LARGE-DEVIATION SAMPLING LAWS FOR CONSTRAINED SIMULATION OPTIMIZATION ON FINITE SETS

Susan R. Hunter
Raghu Pasupathy

Industrial and Systems Engineering
Virginia Tech
Blacksburg, VA 24061, USA

ABSTRACT

We consider the problem of selecting an optimal system from among a finite set of competing systems, based on a "stochastic" objective function and subject to a single "stochastic" constraint. By strategically dividing the competing systems, we derive a large deviations sampling framework that asymptotically minimizes the probability of false selection. We provide an illustrative example where a closed-form sampling law is obtained after relaxation.

1 INTRODUCTION

The simulation-optimization (SO) problem is a nonlinear optimization problem where the objective and constraint function(s) involved are observable only through a stochastic simulation. Numerous SO variations based on the structure of the domain of the objective function, the nature of the constraint set, and the type of requested solution have recently received great attention within the simulation literature. (See Henderson and Nelson (2006) for a comprehensive reference list and examples to motivate SO's wide applicability.) Our interest is limited to a simple subclass of these SO variations — SO problems where the domain of the objective function is a "small" finite set, and a single observable constraint is imposed. Currently available solutions within the subclass of SO problems on "small" finite spaces seem to organize themselves based on the presence (or lack) of objective/constraint function(s) and the distributional assumptions on the simulation output. Table 1 places the current work within this larger problem context.

Table 1: Example papers in the area of simulation optimization on finite sets can be categorized by the nature of the distributional assumption and the presence of objective function or constraints. The problem considered in this paper falls in the bottom right-hand cell, for which there is currently no example paper.

Distributional Assumption	Optimization: only objective(s)	Feasibility: only constraint(s)	Constrained Optimization: objective(s) & constraint(s)
Normal	Kim and Nelson (2006)	Batur and Kim (2010)	Andradóttir and Kim (2010)
General	Glynn and Juneja (2004)	Szechtman and Yücesan (2008)	?

The rows of Table 1 divide the research in this area into papers by the required distributional assumption. For instance, the first row of the table includes papers that make a normality assumption on the simulation output. These papers also happen to be, almost invariably, in contexts where a solution to the SO problem is required in finite time and with a stipulated probabilistic guarantee. In contrast, the second row represents a much more recent line of work that relaxes the distributional assumption on the simulation output, but provides only asymptotic guarantees on attainment of the correct solution. Since Glynn and Juneja (2004) show that a failed normality assumption may provide misleading allocations, this more general line of research can be applied in cases of distributional uncertainty, albeit with sacrifices on finite-time guarantees.

The work we present in this paper resides in the bottom right-hand cell of Table 1, and to our knowledge is the first to attempt this SO variation. It lies alongside a relatively new line of research on constrained ranking and selection (R&S), most notably explored by Andradóttir and Kim (2010), and follows directly as an extension of previous large deviations work in ordinal optimization by Glynn and Juneja (2004) and feasibility determination by Szechtman and Yücesan (2008). Specifically, our work adds a constraint to the work of Glynn and Juneja (2004),

to begin literature in the bottom right-hand cell of Table 1. Like Andradóttir and Kim (2010), we are concerned with finding the best feasible system in the presence of a single constraint. Also, like Glynn and Juneja (2004) and Szechtman and Yücesan (2008), we derive conditions for the limiting optimal allocation in the context of general light-tailed distributions.

1.1 Problem Statement

We consider a finite set $i = 1, \ldots, r$ of systems, each with an unknown objective value $h_i \in \mathbb{R}$ and unknown constraint value $g_i \in \mathbb{R}$. Given a constant $\gamma \in \mathbb{R}$, we wish to select the system with the lowest objective value h_i, subject to a single constraint $g_i \geq \gamma$, assuming at least two systems are feasible. Formally, we consider

$$\arg \min_{i=1,\ldots,r} \quad h_i$$
$$\text{s.t.} \quad g_i \geq \gamma,$$

where a unique solution exists, h_i and g_i are expectations, and estimates of h_i and g_i may observed through simulation as sample means. In this context, we derive a framework for asymptotically optimal sample allocation which minimizes the probability of false selection, that is, the probability of returning to the user a system other than the best feasible system.

1.2 Organization

In Section 2 we discuss the notation and assumptions for the paper. In Section 3 we derive an expression for the rate function of the probability of false selection, and in Section 4 we present both a general sampling framework, as well as a sampling framework for a user-specified sampling allocation to the best system. Section 5 contains concluding remarks. For brevity, we present results without proofs.

