INFORMS Journal on Computing
Vol. 25, No. 3, Summer 2013, pp. 527–542
ISSN 1091-9856 (print) | ISSN 1526-5528 (online)

http://dx.doi.org/10.1287/ijoc.1120.0519

Optimal Sampling Laws for Stochastically Constrained Simulation Optimization on Finite Sets

Susan R. Hunter
School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853,
hunter@cornell.edu

Raghu Pasupathy
The Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia 24061,
pasupath@vt.edu

Consider the context of selecting an optimal system from among 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. We provide a consistent estimator for the optimal allocation and a corresponding sequential algorithm fit for implementation. Various numerical examples demonstrate how the proposed allocation differs from competing algorithms.

Key words: constrained simulation optimization; optimal allocation; ranking and selection
History: Accepted by Marvin Nakayama, Area Editor for Simulation; received March 2011; revised December 2011, (twice in) May 2012; accepted May 2012. Published online in *Articles in Advance* September 14, 2012.

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 through a stochastic simulation model—a formulation that 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), and Ólafsson and Kim (2002) for overviews and entry points into this literature, and Pasupathy and Henderson (2006, 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 (R&S), the currently available solution methods are reliable and have stable digital implementations (Kim and Nelson 2006). 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 the system having the best objective function value from among those systems whose constraint values cross a prespecified threshold, using only the simulation output. The efficiency of a solution to this problem, which we 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 characterize the nature of optimal sampling plans when solving the constrained SO problem on finite sets. 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.

Table 1 Categorization of Research in the Area of Simulation Optimization on Finite Sets by the Nature of the Result, the Required Distributional Assumption, and the Presence of an Objective or Constraints

Result Time	Req'd Dist'n	Optimization: objective(s) only	Feasibility: constraint(s) only	Constrained optimization: objective(s) and constraint(s)
Finite	Normal	R&S (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[a]	OCBA-CO (Lee et al. 2011)
Infinite	General	Glynn and Juneja (2004)	Szechtman and Yücesan (2008)	This work

Note. This table provides example references for each category.
[a]Problems in the infinite-normal row are also solved as applications of solutions in the infinite-general row.

Question 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 among 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?

Question 2. Given a satisfactory answer to Question 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 by 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 two categories:

—*finite-time stopping procedures*, which typically require a normality assumption and provide finite-time guarantees on, for example, the expected opportunity cost (e.g., Branke et al. 2007) or, more traditionally, the probability of false selection (see, e.g., Kim and Nelson 2006, for an overview of R&S procedures), and

—*asymptotically efficient procedures*, such as optimal computing budget allocation (OCBA) (e.g., Chen et al. 2000) which provides an approximately optimal sample allocation under the assumption of normality, and procedures that use a large-deviations (LD) approach (e.g., Glynn and Juneja 2004), to provide an asymptotically optimal sample allocation in the context of general light-tailed distributions.

Corresponding research in the constrained context is taking an analogous route. As illustrated in Table 1, Andradóttir and Kim (2010) provide finite-time guarantees on the probability of false selection for stochastically constrained SO problems and parallels traditional R&S work. Similarly, recent work by Lee et al. (2011) provide an asymptotically efficient procedure to solve stochastically constrained SO problems that parallels the previous OCBA work in the unconstrained context. Our work, which appears in the bottom right-hand cell of Table 1, provides an asymptotically efficient procedure that completely generalizes previous LD 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 $i = 1, 2, \ldots, r$ of systems, each with an unknown objective value $h_i \in \mathbb{R}$ and unknown constraint values $g_{ij} \in \mathbb{R}$, $j = 1, 2, \ldots, s$, and $i = 1, 2, \ldots, r$. Given constants $\gamma_j \in \mathbb{R}$, $j = 1, 2, \ldots, s$, we wish to select the system with the lowest objective value h_i, subject to the constraints $g_{ij} \le \gamma_j$. That is, we consider

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

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.

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 $\alpha_i \ge 0$ for all $i = 1, 2, \ldots, r$. Furthermore, let the system having the smallest estimated objective value among 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?

2. Contributions

This paper addresses the question of identifying the "best" among 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 is a crucial generalization of the work on unconstrained simulation optimization on finite sets by Glynn and Juneja (2004). We contribute the following.

Contribution 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 an LD framework, we derive the probability law for erroneously obtaining a suboptimal solution as a function of the sampling plan. We show that the optimal sampling plan can be identified as the solution to a concave maximization problem.

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

Contribution 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

Let $i \le r$ and $j \le s$ be notational shorthand for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, s$, respectively. Let the feasible system with the lowest objective value be system 1. We partition the set of r systems into the following four mutually exclusive and collectively exhaustive subsets.

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

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

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

$\mathcal{S}_w := \{i\colon h_1 < h_i \text{ and } g_{ij} > \gamma_j \text{ for at least one } j \le s\}$ 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 experience a large deviation in the estimated objective value, systems in $\mathcal{S}_b$ must experience a large deviation in estimated constraints, and systems in $\mathcal{S}_w$ must experience a large deviation in estimated objective and constraint values. 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.

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

$\mathcal{C}_I^i := \{j\colon g_{ij} > \gamma_j\}$ 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 $(\mathcal{A}_m \cap \mathcal{A}^c) \cup (\mathcal{A} \cap \mathcal{A}_m^c)$ is the empty set.

To aid readability, we adopt the following notational convention throughout: 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^*; vectors appear in bold type, e.g., $\boldsymbol{\alpha}$.

3.2. Assumptions

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

Assumption 1. (1) *The random variables* $(H_i, G_{i1}, \ldots, G_{is})$ *are mutually independent for all* $i \le r$, *and* (2) *for any particular system* i, *the 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. A discussion of the independence Assumption 1(2) and its relaxation is provided in §8.

Let $\bar{H}_i(n) = \frac{1}{n}\sum_{k=1}^{n} H_{ik}$ and $\bar{G}_{ij}(n) = \frac{1}{n}\sum_{k=1}^{n} G_{ijk}$. We define $\hat{H}_i \equiv \bar{H}_i(\alpha_i n)$ and $\hat{G}_{ij} \equiv \bar{G}_{ij}(\alpha_i n)$ as shorthand for the estimators of h_i and g_{ij} after scaling the sample size by $\alpha_i > 0$, the proportion of the total sample n that is allocated to system i. We ignore issues due to the stipulation that $\alpha_i n$ is an integer. Let $\Lambda^{(n)}_{H_i}(\theta) = \log E[e^{\theta \bar{H}_i(n)}]$ and $\Lambda^{(n)}_{G_{ij}}(\theta) = \log E[e^{\theta \bar{G}_{ij}(n)}]$ be the cumulant generating functions of $\bar{H}_i(n)$ and $\bar{G}_{ij}(n)$, respectively. Let the effective domain of a function $f(\cdot)$ be denoted $\mathcal{D}_f = \{x\colon f(x) < \infty\}$, and its interior $\mathcal{D}^\circ_f$. Let $f'(x)$ denote the derivative of f with respect to the argument x. We make the following assumption, which is standard in LD contexts (see, e.g., Dembo and Zeitouni 1998).

Assumption 2. *For each system $i \le r$ and constraint $j \le s$,*

(1) *the limits $\Lambda_{H_i}(\theta) = \lim_{n\to\infty} \frac{1}{n}\Lambda^{(n)}_{H_i}(n\theta)$ and $\Lambda_{G_{ij}}(\theta) = \lim_{n\to\infty} \frac{1}{n}\Lambda^{(n)}_{G_{ij}}(n\theta)$ exist as extended real numbers for all θ;*

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

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

(4) *$\Lambda_{H_i}(\theta)$ and $\Lambda_{G_{ij}}(\theta)$ are steep (e.g., for any sequence $\{\theta_n\} \in \mathcal{D}_{\Lambda_{H_i}}$ that converges to a boundary point of $\mathcal{D}_{\Lambda_{H_i}}$, $\lim_{n\to\infty} |\Lambda'_{H_i}(\theta_n)| = \infty$).*

The postulates of Assumption 2 imply that $\bar{H}_i(n) \to h_i$ wp1 and $\bar{G}_{ij}(n) \to g_{ij}$ wp1 (see Bucklew 2003, Remark 3.2.1). Furthermore, Assumption 2 ensures that, by the Gärtner-Ellis theorem (Dembo and Zeitouni 1998, p. 44), the probability measures governing $\bar{H}_i(n)$ and $\bar{G}_{ij}(n)$ satisfy the large deviations principle (LDP) with good rate functions $I_i(x) = \sup_{\theta\in\mathbb{R}}\{\theta x - \Lambda_{H_i}(\theta)\}$ and $J_{ij}(y) = \sup_{\theta\in\mathbb{R}}\{\theta y - \Lambda_{G_{ij}}(\theta)\}$, respectively. Assumption 2(3) is stronger than what is needed for the Gärtner-Ellis theorem to hold. However, we require $\Lambda_{H_i}(\theta)$ and $\Lambda_{G_{ij}}(\theta)$ to be strictly convex and C^∞ on the interiors of their respective domains so that $I_i(x)$ and $J_{ij}(y)$ are strictly convex and C^∞ for $x \in \mathcal{F}^\circ_{H_i} = \text{int}\{\Lambda'_{H_i}(\theta)\colon \theta \in \mathcal{D}^\circ_{\Lambda_{H_i}}\}$ and $y \in \mathcal{F}^\circ_{G_{ij}} = \text{int}\{\Lambda'_{G_{ij}}(\theta)\colon \theta \in \mathcal{D}^\circ_{\Lambda_{G_{ij}}}\}$, respectively. The postulates of Assumption 2 hold if each replicate of H_i and G_{ij} is an independent and identically distributed (iid) copy from a distribution with moment-generating function defined everywhere.

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

Assumption 3. (1) *the interval $[h_\ell, h_u] \subset \bigcap_{i=1}^{r} \mathcal{F}^\circ_{H_i}$, and* (2) *$\gamma_j \in \bigcap_{i=1}^{r} \mathcal{F}^\circ_{G_{ij}}$ for all $j \le s$.*

Assumption 3 ensures that there is nonzero probability of a false selection event. As in Glynn and Juneja (2004), Assumption 3(1) ensures that $\hat{H}_i$ may take any value in the interval $[h_\ell, h_u]$ and that $P(\hat{H}_i \le \hat{H}_1) > 0$ for $2 \le i \le r$. Assumption 3(2) likewise ensures there is a nonzero probability that each system will be deemed feasible or infeasible on any of its constraints; particularly, it ensures that $P(\bigcap_{j\in\mathcal{C}^i_l} \hat{G}_{ij} \le \gamma_j) > 0$ for $i \in \mathcal{S}_b \cup \mathcal{S}_w$ and $P(\hat{G}_{1j} > \gamma_j) > 0$ for all $j \le s$. Assumption 3 is easy to satisfy in practice. For example, any of the commonly encountered light-tailed distributions with overlapping support satisfy Assumption 3(1).

