quickly. In figure 6, while the proposed allocation achieves a rate of decay that remains constant as $\sigma^2_{h_1}$ increases beyond approximately $\sigma^2_{h_1} = 1.5$, the rate of decay of $P\{FS\}$ for the OCBA-CO allocation continues to decrease as a function of $\sigma^2_{h_1}$.

8. Summary and Concluding Remarks

The constrained SO problem on finite sets is an important SO variation about which little is currently known. Questions surrounding the relationship between sampling and error-probability decay, sampling rates to ensure optimal convergence to the correct solution, and minimum sample size rules that probabilistically guarantee attainment of the correct solution remain largely unexplored. Following recent work by Glynn and Juneja [2004] for the unconstrained SO context and Szechtman and Yücesan [2008] for the context of detecting feasibility, we take the first steps toward answering these questions.

To identify the relationship between sampling and error-probability decay, we strategically divide the competing systems into four sets: best feasible, feasible and worse, infeasible and better, and infeasible and worse. Such strategic division facilitates expressing the rate function of the probability of false selection as the minimum of rate functions over these four sets. Finding the optimal sampling allocation then reduces to solving one of two nonlinear systems of equations.

Two other comments are noteworthy:

(i) We re-emphasize a point relating to implementation. In settings where the underlying distributions of the simulation observations is known or assumed, the rate function estimators used within the sequential algorithm should reflect the rate function of the known or assumed distributions, in contrast to estimating the rate functions generically through the Legendre-Fenchel transform. Our numerical experience suggests that this policy facilitates implementation quite dramatically. Further, in settings where the underlying distribution is not known or assumed, this experience suggests that estimating the underlying rate function using a Taylor's series approximation up to a few terms might prove a viable alternative to estimating rate functions through the Legendre-Fenchel transform.

(ii) An important assumption made in this paper is that of independence between the objective function and constraint estimators for each system. While such assumption is true in certain contexts, it is violated in a number of other "real-world" contexts. In such contexts, the framework presented in this paper should be seen as an approximate guide to simulation allocation obtained through the analysis of an imperfect but tractable model. The question of extending the proposed framework to more general dependence settings is inherently less tractable and is currently being investigated.

Acknowledgement

The authors were supported in part by the Office of Naval Research grants N000140810066, N000140910997, and N000141110065.

Appendix

In this section, we provide two useful results and the proofs that were omitted in the main text.

8.1 Useful Results

In many of the results we present, we repeatedly cite two useful propositions. The first is the principle of the slowest term, which, loosely speaking, states that the rate function of a sum of probabilities is equivalent to the rate function of the slowest converging term in the sum.

Proposition 4 (Principle of the slowest term [see, e.g., Ganesh et al., 2004, Lemma 2.1])**.** *Let $a_n^i, i = 1, 2, \ldots, k$, be a finite number of sequences in $\mathbb{R}^+$, the set of positive reals. If $\lim_{n\to\infty} \frac{1}{n} \log a_n^i$ exists for all i, then*

$$\lim_{n\to\infty} \frac{1}{n} \log \sum_{i=1}^{k} a_n^i = \max_i \left(\lim_{n\to\infty} \frac{1}{n} \log a_n^i \right)$$

A consequence of the principle of the slowest term, Proposition 5 states that the slowest amongst a set of rate functions is equivalent to the rate function of the slowest sequence.