2 PRELIMINARIES

For notational convenience, we refer to the feasible system with the lowest objective value as system 1. We introduce the following notation to partition the set of r systems into four mutually exclusive and collectively exhaustive subsets:

$\{1\}$:   the set consisting of the unique best feasible system, $\{1\} = \{i : g_i > \gamma,\ h_i < h_j\ \forall j \leq r\}$
Γ :   the set of suboptimal feasible systems, that is, $\Gamma = \{i : g_i \geq \gamma,\ i \neq 1\}$,
$\mathcal{S}_b$:   the set of infeasible systems that have *better* (lower) objective values than system 1, that is,
     $\mathcal{S}_b = \{i : g_i < \gamma,\ h_1 \geq h_i\}$, and,
$\mathcal{S}_w$:   the set of infeasible systems that have *worse* (higher) objective values than system 1, that is,
     $\mathcal{S}_w = \{i : g_i < \gamma,\ h_1 < h_i\}$.

To be falsely selected as the best feasible system, systems in Γ must pretend to be optimal, systems in $\mathcal{S}_b$ must pretend to be feasible, and systems in $\mathcal{S}_w$ must pretend to be both optimal and feasible. The careful construction of these sets ensures systems are distinguishable from the quantity on which their potential false evaluation as the "best" system depends. Therefore no system will require all of the simulation budget. A similar assumption is made in Szechtman and Yücesan (2008).

To estimate the unknown quantities h_i and g_i, we assume we may obtain replicates of the output random variables (H_i, G_i) from each system, where

Assumption 1. *(1) for a particular system, H_i and G_i are independent; and (2) the systems are independent of each other, that is, the output random variables (H_i, G_i) are mutually independent for all $i \leq r$.*

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

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

be the log-moment generating functions of $\hat{H}_i$ and $\hat{G}_i$, respectively. Let the effective domain of a function $f(x)$ be denoted by $\mathcal{D}_f = \{x : f(x) < \infty\}$, and its interior by $\mathcal{D}_f^{\circ}$. We assume the following.

Assumption 2. *For each system,*

(1)   the limits

$$\Lambda_i^H(\theta) = \lim_{n\to\infty} \frac{1}{\alpha_i n} \Lambda_i^{\hat{H}}(\alpha_i n \theta) \;\; and \;\; \Lambda_i^G(\theta) = \lim_{n\to\infty} \frac{1}{\alpha_i n} \Lambda_i^{\hat{G}}(\alpha_i n \theta),$$

exist as exended real numbers for all θ;

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

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

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

Assumption 2 implies that $\hat{H}_i \to h_i$ with probability one (w.p.1) and $\hat{G}_i \to g_i$ w.p.1 (see Remark 3.2.1, Bucklew (2003)). By the Gärtner-Ellis theorem, for each system, $\hat{H}_i$ and $\hat{G}_i$ satisfy the large deviations principle with good rate functions

$$I_i(x) = \sup_{\theta\in\mathbb{R}}\{\theta x - \Lambda_i^H(\theta)\} \;\; and \;\; J_i(y) = \sup_{\theta\in\mathbb{R}}\{\theta y - \Lambda_i^G(\theta)\},$$

respectively. We note that Assumption 2(3) is stronger than we typically need for the Gärtner-Ellis theorem. However we require $\Lambda_i^H(\theta)$ and $\Lambda_i^G(\theta)$ to be strictly convex and C^∞ on $\mathcal{D}_{\Lambda_i^H}^\circ$ and $\mathcal{D}_{\Lambda_i^G}^\circ$, respectively, so that $I_i(x)$ and $J_i(y)$ are strictly convex and C^∞ for $x \in \mathcal{F}_i^{H\circ} = \text{int}\{\Lambda_i^{H'}(\theta) : \theta \in \mathcal{D}_{\Lambda_i^H}^\circ\}$ and $y \in \mathcal{F}_i^{G\circ} = \text{int}\{\Lambda_i^{G'}(\theta) : \theta \in \mathcal{D}_{\Lambda_i^G}^\circ\}$, respectively. Without loss of generality, let $h_1 < h_2 \leq \ldots \leq h_r$. We further assume

Assumption 3. *(1) the interval $[h_1,h_r] \subset \cap_{i=1}^r \mathcal{F}_i^{H\circ}$, and (2) $\gamma \in \cap_{i=1}^r \mathcal{F}_i^{G\circ}$.*