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 wp1, we make the following assumption.

Assumption 4. *No system has the same objective value as system 1, and no system lies exactly on a constraint; that is, $h_1 \ne h_i$ for all $i = 2, \ldots, r$ and $g_{ij} \ne \gamma_j$ for all $i \le r$, $j \le s$.*

Since distinguishing two values that are the same through simulation requires infinite sample, this assumption is relatively standard for optimal allocation literature—Glynn and Juneja (2004), Szechtman and Yücesan (2008), and Lee et al. (2011) also require assumptions of this type.

4. Rate Function of Probability of False Selection

The false selection (FS) event is when 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. Since system 1 is truly feasible, the event that no systems are estimated feasible is considered a false selection event. Therefore,

$$P\{FS\} = P\Bigg\{\overbrace{\bigcup_{j=1}^{s} \hat{G}_{1j} > \gamma_j}^{\substack{\text{system 1}\\ \text{estimated}\\ \text{infeasible}}}\Bigg\}$$

$$+ P\Bigg\{\overbrace{\bigcup_{i=2}^{r}\Bigg[\underbrace{(\hat{H}_i \le \hat{H}_1)}_{\substack{\text{system i}\\ \text{"beats"}\\ \text{system 1}}} \cap \underbrace{\Bigg(\bigcap_{j=1}^{s} \hat{G}_{ij} \le \gamma_j\Bigg)}_{\substack{\text{system i}\\ \text{estimated}\\ \text{feasible}}} \cap \underbrace{\Bigg(\bigcap_{j=1}^{s} \hat{G}_{1j} \le \gamma_j\Bigg)}_{\substack{\text{system 1}\\ \text{estimated}\\ \text{feasible}}}\Bigg]}^{\substack{\text{system 1 estimated feasible and other}\\ \text{estimated-feasible system(s) "beat" system 1}}}\Bigg\}$$

$$= P\{FS_1\} + P\Bigg\{\bigcup_{i=2}^{r} FS_i\Bigg\}. \tag{1}$$

Bounds on the $P\{FS\}$ are

$$P\{FS_1\} + \max_{2\le i\le r} P\{FS_i\} \le P\{FS\} \le r(P\{FS_1\} + \max_{2\le i\le r} P\{FS_i\}).$$

By Propositions A.1 and A.2 (available as supplemental material at http://dx.doi.org/10.1287/ijoc.1120.0519), assuming the relevant limits exist,

$$-\lim_{n\to\infty}\frac{1}{n}\log P\{FS\} = \min\left(-\lim_{n\to\infty}\frac{1}{n}\log P\{FS_1\}, \min_{2\le i\le r}\left(-\lim_{n\to\infty}\frac{1}{n}\log P\{FS_i\}\right)\right). \quad (2)$$

Therefore the rate function of $P\{FS\}$ is governed by the slowest of (i) the rate of decay of $P\{FS_1\}$, where $P\{FS_1\}$ is the probability that system 1 is estimated infeasible, and (ii) the slowest across infeasible and suboptimal systems of the rate of decay of $P\{FS_i\}$, where $P\{FS_i\}$ is the probability that system 1 is estimated feasible but "beaten" in objective value by another estimated-feasible system i. In the following Theorems 1 and 2, we individually derive the rate functions for $P\{FS_1\}$ and $P\{FS_i\}$ 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 rate function of $P\{FS_1\}$ as the rate function corresponding to the constraint that is most likely to qualify system 1 as infeasible.

THEOREM 1. *The rate function for $P\{FS_1\}$ is*

$$-\lim_{n\to\infty}\frac{1}{n}\log P\{FS_1\} = \min_{j\in\mathcal{C}_F^1}\alpha_1 J_{1j}(\gamma_j).$$

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

$$\max_{j\in\mathcal{C}_F^1} P\{\hat{G}_{1j} > \gamma_j\} \le P\left\{\bigcup_{j=1}^{s}\hat{G}_{1j} > \gamma_j\right\} \le s\max_{j\in\mathcal{C}_F^1} P\{\hat{G}_{1j} > \gamma_j\}.$$

It follows from Proposition A.2 (see online supplement) that

$$\lim_{n\to\infty}\frac{1}{n}\log\max_{j\in\mathcal{C}_F^1} P\{\hat{G}_{1j} > \gamma_j\} = \max_{j\in\mathcal{C}_F^1}\lim_{n\to\infty}\frac{1}{n}\log P\{\hat{G}_{1j} > \gamma_j\}.$$

By Assumption 2 and the Gärtner-Ellis theorem, an analysis similar to that of Szechtman and Yücesan (2008) yields

$$\lim_{n\to\infty}\frac{1}{n}\log P\{FS_1\} = \max_{j\in\mathcal{C}_F^1}\lim_{n\to\infty}\frac{1}{n}\log P\{\hat{G}_{1j} > \gamma_j\} = -\min_{j\in\mathcal{C}_F^1}\alpha_j J_{1j}(\gamma_j). \quad \square$$

Now consider $P\{FS_i\}$. Since system 1 can be beaten in objective value by worse feasible systems ($i\in\Gamma$), better infeasible systems ($i\in\mathcal{S}_b$), or 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 Γ, $\mathcal{S}_b$, or $\mathcal{S}_w$ separately. Theorem 2 states the rate function of $P\{FS_i\}$.

THEOREM 2. *The rate function for $P\{FS_i\}$ is given by*

$$-\lim_{n\to\infty}\frac{1}{n}\log P\{FS_i\} = \begin{cases}\inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x)), & i\in\Gamma,\\ \alpha_i\sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j), & i\in\mathcal{S}_b,\\ \inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x)) + \alpha_i\sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j), & i\in\mathcal{S}_w.\end{cases}$$

PROOF. From Equation (1), assuming the relevant limits exist,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log P\{FS_i\} &= \lim_{n\to\infty}\frac{1}{n}\log P\left\{(\hat{H}_i\le\hat{H}_1)\cap\left(\bigcap_{j=1}^{s}\hat{G}_{ij}\le\gamma_j\right)\cap\left(\bigcap_{j=1}^{s}\hat{G}_{1j}\le\gamma_j\right)\right\}\\ &= \lim_{n\to\infty}\frac{1}{n}\log P\{\hat{H}_i\le\hat{H}_1\} + \sum_{j=1}^{s}\lim_{n\to\infty}\frac{1}{n}\log P\{\hat{G}_{ij}\le\gamma_j\}\\ &\quad + \lim_{n\to\infty}\frac{1}{n}\log P\left\{\bigcap_{j=1}^{s}\hat{G}_{1j}\le\gamma_j\right\} \quad (3)\\ &= \lim_{n\to\infty}\frac{1}{n}\log P\{\hat{H}_i\le\hat{H}_1\} + \sum_{j\in\mathcal{C}_I^i}\lim_{n\to\infty}\frac{1}{n}\log P\{\hat{G}_{ij}\le\gamma_j\}, \quad (4)\end{aligned}$$

where Equation (3) holds by Assumption 1, and Equation (4) holds since the probability that system 1 is estimated feasible and the probability that system i is estimated feasible on constraints $j\in\mathcal{C}_F^i$ tend to one. The result follows by considering the systems by their classification as elements of Γ, $\mathcal{S}_b$, or $\mathcal{S}_w$ and noting the probabilities that tend to one. The result for systems $i\in\Gamma$ follows directly from Glynn and Juneja (2004). For systems $i\in\mathcal{S}_b\cup\mathcal{S}_w$, as in Szechtman and Yücesan (2008), under Assumption 2 one can show that for any system i not satisfying constraint $j\in\mathcal{C}_I^i$, the rate function for the probability that system i is incorrectly estimated feasible on constraint j is $-\lim_{n\to\infty}\frac{1}{n}\log P\{\hat{G}_{ij}\le\gamma_j\} = \alpha_i J_{ij}(\gamma_j)$ for all $j\in\mathcal{C}_I^i$, and the result follows. $\square$

In Theorem 2, the rate function of $P\{FS_i\}$ is determined by whether the competing system i is truly feasible and worse ($i\in\Gamma$), infeasible and better ($i\in\mathcal{S}_b$), or infeasible and worse ($i\in\mathcal{S}_w$). However systems in Γ, $\mathcal{S}_b$, 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 now combine the results for $P\{FS_1\}$ and $P\{FS_i\}$ to derive the rate function for $P\{FS\}$. Recalling from Equation (2) that the rate function of the $P\{FS\}$ is given by the minimum of the rate functions for $P\{FS_1\}$ and $P\{FS_i\}$ for all $2 \le i \le r$ yields Theorem 3.

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*

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

Theorem 3 asserts that the overall rate function of the probability of false selection is determined by *the* most likely false selection event.

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:

$$\begin{aligned}\max\ \min\Bigg(&\min_{j\in\mathcal{C}_F^1}\alpha_1 J_{1j}(\gamma_j),\ \min_{i\in\Gamma}\left(\inf_x(\alpha_1 I_1(x)+\alpha_i I_i(x))\right),\\ &\min_{i\in\mathcal{S}_b}\alpha_i\sum_{j\in\mathcal{C}_I^i}J_{ij}(\gamma_j),\\ &\min_{i\in\mathcal{S}_w}\left(\inf_x(\alpha_1 I_1(x)+\alpha_i I_i(x))+\alpha_i\sum_{j\in\mathcal{C}_I^i}J_{ij}(\gamma_j)\right)\Bigg)\\ \text{s.t.}\ &\sum_{i=1}^r\alpha_i=1,\quad \alpha_i\ge 0 \text{ for all } i\le r.\end{aligned}\tag{5}$$

By Glynn and Juneja (2006), $\inf_x(\alpha_1 I_1(x)+\alpha_i I_i(x))$ is a concave, 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, C^∞ functions of α_1 and α_i. Since the minimum of concave functions is also concave, the problem in (5) is a concave maximization problem. Equivalently, we may rewrite the problem in (5) as the following Problem Q, where we 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.