Proposition 5. *Let a_n^i be defined as in Proposition 4. If $\lim_{n\to\infty} \frac{1}{n} \log a_n^i$ exists for all i, then*

$$\max_i \left(\lim_{n\to\infty} \frac{1}{n} \log a_n^i \right) = \lim_{n\to\infty} \frac{1}{n} \log \left(\max_i a_n^i \right)$$

Proof. By the principle of the slowest term, the lower bound is

$$\max_i \left(\lim_{n\to\infty} \frac{1}{n} \log a_n^i \right) = \lim_{n\to\infty} \frac{1}{n} \log \sum_{i=1}^{k} a_n^i \geq \lim_{n\to\infty} \frac{1}{n} \log \max_i a_i^n.$$

Now the upper bound is given by

$$\max_i \left(\lim_{n\to\infty} \frac{1}{n} \log a_n^i \right) = \lim_{n\to\infty} \frac{1}{n} \log \sum_{i=1}^{k} a_n^i \leq \lim_{n\to\infty} \frac{1}{n} \log \left(k \max_i a_i^n \right) = \lim_{n\to\infty} \frac{1}{n} \log \max_i a_i^n. \quad \square$$

8.2 Proof of Theorem 2 and Proposition 3

The rate function for $P\{FS_2\}$ is the rate function for the probability that system 1 is estimated feasible, but another estimated-feasible system has a better estimated objective value. Since the estimated set of feasible systems $\bar{\Gamma}$ may contain worse feasible systems ($i \in \Gamma$), better infeasible systems ($i \in \mathcal{S}_b$), and worse infeasible systems ($i \in \mathcal{S}_w$), in Lemma 3 we strategically consider the rate functions for the probability that system 1 is beaten by a system in $\bar{\Gamma} \cap \Gamma$, $\bar{\Gamma} \cap \mathcal{S}_b$, or $\bar{\Gamma} \cap \mathcal{S}_w$ separately. Lemmas 5 – 7 provide specific statements of these three rate functions over the sets $\Gamma, \mathcal{S}_b$, and $\mathcal{S}_w$, respectively. Lemma 4 provides a useful bookkeeping-type result that is the starting point for Lemmas 5 – 7.

Assuming for now that the required limits exist, Lemma 3 states that the rate function of $P\{FS_2\}$ is determined by the slowest-converging probability that system 1 will be "beaten" by an estimated-feasible system from $\Gamma, \mathcal{S}_b$, or $\mathcal{S}_w$.

Lemma 3. *The rate function for $P\{FS_2\}$ is given by the minimum rate function of the probability that system 1 is beaten by an estimated-feasible system that is (i) feasible and worse, (ii) infeasible and better, or (iii) infeasible and worse. That is,*

$$-\lim_{n\to\infty}\frac{1}{n}\log P\{FS_2\} = \min\Big(-\lim_{n\to\infty}\frac{1}{n}\log P\{\cup_{i\in\bar{\Gamma}\cap\Gamma}\hat{H}_1 \geq \hat{H}_i\},$$
$$-\lim_{n\to\infty}\frac{1}{n}\log P\{\cup_{i\in\bar{\Gamma}\cap\mathcal{S}_b}\hat{H}_1 \geq \hat{H}_i\}, -\lim_{n\to\infty}\frac{1}{n}\log P\{\cup_{i\in\bar{\Gamma}\cap\mathcal{S}_w}\hat{H}_1 \geq \hat{H}_i\}\Big). \tag{16}$$

Proof. From equation (1), the probability that system 1 is beaten by another estimated-feasible system can be written as

$$P\{\cup_{i\in\bar{\Gamma}}\, \hat{H}_1 \geq \hat{H}_i\} = P\{(\cup_{i\in\bar{\Gamma}\cap\Gamma}\hat{H}_1 \geq \hat{H}_i) \cup (\cup_{i\in\bar{\Gamma}\cap\mathcal{S}_b}\hat{H}_1 \geq \hat{H}_i) \cup (\cup_{i\in\bar{\Gamma}\cap\mathcal{S}_w}\hat{H}_1 \geq \hat{H}_i)\}.$$