Assumption 3(1) ensures that $\hat{H}_i$ may take any value in the interval $[h_1,h_r]$ and that $P(\hat{H}_1 \geq \hat{H}_i) > 0$ for $2 \leq i \leq r$. Assumption 3(2) ensures there is a nonzero probability that a feasible system will be estimated-infeasible and a nonzero probability that an infeasible system will be estimated-feasible, that is, $P(\hat{G}_i \geq \gamma) > 0$ for $i \in \mathcal{S}_b \cup \mathcal{S}_w$ and $P(\hat{G}_i < \gamma) > 0$ for $i \in \{1\} \cup \Gamma$. We note that Assumptions 1(2), 2, and 3(1) are similar to the assumptions of Glynn and Juneja (2006).

3   RATE FUNCTION OF PROBABILITY OF FALSE SELECTION

The probability of false selection (FS), that is, the probability that we return to the user a system other than system 1, is the probability that the optimal system is incorrectly categorized as infeasible, or the optimal system is categorized as feasible, but another estimated-feasible system has a lower estimated objective value. Let $\hat{\Gamma} = \{i : \hat{G}_i \geq \gamma, i \neq 1\}$ be the estimated set of suboptimal feasible systems. Then the probability of false selection is

$$\begin{aligned} P\{FS\} &= P\{(\hat{G}_1 < \gamma) \cup ((\hat{G}_1 \geq \gamma) \cap (\hat{H}_1 \geq \min_{i\in\hat{\Gamma}} \hat{H}_i))\} \\ &= P\{\hat{G}_1 < \gamma\} + P\{\hat{G}_1 \geq \gamma\}P\{\hat{H}_1 \geq \min_{i\in\hat{\Gamma}} \hat{H}_i\} \\ &= P\{FS_1\} + P\{FS_2\}, \end{aligned} \tag{1}$$

where FS_1 denotes the event that the optimal system is categorized as infeasible, and FS_2 denotes the event that the optimal system is correctly classified as feasible, but is "beaten" in objective function value by another estimated-feasible system. Consistent with Dembo and Zeitouni (1998), we interpret the minimum over the null set as infinity. We wish to obtain an expression for the rate at which the probability of false selection tends to zero with increasing computing budget n. We will first derive expressions for the rate functions of $P\{FS_1\}$ and $P\{FS_2\}$, and combine these results to find the rate function of $P\{FS\}$.

Under Assumptions 1, 2, and 3 (2), the rate function for $P\{FS_1\}$ is straightforward. By application of the Gärtner-Ellis theorem, we find the following.

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

$$\lim_{n\to\infty} \frac{1}{n}\log P\{FS_1\} = -\alpha_1 J_1(\gamma).$$

Therefore the rate at which the probability that system 1 is incorrectly classified as infeasible goes to zero is governed by the constant $J_1(\gamma)$. We note that a similar result holds for any $i \in \mathcal{S}_b \cup \mathcal{S}_w$, such that $\lim_{n\to\infty} \frac{1}{n}\log P\{i \in \hat{\Gamma}\} = -\alpha_i J_i(\gamma)$.

Considering the rate function for $P\{FS_2\}$, we note that since $\hat{G}_1 \to g_1$ w.p.1 and $g_1 \leq \gamma$, then $P\{\hat{G}_1 \geq \gamma\} \to 1$. Therefore the rate function of $P\{FS_2\}$ will be governed by the rate function of $P\{\hat{H}_1 \geq \min_{i \in \hat{\Gamma}} \hat{H}_i\}$ in (1). Since the estimated set of feasible systems $\hat{\Gamma}$ may contain worse feasible systems ($i \in \Gamma$), better infeasible systems ($i \in \mathcal{S}_b$), and worse infeasible systems ($i \in \mathcal{S}_w$), we strategically consider the rate functions for the probability that system 1 is beaten by a system in $\hat{\Gamma} \cap \Gamma$, $\hat{\Gamma} \cap \mathcal{S}_b$, or $\hat{\Gamma} \cap \mathcal{S}_w$ separately. By rewriting the probability that system 1 is beaten by an estimated-feasible system as a sum of probabilities over the partitioned sets, we are able to employ the "principle of the largest term," which states that if $a_n^i, i = 1, \ldots, k$ are sequences in $\mathbb{R}^+$ such that $n^{-1} \log a_n^i \to a^i$ for all $i \leq k$, then $n^{-1} \log(\sum_{i=1}^k a_n^i) \to \max_i a^i$ (see, e.g., Lemma 2.1, Ganesh, O'Connell, and Wischik (2004)). Specifically, this principle states that if each term a_n^i represents a probability tending to zero as n tends to infinity, then the slowest converging term dominates the limit of the scaled sum. Therefore the rate function of $P\{FS_2\}$ will be dominated by the slowest converging probability that system 1 will be beaten by an estimated-feasible system in Γ, $\mathcal{S}_b$, or $\mathcal{S}_w$. Lemma 2 states this result formally, assuming for now that the required limits exist.