Problem Q:

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

Slater's Condition (see, e.g., Boyd and Vandenberghe 2004, p. 226) holds for Problem Q; that is, there exists a point $\boldsymbol{\alpha}$ in the relative interior of the feasible set such that the inequality constraints hold strictly. For example, such a point is $z=0$, $\alpha_i = 1/r$ for all $i\le r$. Since Problem Q is concave with differentiable objective function and constraints and Slater's condition holds, the Karush-Kuhn Tucker (KKT) conditions are necessary and sufficient for global optimality (see, e.g., Boyd and Vandenberghe 2004).

From the KKT conditions on Problem Q, we define Problem Q^* by replacing the inequality constraints corresponding to systems in Γ, $\mathcal{S}_b$, and $\mathcal{S}_w$ with equality constraints, and forcing each α_i to be strictly greater than zero.

Problem Q^*:

max z

s.t.

$$\alpha_1 J_{1j}(\gamma_j) \geq z, \quad j \in \mathcal{C}_F^1,$$
$$\alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i)) = z, \quad i \in \Gamma,$$
$$\alpha_i \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j) = z, \quad i \in \mathcal{S}_b,$$
$$\alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i)) + \alpha_i \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j) = z, \quad i \in \mathcal{S}_w,$$
$$\sum_{i=1}^{r} \alpha_i = 1, \quad \alpha_i > 0 \text{ for all } i \leq r.$$

Proposition 1 below states the equivalence of Problems Q and Q^*.

Proposition 1. *Problems Q and Q^* are equivalent; that is, a solution $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_r^*)$ is optimal for Problem Q if and only if $\boldsymbol{\alpha}^*$ is optimal for Problem Q^*.*

Proof. We argue that the forward and backward assertions of the theorem hold if a point satisfying the KKT conditions of Problem Q is feasible for Problem Q^*. We then complete the proof by showing that any point satisfying the KKT conditions of Problem Q is indeed feasible for Problem Q^*.

($\Rightarrow$) Since the feasible region of Problem Q^* is a subset of that of Problem Q, then if an optimal solution to Problem Q is feasible to Problem Q^*, it must be optimal for Problem Q^*. Since the KKT conditions are necessary and sufficient for optimality in Problem Q, if a point satisfying the KKT conditions of Problem Q must be feasible for Problem Q^*, the result holds.

($\Leftarrow$) Suppose $\boldsymbol{\alpha}^*$ is optimal for Problem Q^* and $\check{\boldsymbol{\alpha}}^* = (\check{\alpha}_1^*, \check{\alpha}_2^*, \ldots, \check{\alpha}_r^*)$ is optimal for Problem Q, where $\boldsymbol{\alpha}^* \neq \check{\boldsymbol{\alpha}}^*$. Since the KKT conditions are necessary and sufficient for optimality in Problem Q, $\check{\boldsymbol{\alpha}}^*$ satisfies the KKT conditions of Problem Q. Since the objective functions for Problems Q and Q^* are identical and the feasible region for Problem Q^* is a subset of that of Problem Q, then $\check{\boldsymbol{\alpha}}^*$ must not be feasible for Problem Q^*. Therefore if a point satisfying the KKT conditions of Problem Q must be feasible for Problem Q^*, we have a contradiction and the result holds.

We now show that a point satisfying the KKT conditions of Problem Q must be feasible for Problem Q^*. First, note that for $\alpha_i = 1/r$, $i \leq r$, we have $z > 0$ in Problem Q. Therefore $\alpha_i = 0$ for some $i \in \{1\} \cup \mathcal{S}_b$ is suboptimal since $z = 0$. Now consider $\alpha_i = 0$ for some $i \in \Gamma \cup \mathcal{S}_w$. 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 $i \leq r$.

Let ν and $\boldsymbol{\lambda} = (\lambda_j^1 \geq 0, \lambda_i \geq 0 \colon j \in \mathcal{C}_F^1,\ i = 2, \ldots, r)$ be dual variables for Problem Q. Since $x(\alpha_1, \alpha_i)$ solves $\alpha_1 I_1'(x) + \alpha_i I_i'(x) = 0$, then as in Glynn and Juneja (2004),

$$\begin{aligned} &\frac{\partial}{\partial \alpha_1}(\alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i))) \\ &\qquad = I_1(x(\alpha_1, \alpha_i)) > 0, \\ &\frac{\partial}{\partial \alpha_i}(\alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i))) \\ &\qquad = I_i(x(\alpha_1, \alpha_i)) > 0. \end{aligned} \tag{6}$$

Then we have the following stationarity conditions,

$$\sum_{j \in \mathcal{C}_F^1} \lambda_j^1 + \sum_{i=2}^{r} \lambda_i = 1, \tag{7}$$

$$\sum_{j \in \mathcal{C}_F^1} \lambda_j^1 J_{1j}(\gamma_j) + \sum_{i \in \Gamma \cup \mathcal{S}_w} \lambda_i I_1(x(\alpha_1^*, \alpha_i^*)) = \nu, \tag{8}$$

$$\lambda_i I_i(x(\alpha_1^*, \alpha_i^*)) = \nu, \quad i \in \Gamma, \tag{9}$$

$$\lambda_i \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j) = \nu, \quad i \in \mathcal{S}_b, \tag{10}$$

$$\lambda_i \left[I_i(x(\alpha_1^*, \alpha_i^*)) + \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j) \right] = \nu, \quad i \in \mathcal{S}_w, \tag{11}$$

and the complementary slackness conditions,

$$\lambda_j^1 [\alpha_1^* J_{1j}(\gamma_j) - z] = 0, \quad j \in \mathcal{C}_F^1, \tag{12}$$

$$\lambda_i [\alpha_1^* I_1(x(\alpha_1^*, \alpha_i^*)) + \alpha_i^* I_i(x(\alpha_1^*, \alpha_i^*)) - z] = 0, \quad i \in \Gamma, \tag{13}$$

$$\lambda_i \left[\alpha_i^* \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j) - z \right] = 0, \quad i \in \mathcal{S}_b, \tag{14}$$

$$\lambda_i \left[\alpha_1^* I_1(x(\alpha_1^*, \alpha_i^*)) + \alpha_i^* I_i(x(\alpha_1^*, \alpha_i^*)) + \alpha_i^* \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j) - z \right] = 0, \quad i \in \mathcal{S}_w. \tag{15}$$

Suppose $\lambda_i = 0$ for some $i \in \Gamma \cup \mathcal{S}_b \cup \mathcal{S}_w$. Since $\alpha_i > 0$ for all $i \leq r$, the rate functions in Equations (9) and (11) are strictly greater than zero. By assumption, the rate functions $J_{ij}(\gamma_j)$ in (10) are also strictly greater than zero. Thus $\nu = 0, \lambda_i = 0$ for all $i \in \Gamma \cup \mathcal{S}_b \cup \mathcal{S}_w$, and $\sum_{j \in \mathcal{C}_F^1} \lambda_j^1 = 1$. Therefore at least one $\lambda_j^1 > 0$. Then in Equation (8), it must hold 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 (13)–(15), then complementary slackness implies each of these constraints is binding. Therefore a solution satisfying the KKT conditions of Problem Q must satisfy the equality constraints corresponding to $i \in \Gamma \cup \mathcal{S}_b \cup \mathcal{S}_w$ in Problem Q^*. □

Since the objective of the problem in (5) is a continuous function of $\boldsymbol{\alpha}$ on a compact set, an optimal solution $\boldsymbol{\alpha}^*$ to Problem Q^* exists. Let the optimal value at $\boldsymbol{\alpha}^*$ be denoted z^*. Proposition 2 states that $\boldsymbol{\alpha}^*$ is unique (see online supplement, §B for a proof).

Proposition 2. *The optimal solution $\boldsymbol{\alpha}^*$ to Problem Q^* is unique.*

The structure of 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 $\tilde{Q}^*$, and let $\tilde{z}^*$ be the optimal value at the optimal solution $\tilde{\boldsymbol{\alpha}}^* = (\tilde{\alpha}_1^*, \ldots, \tilde{\alpha}_r^*)$ to Problem $\tilde{Q}^*$. (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 result formally, where we note that Problem $\tilde{Q}^*$ also has a unique optimal solution by a proof similar to that of Proposition 2.

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

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

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

$$= \alpha_1 I_1(x(\alpha_1, \alpha_\ell)) + \alpha_\ell \left[I_\ell(x(\alpha_1, \alpha_\ell)) + \sum_{j \in \mathcal{C}_I^l} J_{\ell j}(\gamma_j) \right],$$

$$\textit{for all } i \in \Gamma,\ k \in \mathcal{S}_b,\ \ell \in \mathcal{S}_w,$$

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

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

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. Let us simplify the KKT equations for Problem Q as follows. 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 1, it follows that $\nu > 0$. Dividing (8) by ν and appropriately substituting in values from Equations (9)–(11), we find

$$\frac{\sum_{j \in \mathcal{C}_F^1} \lambda_j^1 J_{1j}(\gamma_j)}{\nu} + \sum_{i \in \Gamma} \frac{I_1(x(\alpha_1^*, \alpha_i^*))}{I_i(x(\alpha_1^*, \alpha_i^*))} + \sum_{i \in \mathcal{S}_w} \frac{I_1(x(\alpha_1^*, \alpha_i^*))}{I_i(x(\alpha_1^*, \alpha_i^*)) + \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j)} = 1. \quad (16)$$

By logic similar to that given in the proof of Proposition 1 and the simplification provided in (16), omitting terms with λ_j^1 in Equation (16) yields condition C1 as a KKT condition for Problem $\tilde{Q}^*$. Taken together, C0 and C1 create a fully-specified system of equations that form the KKT conditions for Problem $\tilde{Q}^*$. A solution $\boldsymbol{\alpha}$ is thus optimal to Problem $\tilde{Q}^*$ if and only if it solves C0 and C1. Let $\mathcal{D}(Q^*)$ and $\mathcal{D}(\tilde{Q}^*)$ denote the feasible regions of Problems Q^* and $\tilde{Q}^*$, respectively.