We have

$$\frac{1}{n}\log\Big(\max\Big(P\{\cup_{i\in\bar{\Gamma}\cap\Gamma}\hat{H}_1 \geq \hat{H}_i\}, P\{\cup_{i\in\bar{\Gamma}\cap\mathcal{S}_b}\hat{H}_1 \geq \hat{H}_i\}, P\{\cup_{i\in\bar{\Gamma}\cap\mathcal{S}_w}\hat{H}_1 \geq \hat{H}_i\}\Big)\Big)$$
$$\leq \frac{1}{n}\log P\{\cup_{i\in\bar{\Gamma}}\hat{H}_1 \geq \hat{H}_i\}$$
$$\leq \frac{1}{n}\log\Big(P\{\cup_{i\in\bar{\Gamma}\cap\Gamma}\hat{H}_1 \geq \hat{H}_i\} + P\{\cup_{i\in\bar{\Gamma}\cap\mathcal{S}_b}\hat{H}_1 \geq \hat{H}_i\} + P\{\cup_{i\in\bar{\Gamma}\cap\mathcal{S}_w}\hat{H}_1 \geq \hat{H}_i\}\Big).$$

Assuming the relevant limits exist, the conclusion is reached by noting that the limit of the left-hand and right-hand sides are equivalent by Proposition 5 and the principle of the slowest term, respectively. □

Next, we will individually consider each of the terms on the right-hand side of equation (16), and establish their respective limits. However before proceeding to these results, we first present the following lemma which is a preliminary step for the proofs that follow. Lemma 4 uses the law of total probability to further separate the events involved in system 1 being "beaten" by another estimated-feasible system.

Lemma 4. *For sets of systems $\mathcal{S} \in \{\Gamma, \mathcal{S}_b, \mathcal{S}_w\}$ and $C \subseteq \mathcal{S}$,*

$$P\{\cup_{i\in\bar{\Gamma}\cap\mathcal{S}}\, \hat{H}_1 \geq \hat{H}_i\}$$
$$= \sum_C P\{(\cup_{i\in C}\hat{H}_1 \geq \hat{H}_i) \cap (\cap_{i\in C}\cap_{j\in\mathcal{C}_F^i}\hat{G}_{ij} \geq \gamma_j) \cap (\cap_{i\in C}\cap_{j\in\mathcal{C}_I^i}\hat{G}_{ij} \geq \gamma_j) \cap (\cap_{i\in\mathcal{S}\setminus C}\cup_j \hat{G}_{ij} < \gamma_j)\} \tag{17}$$

Proof. By the law of total probability, for some set of systems $C \subseteq \mathcal{S}$,

$$P\{\cup_{i\in\bar{\Gamma}\cap\mathcal{S}}\, \hat{H}_1 \geq \hat{H}_i\} = \sum_C P\{(\cup_{i\in\bar{\Gamma}\cap\mathcal{S}}\hat{H}_1 \geq \hat{H}_i) \cap (\bar{\Gamma}\cap\mathcal{S} = C)\}$$

$$= \sum_C P\{(\cup_{i\in C}\hat{H}_1 \geq \hat{H}_i) \cap (\bar{\Gamma} \cap \mathcal{S} = C)\} = \sum_C P\{(\cup_{i\in C}\hat{H}_1 \geq \hat{H}_i) \cap_{i\in C} \ (i \in \bar{\Gamma}) \cap_{i\in \mathcal{S}\setminus C} (i \notin \bar{\Gamma})\}$$
$$= \sum_C P\{(\cup_{i\in C}\hat{H}_1 \geq \hat{H}_i) \cap (\cap_{i\in C} \ (\cap_{j\in \mathcal{C}_F^i}\hat{G}_{ij} \geq \gamma_j \cap_{j\in \mathcal{C}_I^i} \hat{G}_{ij} \geq \gamma_j)) \cap (\cap_{i\in \mathcal{S}\setminus C} \cup_j \hat{G}_{ij} < \gamma_j)\}. \qquad \square$$

Let us now consider the rate function of the probability that system 1 is "beaten" by a worse estimated-feasible system from Γ. Since $\bar{\Gamma}$ is equivalent to Γ in the limit, and we are considering only the probability that system 1 is beaten by another truly feasible system, we expect that the rate function will be the same as in the unconstrained case presented by Glynn and Juneja [2004]. Also, since system 1 can be beaten by any system in $\bar{\Gamma} \cap \Gamma$, we intuitively expect the rate function to be the minimum rate function across all systems in Γ, corresponding to the system that is "best" at crossing the optimality hurdle. Lemma 5 states that this is indeed the case.