Proof of Claim (i). ($\Rightarrow$) Suppose $\tilde{\boldsymbol{\alpha}}^*$ solves C0 and C1, and $\min_{j \in \mathcal{C}_F^1} \tilde{\alpha}_1^* J_{1j}(\gamma_j) \geq \tilde{z}^*$. Then $\tilde{\boldsymbol{\alpha}}^* \in \mathcal{D}(Q^*)$. Since the objective functions of Problems Q^* and $\tilde{Q}^*$ are identical, and $\mathcal{D}(Q^*) \subset \mathcal{D}(\tilde{Q}^*)$, we know that $z^* \leq \tilde{z}^*$. Therefore $\tilde{\boldsymbol{\alpha}}^* \in \mathcal{D}(Q^*)$ implies $\tilde{\boldsymbol{\alpha}}^*$ is an 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 $\tilde{Q}^*$, then $\tilde{\boldsymbol{\alpha}}^*$ solves C0 and C1. Furthermore, 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 $\min_{j \in \mathcal{C}_F^1} \tilde{\alpha}_1^* J_{1j}(\gamma_j) < \tilde{z}^*$. Furthermore, 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 (16). By the complementary slackness condition in Equation (12), $\min_{j \in \mathcal{C}_F^1} \alpha_1^* J_{1j}(\gamma_j) = z^*$, and hence C2 holds for $\boldsymbol{\alpha}^*$. □

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 $\boldsymbol{\alpha}^*$)
1. Solve the nonlinear system C0, C1 to obtain $\tilde{\boldsymbol{\alpha}}^*$ and $\tilde{z}^*$;
2. **if** $\min_j \tilde{\alpha}_1^* J_{1j}(\gamma_j) \geq \tilde{z}^*$, **then**
3. **return** $\boldsymbol{\alpha}^* = \tilde{\boldsymbol{\alpha}}^*$.
4. **else**
5. Solve the nonlinear system C0, C2 to obtain $\boldsymbol{\alpha}^*$.
6. **return** $\boldsymbol{\alpha}^*$;
7. **end if**.

An underlying assumption in Theorem 4 is that there is 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

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	$g_1 \in [-1.5, 0)$	1.0
2	2.0	1.0	-1.0	1.0
3	2.0	1.0	-2.0	1.0

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 results of Glynn and Juneja (2004).

To illustrate the proposed conceptual algorithm, we present Example 1 for the case in which the underlying random variables are iid replicates from a normal distribution.

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 $i \le r$. The relevant rate functions for the normal case are

$$\min_{j\in\mathcal{C}_F^1} \alpha_1 J_{1j}(\gamma_j) = \min_{j\in\mathcal{C}_F^1} \frac{\alpha_1(\gamma_j - g_{1j})^2}{2\sigma^2_{g_{1j}}}, \quad i \in \{1\},$$

$$\alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i)) = \frac{(h_1 - h_i)^2}{2(\sigma^2_{h_1}/\alpha_1 + \sigma^2_{h_i}/\alpha_i)}, \quad i \in \Gamma,$$

$$\alpha_i \sum_{j\in\mathcal{C}_I^i} J_{ij}(\gamma_j) = \alpha_i \sum_{j\in\mathcal{C}_I^i} \frac{(\gamma_j - g_{ij})^2}{2\sigma^2_{g_{ij}}}, \quad i \in \mathcal{S}_b, \quad \text{and}$$

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

for $i \in \mathcal{S}_w$. Taking partial derivatives with respect to α_i, we find

$$\frac{\partial}{\partial\alpha_1}\left[\alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i))\right] = I_1(x(\alpha_1, \alpha_i)) = \frac{(\sigma^2_{h_1}/\alpha_1^2)(h_1 - h_i)^2}{2(\sigma^2_{h_1}/\alpha_1 + \sigma^2_{h_i}/\alpha_i)^2},$$

$$\frac{\partial}{\partial\alpha_i}\left[\alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i))\right] = I_i(x(\alpha_1, \alpha_i)) = \frac{(\sigma^2_{h_i}/\alpha_i^2)(h_1 - h_i)^2}{2(\sigma^2_{h_1}/\alpha_1 + \sigma^2_{h_i}/\alpha_i)^2}.$$

Let $\gamma = 0$, and let the mean and variance of each objective and constraint be as in Table 2.

Note that $\Gamma = \{2, 3\}$ and $\mathcal{S}_b = \mathcal{S}_w = \varnothing$. Since the allocation to systems in Γ is based on their "scaled distance" from system 1, and systems 2 and 3 are equal in this respect, we expect that they will receive equal allocation. To demonstrate the effect of the constraint g_1 on the allocation to system 1, we vary g_1 in the interval $[-1.5, 0)$. Solving for the optimal allocation as a function of g_1 yields the allocations in Figure 1 and the rate z^* in Figure 2.

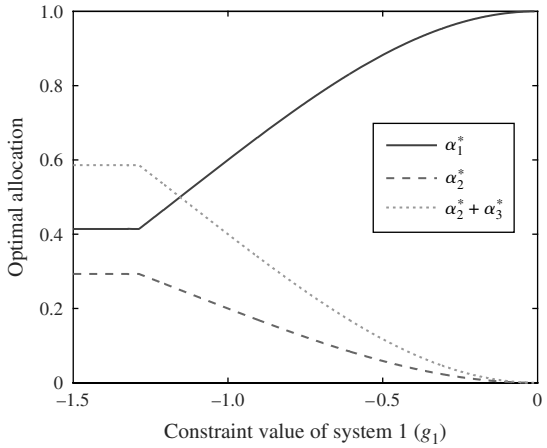

Figure 1 Graph of g_1 vs. Allocation for the Systems in Example 1

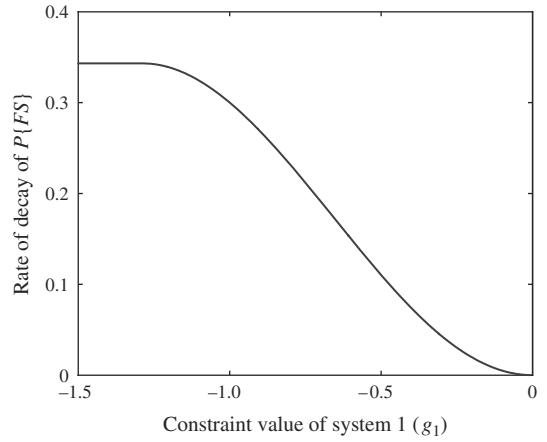

Figure 2 Graph of g_1 vs. Rate of Decay of $P\{FS\}$ for Example 1

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. For values of g_1 smaller than -1.2872, the feasibility of system 1 is no longer binding in this example since the rate function of the feasibility of system 1 is sufficiently large. 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 $z^* = 0.3431$. □

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 §6.2. All proofs for this section appear in the online supplement, §C.

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 $\hat{I}_i^m(x)$, $\hat{J}_{ij}^m(y)$, $i \le r$, $j \le s$, of the corresponding rate functions $I_i(x)$, $J_{ij}(y)$, $i \le r$, $j \le s$. Such consistent estimators are easy to construct when the distributional families underlying the true rate functions $I_i(x)$, $J_{ij}(y)$, $i \le r$, $j \le s$, are known or assumed. For example, suppose H_{ik}, $k = 1, 2, \ldots, m$, 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 $I_i(x) = (x - h_i)^2/(2\sigma_{h_i}^2)$ is then $\hat{I}_i^m(x) = (x - \hat{H}_i)^2/(2\hat{\sigma}_i^2)$, where $\hat{H}_i$ and $\hat{\sigma}_{h_i}$ are the sample mean and sample standard deviation of $H_{ik}, k = 1, 2, \ldots, m$, 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 (see, e.g., Dembo and Zeitouni 1998, p. 26) of the cumulant generating function estimator

$$\hat{I}_i^m(x) = \sup_\theta (\theta x - \hat{\Lambda}_{H_i}^m(\theta)), \tag{17}$$

where $\hat{\Lambda}_{H_i}^m(\theta) = \log\big(\frac{1}{m}\sum_{k=1}^m \exp(\theta H_{ik})\big)$. In what follows, to preserve generality, our discussion pertains to estimators of the type displayed in (17). By arguments analogous to those in Glynn and Juneja (2004), the estimator in (17) is consistent.

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

- $\hat{1}(m) := \arg\min_i\{\hat{H}_i(m)\colon \hat{G}_{ij}(m) \le \gamma_j \text{ for all } j \le s\}$ is the estimated best feasible system;
- $\hat{\Gamma}(m) := \{i\colon \hat{G}_{ij}(m) \le \gamma_j \text{ for all } j \le s,\ i \ne \hat{1}(m)\}$ is the estimated set of suboptimal feasible systems;
- $\hat{\mathcal{S}}_b(m) := \{i\colon \hat{H}_{\hat{1}(m)}(m) \ge \hat{H}_i(m) \text{ and } \hat{G}_{ij}(m) > \gamma_j \text{ for some } j \le s\}$ is the estimated set of infeasible, better systems;
- $\hat{\mathcal{S}}_w(m) := \{i\colon \hat{H}_{\hat{1}(m)}(m) < \hat{H}_i(m) \text{ and } \hat{G}_{ij}(m) > \gamma_j \text{ for some } j \le s\}$ is the estimated set of infeasible, worse systems;
- $\hat{\mathcal{C}}_F^i(m) := \{j\colon \hat{G}_{ij}(m) \le \gamma_j\}$ is the set of constraints on which system i is estimated feasible;
- $\hat{\mathcal{C}}_I^i(m) := \{j\colon \hat{G}_{ij}(m) > \gamma_j\}$ is the set of constraints on which system i is estimated infeasible.

Since Assumption 2 implies $\hat{H}_i(m) \to h_i$ wp1 and $\hat{G}_{ij}(m) \to g_{ij}$ wp1 for all $i \le r$ and $j \le s$, and the numbers of systems and constraints are finite, all estimated sets converge to their true counterparts wp1 as $m \to \infty$. (See §3.1 for a rigorous definition of the convergence of sets.) Proposition 3 formally states this result.

Proposition 3. *Under Assumption* 2, $\hat{1}(m) \to$ *system* 1, $\hat{\Gamma}(m) \to \Gamma$, $\hat{\mathcal{S}}_b(m) \to \mathcal{S}_b$, $\hat{\mathcal{S}}_w(m) \to \mathcal{S}_w$, $\hat{\mathcal{C}}_F^i(m) \to \mathcal{C}_F^i$, *and* $\hat{\mathcal{C}}_I^i(m) \to \mathcal{C}_I^i$ *wp*1 *as* $m \to \infty$.

Let $\hat{\boldsymbol{\alpha}}^*(m)$ denote the estimator of the optimal allocation vector $\boldsymbol{\alpha}^*$ obtained by replacing the rate functions $I_i(x)$, $J_{ij}(x)$, $i \le r$, $j \le s$, appearing in conditions C0, C1, and C2 with their corresponding estimators $\hat{I}_i^m(x)$, $\hat{J}_{ij}^m(x)$, $i \le r$, $j \le s$, obtained through sampling, and then using Algorithm 1. Since the search space $\{\boldsymbol{\alpha}\colon \sum_{i=1}^r \alpha_i = 1, \alpha_i \ge 0 \text{ for all } i \le r\}$ 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, where the proof is a direct application of results in the stochastic root-finding literature (see, e.g., Pasupathy and Kim 2011, Theorem 5.7). Before stating Theorem 5, we state two lemmas.

Lemma 1. *Suppose Assumption* 3 *holds. Then there exists* $\epsilon > 0$ *such that* $\hat{I}_i^m(x) \to I_i(x)$ *as* $m \to \infty$ *uniformly in* $x \in [h_\ell - \epsilon, h_u + \epsilon]$ *wp*1, *for all* $i \in \{1\} \cup \Gamma \cup \mathcal{S}_w$.

Lemma 2. *Let the system of equations C*0 *and C*1 *be denoted* $f_1(\boldsymbol{\alpha}) = 0$, *and let the system of equations C*0 *and C*2 *be denoted by* $f_2(\boldsymbol{\alpha}) = 0$, *where* f_1 *and* f_2 *are vector-valued functions with compact support* $\sum_{i=1}^r \alpha_i = 1$, $\boldsymbol{\alpha} \ge 0$. *Let the estimators* $\hat{F}_1^m(\boldsymbol{\alpha})$ *and* $\hat{F}_2^m(\boldsymbol{\alpha})$ *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 estimators. If Assumption* 3 *holds, then the functional sequences* $\hat{F}_1^m(\boldsymbol{\alpha}) \to f_1(\boldsymbol{\alpha})$ *and* $\hat{F}_2^m(\boldsymbol{\alpha}) \to f_2(\boldsymbol{\alpha})$ *uniformly in* $\boldsymbol{\alpha}$ *as* $m \to \infty$ *wp*1.

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* $m \to \infty$ *wp*1.

In addition to the consistency of $\hat{\boldsymbol{\alpha}}^*(m)$, one may ask what minimum sample size guarantees that

$P\{|\hat{\boldsymbol{\alpha}}^*(m) - \boldsymbol{\alpha}^*| \le \epsilon\} \ge 1 - \beta$ for given $\epsilon, \beta > 0$. While results of this sort are widely available, they tend to be overly conservative and of limited value from the standpoint of implementation (Pasupathy and Kim 2011).

6.2. A Sequential Algorithm for Implementation

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

Algorithm 2 (Sequential algorithm for implementation)

Require: Number of pilot samples $\delta_0 > 0$; number of samples between allocation vector updates $\delta > 0$; and a minimum-sample vector $\boldsymbol{\epsilon} = \{\epsilon_1, \ldots, \epsilon_r\} > 0$.

1: Initialize: collect δ_0 samples from each system $i \le r$.
2: Set $n = r\delta_0$, $n_i = \delta_0$. {Initialize total simulation effort and effort for each system.}
3: Update the sample means $\hat{H}_i$, $\hat{G}_{ij}$, estimated sets $\hat{1}(n)$, $\hat{\Gamma}(n)$, $\hat{\mathcal{S}}_b(n)$, $\hat{\mathcal{S}}_w(n)$, and rate function estimators $\hat{I}_i^{n_i}(x)$, $\hat{J}_{ij}^{n_i}(\gamma_j)$, for all $i \le r$, $j \le s$.
4: **if** no systems are estimated feasible, **then**
5: Set $\hat{\boldsymbol{\alpha}}^*(n) = (1/r, 1/r, \ldots, 1/r)$.
6: **else**
7: Solve the system C0, C1 using rate function estimators to obtain $\hat{\hat{\boldsymbol{\alpha}}}^*(n)$ and $\hat{\hat{z}}^*(n)$.
8: **if** $\min_j \hat{\hat{\alpha}}_{\hat{1}}^*(n)\hat{J}_{\hat{1}j}^{n_{\hat{1}}}(\gamma_j) \ge \hat{\hat{z}}^*(n)$, **then**
9: $\hat{\boldsymbol{\alpha}}^*(n) = \hat{\hat{\boldsymbol{\alpha}}}^*(n)$.
10: **else**
11: Solve the system C0, C2 using rate function estimators to obtain $\hat{\boldsymbol{\alpha}}^*(n)$.
12: **end if**
13: **end if**
14: Collect one sample at each of the systems X_k, $k = 1, 2, \ldots, \delta$, where the X_k's are iid random variates having probability mass function $\hat{\boldsymbol{\alpha}}^*(n)$ on support $\{1, 2, \ldots, r\}$, and update $n_{X_k} = n_{X_k} + 1$.
15: Set $n = n + \delta$ and update $\bar{\boldsymbol{\alpha}}_n = \{\bar{\alpha}_{1,n}, \ldots, \bar{\alpha}_{r,n}\} = \{n_1/n, n_2/n, \ldots, n_r/n\}$.
16: **if** $\bar{\boldsymbol{\alpha}}_n > \boldsymbol{\epsilon}$, **then**
17: Set $\delta^+ = 0$.
18: **else**
19: Collect one sample from each system in the set of systems receiving insufficient sample $\mathcal{I}_n = \{i\colon \bar{\alpha}_{i,n} < \epsilon_i\}$.
20: Update $n_i = n_i + 1$ for all $i \in \mathcal{I}_n$. Set $\delta^+ = |\mathcal{I}_n|$.
21: **end if**
22: Set $n = n + \delta^+$ and go to step 3.

The essential idea in Algorithm 2 is straightforward. At the end of each iteration, the optimal allocation vector is estimated using the estimated rate functions constructed using samples obtained from the various systems. Systems are chosen for sampling in the subsequent iteration by using the estimated optimal allocation vector as the sampling distribution. Since $\boldsymbol{\alpha}^* > 0$, the sequential algorithm should sample from each system infinitely often. To ensure systems with small allocations continue to be sampled, we assume knowledge of an "indifference zone" vector $\boldsymbol{\epsilon} > 0$ such that if the actual proportion of sample expended at each system in Algorithm 2, defined as $\bar{\boldsymbol{\alpha}}_n = \{n_1/n, n_2/n, \ldots, n_r/n\}$, falls below $\boldsymbol{\epsilon}$, we sample once from each system receiving insufficient sample. All elements of $\boldsymbol{\epsilon}$ should be "small" relative to $1/r$.

In a context where the distributional family underlying the simulation observations is known or assumed, the rate function estimators should be estimated in step 3 accordingly—by simply estimating the distributional parameters appearing within the expression for the rate function. Also, Algorithm 2 provides flexibility on how often the optimal allocation vector is re-estimated through the algorithm parameter δ. The choice of the parameter δ will depend on the particular problem, and specifically, on how expensive the simulation execution is relative to solving the nonlinear systems in steps 7 and 11. Lastly, Algorithm 2 relies on fully sequential and simultaneous observation of the objective and constraint functions. Deviation from these assumptions renders the present context inapplicable.

7. Numerical Examples

To illustrate the proposed allocation, we present the following numerical examples. First, we compare our proposed optimal allocation to the OCBA-CO allocation presented by Lee et al. (2011) under a normality assumption. In this comparison, we use the actual rate functions governing the simulation estimators since our primary objective is to highlight the theoretical differences between the proposed allocation and that of OCBA-CO. While both methods handle multiple constraints, we use the one-constraint case for ease of exposition. Then we present results of the implementation of the sequential estimator outlined in Algorithm 2 when the underlying random variables follow a normal distribution.

7.1. Comparison with OCBA-CO

Lee et al. (2011) describe an OCBA framework for simulation budget allocation in the context of constrained SO on finite sets under the assumption of normality. The work by Lee et al. (2011) is the only other asymptotic sample allocation result for constrained

simulation optimization on finite sets in the literature. Lee et al. (2011) divide the suboptimal systems into a "feasibility dominance" set and an "optimality dominance" set, defined as

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

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

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

$$\frac{\alpha_i}{\alpha_k} = \frac{((h_1 - h_k)/\sigma_{h_k})^2 \mathbb{I}_{k \in \mathcal{S}_O} + ((\gamma - g_k)/\sigma_{g_k})^2 \mathbb{I}_{k \in \mathcal{S}_F}}{((h_1 - h_i)/\sigma_{h_i})^2 \mathbb{I}_{i \in \mathcal{S}_O} + ((\gamma - g_i)/\sigma_{g_i})^2 \mathbb{I}_{i \in \mathcal{S}_F}} \quad \text{for all } i, k = 2, \ldots, r, \quad (18)$$

where $\mathbb{I}$ denotes the indicator variable. In OCBA-CO, only one term in each of the numerator and denominator of the right-hand side of Equation (18) is active at a time. This artifact of the set definitions and the approximations used in OCBA-CO may lead to suboptimal allocations for infeasible and worse systems. The following two examples are designed to highlight the theoretical differences between OCBA-CO and the proposed allocation.

Example 2. Suppose there are two systems and one constraint, with each (H_i, G_i) iid normally distributed. Let the means and variances be as given in Table 3, and let $\gamma = 0$.

Note the following features of this example: (i) since system 2 belongs to $\mathcal{S}_O$ for large enough n and $g_2 \in (0, 1.9]$, 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 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 in Figure 3, the allocation to system 2 should not remain constant as a function of g_2. In fact, for certain values of g_2, we give nearly all of the sample to system 2. □

Example 3. We retain the two systems from Example 2, except we fix $g_2 = 1.6$ and vary $\sigma^2_{h_1}$ in the interval $[0.2, 4]$ to explore the allocation to system 1 as a function of σ_{h_1}. 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.