Lemma 5. *The rate function for the probability that system 1 is estimated feasible and has a worse estimated objective value than an estimated-feasible system from Γ (feasible and worse) is*

$$-\lim_{n\to\infty} \frac{1}{n} \log P\{\cup_{i\in\bar{\Gamma}\cap\Gamma} \hat{H}_1 \geq \hat{H}_i\} = \min_{i\in\Gamma} \left(\inf_x (\alpha_1 I_1(x) + \alpha_i I_i(x)) \right).$$

Proof. From Lemma 4, let $\mathcal{S} = \Gamma$ and therefore $C \subseteq \Gamma$. Then

$$P\{\cup_{i\in\bar{\Gamma}\cap\Gamma} \hat{H}_1 \geq \hat{H}_i\}$$
$$= \sum_C P\{(\cup_{i\in C}\hat{H}_1 \geq \hat{H}_i) \cap (\cap_{i\in C} \cap_{j\in \mathcal{C}_F^i} \hat{G}_{ij} \geq \gamma_j) \cap (\cap_{i\in C} \cap_{j\in \mathcal{C}_I^i} \hat{G}_{ij} \geq \gamma_j) \cap (\cap_{i\in \Gamma\setminus C} \cup_j \hat{G}_{ij} < \gamma_j)\}$$

We derive a lower bound bound by letting $C = \Gamma$ and noticing that all constraints are feasible for all $i \in \Gamma$. Then

$$P\{\cup_{i\in\bar{\Gamma}\cap\Gamma} \hat{H}_1 \geq \hat{H}_i\} \geq P\{(\cup_{i\in\Gamma}\hat{H}_1 \geq \hat{H}_i) \cap (\cap_{i\in\Gamma} \cap_j \hat{G}_{ij} \geq \gamma_j)\}$$
$$\geq \max_{i\in\Gamma} P\{(\hat{H}_1 \geq \hat{H}_i) \cap (\cap_{i\in\Gamma} \cap_j \hat{G}_{ij} \geq \gamma_j)\}.$$

We derive an upper bound by noting that,

$$P\{\cup_{i\in\bar{\Gamma}\cap\Gamma} \hat{H}_1 \geq \hat{H}_i\} \leq P\{\cup_{i\in\Gamma} \hat{H}_1 \geq \hat{H}_i\} \leq |\Gamma| \max_{i\in\Gamma} P\{\hat{H}_1 \geq \hat{H}_i\}. \tag{18}$$

By Proposition 5 and the independence assumption, the rate function for the lower bound is,

$$\lim_{n\to\infty} \frac{1}{n} \log \max_{i\in\Gamma} P\{(\hat{H}_1 \geq \hat{H}_i) \cap (\cap_{i\in\Gamma} \cap_j \hat{G}_{ij} \geq \gamma_j)\}$$
$$= \max_{i\in\Gamma} \lim_{n\to\infty} \frac{1}{n} \log P\{\underbrace{(\hat{H}_1 \geq \hat{H}_i)}_{\text{pr } \to 0} \cap \underbrace{(\cap_{i\in\Gamma} \ \cap_j \hat{G}_{ij} \geq \gamma_j)}_{\text{pr } \to 1}\} = \max_{i\in\Gamma} \lim_{n\to\infty} \frac{1}{n} \log P\{\hat{H}_1 \geq \hat{H}_i\}.$$

Likewise applying Proposition 5 to equation (18), we find that the rate function for the upper bound is equivalent to the rate function for the lower bound. By Glynn and Juneja [2004],

$$-\lim_{n\to\infty}\frac{1}{n}\log P\{\hat{H}_1 \geq \hat{H}_i\} = \inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x)),$$

and hence the conclusion follows. □

We now consider the rate function of the probability that system 1 has a worse estimated objective value than an estimated-feasible system from $\mathcal{S}_b$ (infeasible but better). We state Lemma 6 without proof as it is similar to the proof of Lemma 7, which immediately follows.