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	$g_2 \in (0, 1.9]$	1.0

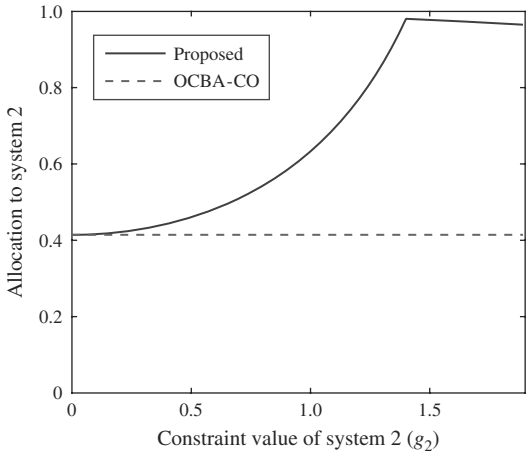

Figure 3 **Graph of g_2 vs. Allocation for the Systems in Example 2**

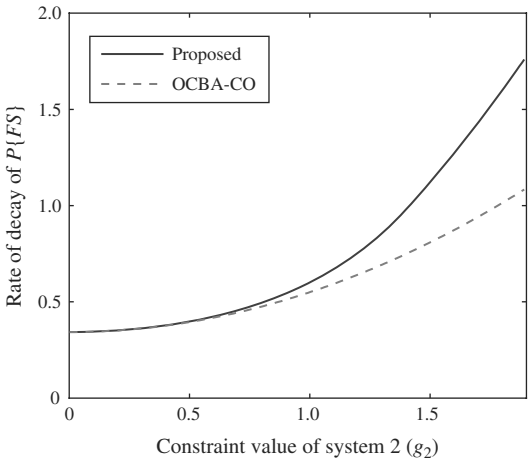

Figure 4 **Graph of g_2 vs. Rate of Decay of $P\{FS\}$ for the Systems in Example 2**

From Figure 5, 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. However, as a result of the $\alpha_1 \gg \alpha_i$ assumption, the OCBA-CO allocation constantly increases as $\sigma^2_{h_1}$ increases. 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 $\sigma^2_{h_1}$. In this scenario, the OCBA-CO allocation does not exploit the fact that when optimality is difficult to determine based on the objective,

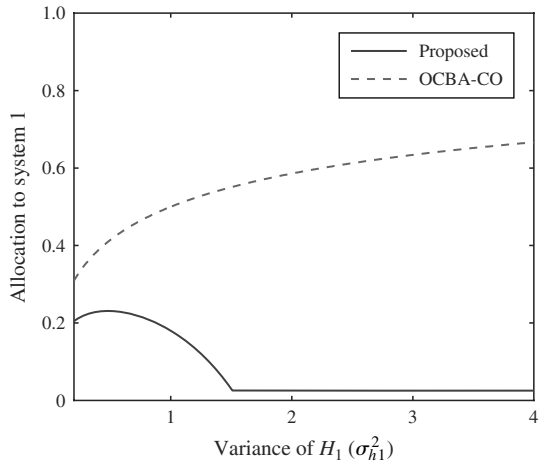

Figure 5 **Graph of $\sigma^2_{h_1}$ vs. Allocation for the Systems in Example 3**

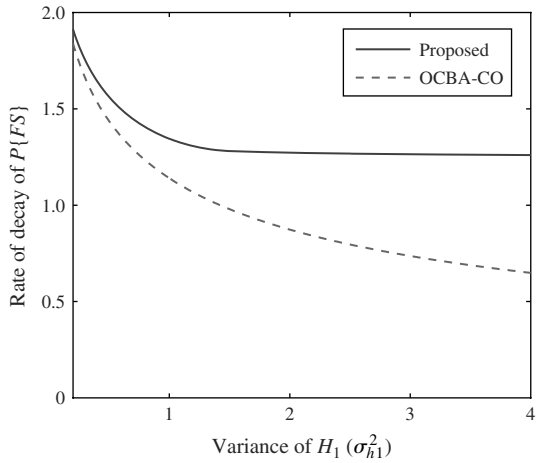

Figure 6 **Graph of $\sigma^2_{h_1}$ vs. Rate of Decay of $P\{FS\}$ for the Systems in Example 3**

a more efficient allocation can be achieved based on feasibility. □

7.2. Implementation of the Sequential Algorithm

We now present two examples of the implementation of the sequential Algorithm 2. In the first example, we calculate 500 sample paths of the sequential algorithm for one Problem P (see §1.2) to show the variability in convergence across sample paths. In the second, we calculate one sample path of the sequential algorithm for each of 500 Problems P to show variability in convergence across problem instances.

Example 4 (Normal: 500 Sample Paths of One Problem P). Suppose that for each system i and constraint j, we may obtain iid replicates of the random variables H_i and G_{ij}, where H_i has distribution normal$(h_i, \sigma^2_{h_i})$ and G_{ij} has distribution normal$(g_{ij}, \sigma^2_{g_{ij}})$. We use the sequential Algorithm 2 to solve the Problem P for five systems in the presence of two constraints, $\gamma_1 = \gamma_2 = 0$, using algorithm parameters of initial sample size $\delta_0 = 20$, sample size between estimated optimal allocation updates $\delta = 20$, and minimum-sample parameter $\epsilon_i = 10^{-6}$ for all $i \le r$. Randomly generated means for the five systems are given in Table 4, where $\sigma^2_{h_i} = \sigma^2_{g_{i1}} = \sigma^2_{g_{i2}} = 1$ for all $i \le r$, $j \le s$.

Table 4 **Means for Example 4, $z^* = 0.1113$**

Set	α^*_i	h_i	g_{i1}	g_{i2}
$\mathcal{S}_b$	0.3526	−2.0127	0.7946	−0.8223
System 1	0.1835	−1.3651	−1.3272	−1.2408
$\mathcal{S}_w$	0.3407	−0.3887	0.3792	0.4194
$\mathcal{S}_w$	0.1078	−0.3909	−0.1299	1.2115
Γ	0.0154	2.5915	−0.9288	−1.4952

Figure 7 displays the 90th, 75th, 50th, 25th, and 10th sample percentiles of the optimality gap in the rate of decay of the $P\{FS\}$ of the sampling algorithm, $z^* - z(\bar{\boldsymbol{\alpha}}_n)$, calculated across 500 sample paths. Since z^* is the fixed optimal rate of decay of the $P\{FS\}$ for the Problem P specified in Table 4, the optimality gap is necessarily positive. Figure 7 also displays the optimality gap for equal allocation, which remains fixed across all n. As expected, the optimality gap for Algorithm 2 appears to converge to near zero, with 90 percent of sample paths achieving a faster rate of decay of $P\{FS\}$ than equal allocation by sample size $n = 300$. □

Example 5 (Normal: One Sample Path of Each of 500 Problems P). Now suppose we have 500 randomly generated problems like the one presented in Example 4. For one sample path of each of these 500 problems, we calculate the sample quantiles of the optimality gap. This example is intended to provide support for the robustness of the convergence across different problem scenarios.

We retain the values of the algorithm parameters δ_0, δ, and $\boldsymbol{\epsilon}$ from the previous example, and let the constraints $\gamma_1 = \gamma_2 = 0$. Each Problem P was created by uniformly and independently generating objective values h_i, $i \le 5$, in the interval $[-3, 3]$. To ensure the

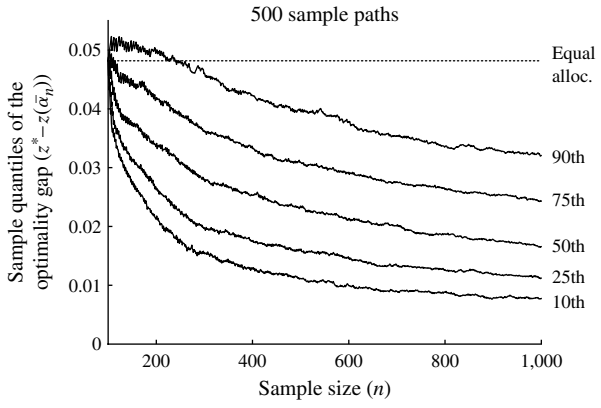

Figure 7 **Sample Distribution of the Optimality Gap for Example 4**

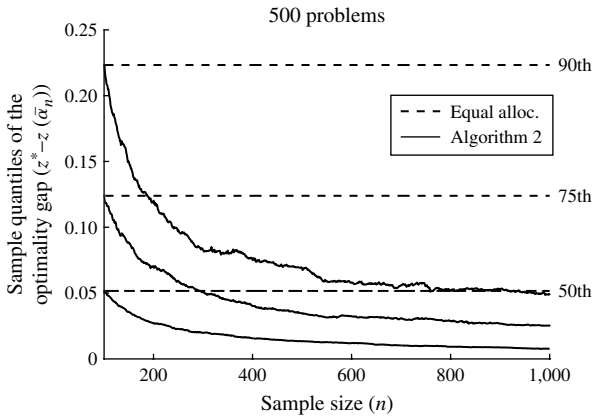

Figure 8 Sample Distribution of the Optimality Gap for Example 5

existence of system 1 and nonempty Γ, two of the five systems had both of their constraint values g_{ij}, $j \le 2$, independently and uniformly generated in the interval $[-3, 0]$, which is the feasible region. The constraint values g_{ij}, $j \le 2$, for all other systems were generated in the interval $[-3, 3]$. As in the previous example, all variances equal one. To ensure numerically distinguishable systems and constraints in the context of the scaling of these problems, we ensured $|h_1 - h_i| > 0.05$ and $|g_{ij} - \gamma_j| > 0.05$ for all $i \le 5$, $j \le 2$.

Figure 8 shows the 90th, 75th, and 50th sample percentiles of the optimality gap, $z^* - z(\bar{\boldsymbol{\alpha}}_n)$, as a function of sample size n for equal allocation and for the sequential Algorithm 2. The sample paths appear to converge across a variety of problems. □

Preliminary results with regard to implementation in the context of Bernoulli random variables indicates that, at least for the random problems we tested, there is a large initial variance in the optimality gap of Algorithm 2 across sample paths. This finding is consistent with that of Broadie et al. (2007), who find that a modified version of the algorithm provided in Glynn and Juneja (2004) is highly sensitive to initial sample size in the context of exponential random variables. We suspect that this sensitivity occurs because for some problem instances, even small error in estimating $\boldsymbol{\alpha}^*$ yields a highly suboptimal true rate of convergence due to a steep gradient of the rate function. Further investigation of this phenomenon is a topic of future research.

8. Remarks on the Independence Assumption

Throughout this paper, we have assumed independence of the objective and constraint function estimators obtained from the simulation. This scenario naturally arises in the context of a single normally distributed performance measure where the mean is the objective and the variance is used as a constraint. Additional examples appear in Lee et al. (2010), where independence between the performance measures is assumed in the context of multiobjective models. However, we acknowledge that such scenarios are relatively uncommon. In this section, we provide a relaxation of the independence assumption and a discussion of the scope of our results.

8.1. Asymptotic Independence

To see that the independence assumption, Assumption 1(2), may be relaxed, note that we require independence only to analyze the rate function of $P\{FS_i\}$ in §4. Under Assumption 1,

$$P\{FS_i\} = P\{\hat{H}_i \le \hat{H}_1\} \prod_{j=1}^{s} P\{\hat{G}_{ij} \le \gamma_j\} \prod_{j=1}^{s} P\{\hat{G}_{1j} \le \gamma_j\}, \quad (19)$$

which directly results in the rate function for the $P\{FS_i\}$ presented in Equation (3) and the final result in Theorem 2. However, to arrive at the rate function presented in Theorem 2, we do not actually require (19). We require only that

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \log P\{FS_i\} &= \lim_{n\to\infty} \frac{1}{n} \log P\{\hat{H}_i \le \hat{H}_1\} + \sum_{j=1}^{s} \lim_{n\to\infty} \frac{1}{n} \log P\{\hat{G}_{ij} \le \gamma_j\} \\ &\quad + \lim_{n\to\infty} \frac{1}{n} \log P\left\{\bigcap_{j=1}^{s} \hat{G}_{1j} \le \gamma_j\right\}.\end{aligned}$$

Thus, in the limit, we require the random variables to behave *as if* they were independent.