Lemma 6. *The rate function for the probability that system 1 is estimated feasible and has a worse estimated objective value than an estimated-feasible system from $\mathcal{S}_b$ (infeasible and better) is*

$$-\lim_{n\to\infty}\frac{1}{n}\log P\{\cup_{i\in\bar{\Gamma}\cap\mathcal{S}_b}\,\hat{H}_1 \geq \hat{H}_i\} = \min_{i\in\mathcal{S}_b}\alpha_i\sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j).$$

Finally, we consider the rate function for the probability that system 1 has a worse estimated objective value than an estimated-feasible system from $\mathcal{S}_w$ (infeasible and worse). Lemma 7 states this result formally.

Lemma 7. *The rate function for the probability that system 1 is estimated feasible and has a worse estimated objective value than an estimated-feasible system from $\mathcal{S}_w$ (infeasible and worse) is*

$$-\lim_{n\to\infty}\frac{1}{n}\log P\{\cup_{i\in\bar{\Gamma}\cap\mathcal{S}_w}\hat{H}_1 \geq \hat{H}_i\} = \min_{i\in\mathcal{S}_w}\left(\inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x)) + \alpha_i\sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j)\right).$$

Proof. From Lemma 4, let $\mathcal{S} = \mathcal{S}_w$ and therefore $C \subseteq \mathcal{S}_w$. Then we derive an upper bound as,

$$\begin{aligned}
&P\{\cup_{i\in\bar{\Gamma}\cap\mathcal{S}_w}\hat{H}_1 \geq \hat{H}_i\}\\
&= \sum_C P\{(\cup_{i\in C}\hat{H}_1 \geq \hat{H}_i)\cap(\cap_{i\in C}\cap_{j\in\mathcal{C}_F^i}\,\hat{G}_{ij}\geq\gamma_j)\cap(\cap_{i\in C}\cap_{j\in\mathcal{C}_I^i}\,\hat{G}_{ij}\geq\gamma_j)\cap(\cap_{i\in\mathcal{S}_w\setminus C}\cup_j\,\hat{G}_{ij}<\gamma_j)\} \quad (19)\\
&\leq \sum_C P\{(\cup_{i\in C}\hat{H}_1 \geq \hat{H}_i)\cap(\cap_{i\in C}\cap_{j\in\mathcal{C}_I^i}\,\hat{G}_{ij}\geq\gamma_j)\} \leq \sum_C P\{\cup_{i\in C}(\hat{H}_1 \geq \hat{H}_i \cap(\cap_{j\in\mathcal{C}_I^i}\hat{G}_{ij}\geq\gamma_j))\}\\
&\leq \sum_C |C|\max_{i\in C}P\{(\hat{H}_1 \geq \hat{H}_i)\cap(\cap_{j\in\mathcal{C}_I^i}\hat{G}_{ij}\geq\gamma_j)\} \leq 2^{|\mathcal{S}_w|}|\mathcal{S}_w|\max_{i\in\mathcal{S}_w}P\{(\hat{H}_1 \geq \hat{H}_i)\cap(\cap_{j\in\mathcal{C}_I^i}\hat{G}_{ij}\geq\gamma_j)\}.
\end{aligned}$$

Therefore the rate function for the upper bound is

$$\lim_{n\to\infty}\frac{1}{n}\log\max_{i\in\mathcal{S}_w}P\{(\hat{H}_1 \geq \hat{H}_i)\cap(\cap_{j\in\mathcal{C}_I^i}\hat{G}_{ij}\geq\gamma_j)\}. \quad (20)$$

Let $k^* = \arg\max_{i\in\mathcal{S}_w} P\{(\hat{H}_1 \geq \hat{H}_i) \cap (\cap_{j\in\mathcal{C}_I^i}\hat{G}_{ij} \geq \gamma_j)\}$. We derive a lower bound by letting k^* be the only element in C. Continuing from equation (19),

$$\sum_C P\{(\cup_{i\in C}\hat{H}_1 \geq \hat{H}_i) \cap (\cap_{i\in C}\cap_{j\in\mathcal{C}_F^i}\hat{G}_{ij} \geq \gamma_j) \cap (\cap_{i\in C}\cap_{j\in\mathcal{C}_I^i}\hat{G}_{ij} \geq \gamma_j) \cap (\cap_{i\in\mathcal{S}_w\setminus C}\cup_j \hat{G}_{ij} < \gamma_j)\}$$
$$\geq P\{(\hat{H}_1 \geq \hat{H}_{k^*}) \cap (\cap_{j\in\mathcal{C}_F^{k^*}}\hat{G}_{k^*j} \geq \gamma_j) \cap (\cap_{j\in\mathcal{C}_I^{k^*}}\hat{G}_{k^*j} \geq \gamma_j) \cap (\cap_{i\in\mathcal{S}_w\setminus\{k^*\}}\cup_j \hat{G}_{ij} < \gamma_j)\}.$$

By Proposition 5 and the independence assumption,

$$\lim_{n\to\infty}\frac{1}{n}\log P\{\underbrace{(\hat{H}_1 \geq \hat{H}_{k^*})}_{\text{pr }\to 0} \cap \underbrace{(\cap_{j\in\mathcal{C}_F^{k^*}}\hat{G}_{k^*j} \geq \gamma_j)}_{\text{pr }\to 1} \cap \underbrace{(\cap_{j\in\mathcal{C}_I^{k^*}}\hat{G}_{k^*j} \geq \gamma_j)}_{\text{pr }\to 0} \cap \underbrace{(\cap_{i\in\mathcal{S}_w\setminus\{k^*\}}\cup_j \hat{G}_{ij} < \gamma_j)}_{\text{pr }\to 1}\}$$
$$= \lim_{n\to\infty}\frac{1}{n}\log\max_{i\in\mathcal{S}_w} P\{(\hat{H}_1 \geq \hat{H}_i) \cap (\cap_{j\in\mathcal{C}_I^i}\hat{G}_{ij} \geq \gamma_j)\},$$

which is the rate function for the probability that system k^* is falsely estimated as optimal and feasible on all constraints for which it is truly infeasible. We note that this rate function is equivalent to the rate function for the upper bound in equation (20). By Proposition 5 and the independence assumption, the rate function for the upper and lower bounds is,

$$\lim_{n\to\infty}\frac{1}{n}\log\max_{i\in\mathcal{S}_w} P\{(\hat{H}_1 \geq \hat{H}_i) \cap (\cap_{j\in\mathcal{C}_I^i}\hat{G}_{ij} \geq \gamma_j)\}$$
$$= \max_{i\in\mathcal{S}_w}\big(\lim_{n\to\infty}\frac{1}{n}\log P\{\hat{H}_1 \geq \hat{H}_i\} + \sum_{j\in\mathcal{C}_I^i}\lim_{n\to\infty}\frac{1}{n}\log P\{\hat{G}_{ij} \geq \gamma_j\big)\}\big)$$

Applying previous results, the conclusion follows. □

Proof of Theorem 2. We arrive at Theorem 2 by substituting the results from Lemmas 5–7 into the result presented in Lemma 3. □

Proof of Proposition 3. We will only prove that $\widehat{\mathcal{C}}_F^i \to \mathcal{C}_F^i$ wp1 as $m \to \infty$. The proofs for the other parts of the theorem follow in a very similar fashion.