Toward mitigating the stringency of the independence assumption, we define a less-stringent type of independence which we call asymptotic independence. For a sequence of events A_n and B_n, define $c_n = P\{A_n \cap B_n\}/(P\{A_n\}P\{B_n\})$, where $P\{A_n\} > 0$, $P\{B_n\} > 0$ for all n. Then by definition $P\{A_n \cap B_n\} = c_n P\{A_n\}P\{B_n\}$, and supposing all relevant limits exist,

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \log P\{A_n \cap B_n\} &= \lim_{n\to\infty} \frac{1}{n} \log c_n + \lim_{n\to\infty} \frac{1}{n} \log P\{A_n\} + \lim_{n\to\infty} \frac{1}{n} \log P\{B_n\}.\end{aligned}$$

From this argument, the conditions for asymptotic independence become clear: we require that $\lim_{n\to\infty} \frac{1}{n} \log c_n = 0$. The following definition states this requirement formally and can easily be extended to consider more than two events.

Definition 1. Let A_n and B_n be events such that $P\{A_n\} > 0$ and $P\{B_n\} > 0$ for all n, $P\{A_n\} \to 0$, and $P\{B_n\} \to 0$ or $P\{B_n\} \to 1$. Then A_n and B_n exhibit *asymptotic independence* if

$$\lim_{n\to\infty} \frac{1}{n} \log[P\{A_n \cap B_n\}/(P\{A_n\}P\{B_n\})] = 0.$$

To the best of our knowledge, this type of independence has not been explicitly defined in the setting

of sample means. However, similar concepts exist in the context of extreme values (see, e.g., Resnick 2008, p. 290ff). The following example explores one type of asymptotic independence for bivariate normal distributions.

Example 6. For $k = 1, \ldots, n$, let (X_k, Y_k) be iid copies from a bivariate normal distribution with mean $(0, 0)$, variances equal to 1, and correlation $|\rho| < 1$. Then for $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ and $\bar{Y} = \frac{1}{n}\sum_{k=1}^{n} Y_k$, consider $P\{\bar{X} > a\}$ and $P\{\bar{Y} > b\}$ for some $a > 0$ and $b < 0$, where $a, b < \infty$ and $\rho \geq 0$. Since ρ is nonnegative, $P\{(\bar{X} > a) \cap (\bar{Y} > b)\} \geq P\{\bar{X} > a\}P\{\bar{Y} > b\}$, and therefore $c_n \geq 1$. Let $b < 0$. Then $P\{\bar{Y} > b\} \to 1$, and since

$$c_n = \underbrace{\frac{P\{(\bar{X} > a) \cap (\bar{Y} > b)\}}{P\{\bar{X} > a\}}}_{\leq 1} \underbrace{\frac{1}{P\{\bar{Y} > b\}}}_{\to 1},$$

$\limsup c_n \leq 1$. Since $c_n \geq 1$ and $\limsup c_n \leq 1$, then $\lim_{n\to\infty} c_n = 1$ and $\lim_{n\to\infty} \frac{1}{n} \log c_n = 0$. □

From this analysis, we may deduce that for the two events A_n and B_n, negative correlation results in $c_n \leq 1$, which *increases* the rate function over what is observed in the independent case. Likewise, positive correlation results in $c_n \geq 1$ which *decreases* the rate function below what is observed in the independent case. The interaction between correlation and the rate function becomes more complex as we consider more than two events, as in the case of $P\{FS_i\}$ with multiple constraints. In the next section, we discuss the scope of these results when the asymptotic independence assumption is not satisfied.

8.2. Scope of Results

One may ask, how do the results derived in this paper apply to contexts in which independence between objective and constraint function estimators cannot be guaranteed? To obtain useful results on sampling in the context of our problem statement, we argue that one must assume (i) the objective and constraint function estimators are independent, (ii) the nature of the distributions underlying the estimators are known, or (iii) the number of systems tends to infinity. While theoretically valuable, a characterization of the optimal allocation without such assumptions would have limited practical value due to its abstract nature and issues regarding implementation. As such, sampling frameworks derived under one of these assumptions should be seen as an approximate guide to simulation allocation obtained through the analysis of an imperfect but tractable model.

This paper relies on (i) to make the first strides on the question of efficient sampling within stochastically constrained SO on finite sets. Ongoing research is focused on (ii) and (iii), that is, deriving analogous results by making a distributional assumption

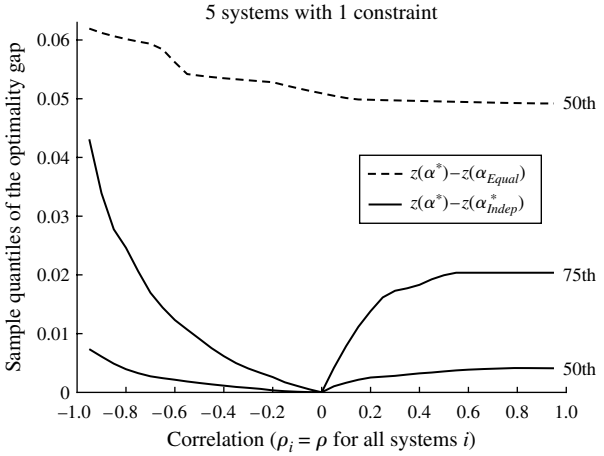

Figure 9 **Sample Quantiles of the Optimality Gap as a Function of Correlation Using the True Optimal Allocation Under the Assumption of Bivariate Normality from Hunter et al. (2011)**

and by sending the number of systems to infinity. While it is difficult to make statements on the effect of correlation under the general terms of this paper, work on (ii) appearing in Hunter (2011) and Hunter et al. (2011) indicates that, under the assumption of bivariate normality, dependence indeed affects the optimal allocation policy. However, the effect of correlation may be "small enough" to matter little during implementation. As evidence, we present Figure 9 to show the effect of correlation in the bivariate normal case. In this figure, let $z(\boldsymbol{\alpha}^*)$ denote the true optimal rate of decay with correlation under the assumption of bivariate normality. For 100-five-system, one-constraint Problems P, generated in a manner similar to the systems in Example 5, Figure 9 shows the sample quantiles of the true optimality gap for the independent model as a function of correlation. The 50th sample quantile of the optimality gap of equal allocation is also shown. For simplicity, at each correlation value, all systems in a randomly generated Problem P have the same correlation between the objective and constraint function.

We present these preliminary results to demonstrate that, even when correlation is present, using the independent model may still result in significant gains over equal allocation. Further analysis of the effect of correlation on the optimal allocation is beyond the scope of this paper.

9. 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) and Szechtman and Yücesan (2008), we take the first steps toward answering these questions.

To identify the relationship between sampling and error-probability decay, we divide the competing systems into four sets. 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.

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. In settings where the underlying distribution is not known or assumed, estimating the underlying rate function using a Taylor's series approximation up to a few terms might prove a viable alternative.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/ijoc.1120.0519.

Acknowledgments

The authors were supported in part by the Office of Naval Research [Grants N000140810066, N000140910997, and N000141110065]. The first author was also supported in part by National Science Foundation [Grants CMMI 0758441 and CMMI 0800688].

References

Andradóttir S (2006) An overview of simulation optimization via random search. Henderson SG, Nelson BL, eds. *Simulation, Handbooks in Oper. Res. and Management Sci.* (Elsevier), 617–631.

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

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

Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press, New York).

Branke J, Chick SE, Schmidt C (2007) Selecting a selection procedure. *Management Sci.* 53(12):1916–1932.

Broadie M, Han M, Zeevi A (2007) Implications of heavy tails on simulation-based ordinal optimization. *Proc. 2007 Winter Simulation Conf.* (IEEE Press, Piscataway, NJ), 439–447.

Bucklew JA (2003) *Introduction to Rare Event Simulation*. Springer Series in Statistics (Springer, New York).

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

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

Glynn PW, Juneja S (2004) A large deviations perspective on ordinal optimization. *Proc. 2004 Winter Simulation Conf.* (IEEE Press, Piscataway, NJ), 577–585.

Glynn PW, Juneja S (2006) Ordinal optimization: A large deviations perspective. Working Paper Series, Indian School of Business. Accessed August 7, 2012, http://www.isb.edu/faculty/Working_Papers_pdfs/Ordinal_Optimization.pdf.

Hunter SR (2011) Sampling laws for stochastically constrained simulation optimization on finite sets. Ph.D. thesis, Virginia Tech, Blacksburg, Virginia.

Hunter SR, Pujowidianto NA, Chen C-H, Lee LH, Pasupathy R, Yap CM (2011) Optimal sampling laws for stochastically constrained simulation optimization on finite sets: The bivariate normal case. *Proc. 2011 Winter Simulation Conf.* (IEEE Press, Piscataway, NJ), 4294–4302.

Kim S-H, Nelson BL (2006) Selecting the best system. Henderson SG, Nelson BL, eds. *Simulation*, Vol. 13. *Handbooks in Oper. Res. and Management Sci.* (Elsevier, Amstrdam, The Nethrlands), 501–534.

Lee LH, Chew EP, Teng S, Goldsman D (2010) Finding the nondominated pareto set for multiobjective simulation models. *IIE Trans.* 42(9):656–674.

Lee LH, Pujowidianto NA, Li L-W, Chen C-H, Yap CM (2011) Approximate simulation budget allocation for selecting the best design in the presence of stochastic constraints. *IEEE Trans. Automatic Control*. Forthcoming.

Ólafsson S, Kim J (2002) Simulation optimization. *Proc. 2002 Winter Simulation Conf.* (IEEE Press, Piscataway, NJ), 79–84.

Pasupathy R, Henderson SG (2006) A testbed of simulation-optimization problems. *Proc. 2006 Winter Simulation Conf.* (IEEE Press, Piscataway, NJ).

Pasupathy R, Henderson SG (2011) SimOpt: A library of simulation optimization problems. *Proc. 2011 Winter Simulation Conf.* (IEEE Press, Piscataway, NJ).

Pasupathy R, Kim S (2011) The stochastic root-finding problem: Overview, solutions, and open questions. *ACM Trans. Model. Comput. Simulation* 21(3):1–23.