By Assumption 3, $\hat{G}_{ij}(m) \to g_{ij}$ wp1 for all $i \leq r$ and $j \leq s$. We know that $g_{ij} > \gamma_j$ for each $j \in \mathcal{C}_F^i$. Since $|\mathcal{C}_F^i| < \infty$, we conclude that for large enough m, $\hat{G}_{ij}(m) > \gamma_j$ uniformly in $j \in \mathcal{C}_F^i$ wp1, and hence the assertion holds. □

References

S. Andradóttir. An overview of simulation optimization via random search. In S. G. Henderson and B. L. Nelson, editors, *Simulation*, Handbooks in Operations Research and Management Science, pages 617–631. Elsevier, 2006.

S. Andradóttir and S.-H. Kim. Fully sequential procedures for comparing constrained systems via simulation. *Naval Research Logistics*, 57(5):403–421, 2010.

R. R. Barton and M. Meckesheimer. Metamodel-based simulation optimization. In S. G. Henderson and B. L. Nelson, editors, *Simulation*, Handbooks in Operations Research and Management Science, pages 535–574. Elsevier, 2006.

D. Batur and S.-H. Kim. Finding feasible systems in the presence of constraints on multiple performance measures. *ACM Transactions on Modeling and Computer Simulation*, 20(3):13:1–26, 2010.

J. A. Bucklew. *Introduction to Rare Event Simulation*. Springer, New York, 2003.

C.-H. Chen, J. Lin, E. Yücesan, and S. E. Chick. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270, 2000.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, New York, 2nd edition, 1998.

M. C. Fu. Optimization for simulation: Theory vs. practice. *INFORMS Journal on Computing*, 14:192–215, 2002.

A. Ganesh, N. O'Connell, and D. Wischik. *Big Queues*. Lecture Notes in Mathematics, Volume 1838. Springer, New York, 2004.

P. Glynn and S. Juneja. A large deviations perspective on ordinal optimization. In R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, editors, *Proceedings of the 2004 Winter Simulation Conference*, pages 577–585, Piscataway, New Jersey, 2004. Institute of Electrical and Electronics Engineers, Inc.

P. Glynn and S. Juneja. Ordinal optimization: A large deviations perspective. 2006. Working Paper Series, Indian School of Business.

S. G. Henderson and R. Pasupathy. Simulation optimization library, 2011. URL `http://www.simopt.org`. Accessed Febuary 24, 2011.

S.-H. Kim and B. L. Nelson. Selecting the best system. In S. G. Henderson and B. L. Nelson, editors, *Simulation*, Handbooks in Operations Research and Management Science, Volume 13, pages 501–534. Elsevier, 2006.

L. H. Lee, N. A. Pujowidianto, L.-W. Li, C.-H. Chen, and C. M. Yap. Asymptotic simulation budget allocation for selecting the best design in the presence of stochastic constraints. 2011. Submitted to *IEEE Transactions on Automatic Control*.

S. Ólafsson and J. Kim. Simulation optimization. In E. Yucesan, C. H. Chen, J. L. Snowdon, and J. M. Charnes, editors, *Proceedings of the 2002 Winter Simulation Conference*, pages 79–84. Institute of Electrical and Electronics Engineers: Piscataway, New Jersey, 2002.

R. Pasupathy and S. Kim. The stochastic root-finding problem: overview, solutions, and open questions. *ACM TOMACS*, 2010. Under revision.

W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1976.

J. C. Spall. *Introduction to Stochastic Search and Optimization*. John Wiley & Sons, Inc., Hoboken, NJ., 2003.

R. Szechtman and E. Yücesan. A new perspective on feasibility determination. In S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, editors, *Proceedings of the 2008 Winter Simulation Conference*, pages 273–280, Piscataway, New Jersey, 2008. Institute of Electrical and Electronics Engineers, Inc.