Resnick SI (2008) *Extreme Values, Regular Variation, and Point Processes*, Springer Series in Oper. Res. and Financial Engrg. (Springer, New York).

Rudin W (1976) *Principles of Mathematical Analysis*, International Series in Pure and Applied Mathematics (McGraw-Hill, New York).

Spall JC (2003) *Introduction to Stochastic Search and Optimization* (John Wiley & Sons, Hoboken, NJ).

Szechtman R, Yücesan E (2008) A new perspective on feasibility determination. *Proc. 2008 Winter Simulation Conf.* (IEEE Press, Piscataway, NJ), 273–280.

Online Supplement for Optimal Sampling Laws for Stochastically Constrained Simulation Optimization on Finite Sets

Susan R. Hunter
School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853, USA, hunter@cornell.edu

Raghu Pasupathy
The Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA 24061, USA, pasupath@vt.edu

---

A. Useful Results

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

A consequence of the principle of the slowest term, Proposition A.2 states that the slowest among a set of rate functions is equivalent to the rate function of the slowest sequence.

Proposition A.2. *Let $a_i(n)$ be defined as in Proposition A.1. If $\lim_{n\to\infty} \frac{1}{n} \log a_i(n)$ exists for all i, then $\max_i \lim_{n\to\infty} \frac{1}{n} \log a_i(n) = \lim_{n\to\infty} \frac{1}{n} \log \left(\max_i a_i(n) \right)$.*

Proof. The lower bound is $\max_i \lim_{n\to\infty} \frac{1}{n} \log a_i(n) = \lim_{n\to\infty} \frac{1}{n} \log \sum_{i=1}^k a_i(n) \geq \lim_{n\to\infty} \frac{1}{n} \log \max_i a_i(n)$. The upper bound is $\lim_{n\to\infty} \frac{1}{n} \log \sum_{i=1}^k a_i(n) \leq \lim_{n\to\infty} \frac{1}{n} \log \left(k \max_i a_i(n) \right) = \lim_{n\to\infty} \frac{1}{n} \log \max_i a_i(n)$. □

B. Proof of Proposition 2

Proof. Suppose $\boldsymbol{\alpha}^*$ is not unique. Then there exists another optimal point $\check{\boldsymbol{\alpha}}^*$ such that $\boldsymbol{\alpha}^* \neq \check{\boldsymbol{\alpha}}^*$, but $z^* = \check{z}^*$. We contradict this statement in each of the following cases.

Case 1: $\alpha_1^* = \check{\alpha}_1^*$. Suppose system 1 receives the same allocation in each of the optimal solutions $\boldsymbol{\alpha}^*$ and $\check{\boldsymbol{\alpha}}^*$. Since $\sum_{i=1}^r \alpha_i^* = \sum_{i=1}^r \check{\alpha}_i^* = 1$, and $\alpha_1^* = \check{\alpha}_1^*$, then there exist systems k and ℓ for which $\alpha_k^* > \check{\alpha}_k^*$ and $\alpha_\ell^* < \check{\alpha}_\ell^*$. Let the rate function for $P\{FS_i\}, i \neq 1$, be denoted $K_i(\alpha_1, \alpha_i)$, where this expression will vary depending on whether $i \in \Gamma, \mathcal{S}_b$, or $\mathcal{S}_w$. Then since the rate functions have strictly positive partial derivatives $\partial K_i(\alpha_1, \alpha_i)/\partial \alpha_i$ for all $i = 2, \ldots, r$ (see equation (6) for the partial derivatives for $i \in \Gamma$; the others follow trivially), it must be the case that $K_k(\alpha_1^*, \alpha_k^*) > K_k(\alpha_1^*, \check{\alpha}_k^*)$ and $K_\ell(\alpha_1^*, \alpha_\ell^*) < K_\ell(\alpha_1^*, \check{\alpha}_\ell^*)$. However this provides a contradiction, since feasibility in Problem Q^* requires $K_k(\alpha_1^*, \alpha_k^*) = K_\ell(\alpha_1^*, \alpha_\ell^*)$ and $K_k(\alpha_1^*, \check{\alpha}_k^*) = K_\ell(\alpha_1^*, \check{\alpha}_\ell^*)$.

Case 2: $\alpha_1^* \neq \check{\alpha}_1^*$ *and* $\alpha_i^* \neq \check{\alpha}_i^*$ *for at least one* $i \in \mathcal{S}_b$. Suppose $\alpha_k^* < \check{\alpha}_k^*$ for $k \in \mathcal{S}_b$. Then feasibility in Problem Q^* implies $\alpha_k^* \sum_{j \in \mathcal{C}_I^k} J_{kj}(\gamma_j) = z^* < \check{\alpha}_k^* \sum_{j \in \mathcal{C}_I^k} J_{kj}(\gamma_j) = \check{z}^*$, which is a contradiction.

Case 3: $\alpha_1^* \neq \check{\alpha}_1^*$ *and* $\alpha_i^* = \check{\alpha}_i^*$ *for all* $i \in \mathcal{S}_b$. Since $\alpha_1^* \neq \check{\alpha}_1^*$, the feasibility constraint for system 1 cannot be binding at optimality in Problem Q^*. Suppose we pre-allocate a total, fixed, and optimal portion of the sample $\alpha_{\mathcal{S}_b}^* = \check{\alpha}_{\mathcal{S}_b}^*$ to the systems in $\mathcal{S}_b$. Then we may consider a modified version of the problem in (5),

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

We now establish the strict concavity of the rate functions of systems in Γ. Denote the rate function as $K_k(\alpha_1, \alpha_k)$ for $k \in \Gamma$, and let the exponent T denote the matrix transpose. Since $(\alpha_1^*, \alpha_k^*) \neq (\check{\alpha}_1^*, \check{\alpha}_k^*)$, consider

$$\begin{aligned} & (\nabla K_k(\alpha_1, \alpha_i)|_{(\alpha_1^*, \alpha_k^*)} - \nabla K_k(\alpha_1, \alpha_i)|_{(\check{\alpha}_1^*, \check{\alpha}_k^*)})^T ((\alpha_1^*, \alpha_k^*)^T - (\check{\alpha}_1^*, \check{\alpha}_k^*)^T) \\ & = (I_1(x(\alpha_1^*, \alpha_k^*)) - I_1(x(\check{\alpha}_1^*, \check{\alpha}_k^*)))(\alpha_1^* - \check{\alpha}_1^*) + (I_k(x(\alpha_1^*, \alpha_k^*)) - I_k(x(\check{\alpha}_1^*, \check{\alpha}_k^*)))(\alpha_k^* - \check{\alpha}_k^*) \\ & = (\alpha_1^* I_1(x(\alpha_1^*, \alpha_k^*)) + \alpha_k^* I_k(x(\alpha_1^*, \alpha_k^*))) - (\alpha_1^* I_1(x(\check{\alpha}_1^*, \check{\alpha}_k^*)) + \alpha_k^* I_k(x(\check{\alpha}_1^*, \check{\alpha}_k^*))) \\ & \quad + (\check{\alpha}_1^* I_1(x(\check{\alpha}_1^*, \check{\alpha}_k^*)) + \check{\alpha}_k^* I_k(x(\check{\alpha}_1^*, \check{\alpha}_k^*))) - (\check{\alpha}_1^* I_1(x(\alpha_1^*, \alpha_k^*)) + \check{\alpha}_k^* I_k(x(\alpha_1^*, \alpha_k^*))). \quad \text{(B.2)} \end{aligned}$$

By the strict convexity of $I_1(x)$ and $I_i(x)$, $x(\alpha_1, \alpha_i)$ uniquely minimizes $\inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x))$. Therefore (B.2) is strictly less than zero, and hence the rate functions in Γ are strictly concave in (α_1, α_i). By a similar proof, the rate functions for $i \in \mathcal{S}_w$ are also strictly concave in (α_1, α_i).

Since the minimum of strictly concave functions is strictly concave, the objective in (B.1) is strictly concave. Since the objective is strictly concave and the feasible set is convex, the problem in (B.1) has a unique solution. Let the rate achieved at this solution be z_b^*. If $z_b^* \neq z^* = \check{z}^*$, then $\alpha_{\mathcal{S}_b}^*$ cannot be optimal to the original Problem Q^*. Therefore $z_b^* = z^* = \check{z}^*$, but z_b^* is the optimal value at the unique optimal solution to (B.1). Therefore $\boldsymbol{\alpha}^* = \check{\boldsymbol{\alpha}}^*$, and we have a contradiction. □

C. Proofs for Section 6

Proof of Proposition 3. We 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 proposition follow in a similar fashion.

By Assumption 2,$\hat{G}_{ij}(m) \to g_{ij}$ 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. □

We omit the proof of Lemma 1 since it follows closely along the lines of the proofs presented in Glynn and Juneja (2004).

Proof of Lemma 2. We prove that the theorem holds in two steps. We first show $\alpha_1 \hat{I}_1^m(\hat{x}_m(\alpha_1, \alpha_i) + \alpha_i \hat{I}_i^m(\hat{x}_m(\alpha_1, \alpha_i))$ converges uniformly in $\boldsymbol{\alpha}$ as $m \to \infty$ wp1 for all $i \in \Gamma \cup \mathcal{S}_w$, where $\hat{x}_m(\alpha_1, \alpha_i) = \arg\inf_x(\alpha_1 \hat{I}_1^m(x) + \alpha_i \hat{I}_i^m(x))$. Next we show $\alpha_i \sum_{j \in \mathcal{C}_I^i} \hat{J}_{ij}^m(\gamma_j), i \in \mathcal{S}_b \cup \mathcal{S}_w$ and $\alpha_1 \hat{J}_{1j}^m(\gamma_j), j \in \mathcal{C}_F^1$ 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, $\hat{I}_i^m(\hat{x}_m(\alpha_1, \alpha_i)) > \delta$, 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 $x(\alpha_1, \alpha_i) = \arg\inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x)) \in [h_\ell, h_u]$. 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 3, 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 $j \leq s$. 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. □

Proof of Theorem 5. 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 maxima. 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, 2011, Theorem 5.7). □

References

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

Glynn, P. W., S. Juneja. 2004. A large deviations perspective on ordinal optimization. R. G. Ingalls, M. D. Rossetti, J. S. Smith, B. A. Peters, eds., *Proc. of the 2004 Winter Simulation Conf.*. IEEE Press, Piscataway, NJ, 577–585.

Pasupathy, R., S. Kim. 2011. The stochastic root-finding problem: overview, solutions, and open questions. *ACM Trans. on Model. and Comput. Simulation* **21** 1–23.