Lemma 2. *The rate function for $P\{FS_2\}$ is given by the minimum rate function of the probability that (1) system 1 is beaten by a worse, feasible, and estimated-feasible system; (2) system 1 is beaten by a better, infeasible, and estimated-feasible system; or (3) system 1 is beaten by a worse, infeasible, and estimated-feasible system. That is,*

$$\lim_{n \to \infty} \frac{1}{n} \log P\{FS_2\} =$$
$$-\min \left(-\lim_{n \to \infty} \frac{1}{n} \log P\{\hat{H}_1 \geq \min_{i \in \hat{\Gamma} \cap \Gamma} \hat{H}_i\}, -\lim_{n \to \infty} \frac{1}{n} \log P\{\hat{H}_1 \geq \min_{i \in \hat{\Gamma} \cap \mathcal{S}_b} \hat{H}_i\}, -\lim_{n \to \infty} \frac{1}{n} \log P\{\hat{H}_1 \geq \min_{i \in \hat{\Gamma} \cap \mathcal{S}_w} \hat{H}_i\} \right). \quad (2)$$

We now individually consider each of the terms on the right hand side of equation (2), and show that these limits exist. First, we consider the rate function of the probability that system 1 is beaten by a worse feasible system. Since $\hat{G}_i \to g_i$ w.p.1 for all i, then in the limit, $\hat{\Gamma} = \Gamma$. Therefore in the limit, we expect the set of feasible systems will operate as in the case of unconstrained ordinal optimization, which is considered in Glynn and Juneja (2004). The following lemma shows that this is, in fact, the case.

Lemma 3. *The rate function for the probability that system 1 is estimated-feasible and beaten by a worse, feasible, and estimated-feasible system is*

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

We note that, intuitively, we may add the rate functions $I_1(x)$ and $I_i(x)$ because of the assumed independence between systems (Assumption 1(2)). Loosely speaking, in the prior lemma, the rate function of the probability that system 1 is beaten by a worse feasible system is dominated by the suboptimal feasible system which is "best" at pretending to be the best feasible system. In this case, the rate at which feasible systems become feasible does not affect the rate function.

We now consider the rate function of the probability that system 1 is beaten by a better infeasible system. Since the only hurdle to a better, infeasible system being declared optimal is feasibility, we find that the rate function is dominated by the better infeasible system which is "best" at pretending to be feasible. Lemma 4 states this result rigorously.

Lemma 4. *The rate function for the probability that system 1 is estimated-feasible and beaten by a better, infeasible, and estimated-feasible system is*

$$\lim_{n \to \infty} \frac{1}{n} \log P\{\hat{H}_1 \geq \min_{i \in \hat{\Gamma} \cap \mathcal{S}_b} \hat{H}_i\} = -\min_{i \in \mathcal{S}_b} \alpha_i J_i(\gamma).$$

Therefore the rate at which better infeasible systems are declared "better" does affect the rate function. Informally, in the limit, we are only concerned with the better infeasible system that jumps into the feasible set most easily.

Finally, we consider the rate function for the probablility that system 1 is beaten by a worse, infeasible system. Systems which are infeasible and worse must pretend to be both feasible and optimal. Therefore the rate function includes both optimality and feasibility terms, where the addition between these terms is a result of the independence between the objective function and constraints (Assumption 1(1)).

Lemma 5. *The rate function for the probability that system 1 is estimated-feasible and beaten by a worse, infeasible, and estimated-feasible system is*

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

As before, we find that the rate function is dominated by the worse infeasible system which is "best" at pretending to be feasible and optimal.

Combining Lemmas 2 through 5 and applying the principle of the largest term, we arrive at the following theorem.

Theorem 1. *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*

$$\lim_{n\to\infty}\frac{1}{n}\log P\{FS\} =$$

$$-\min(\ \overbrace{\alpha_1 J_1(\gamma)}^{\text{declared infeasible}},\ \underbrace{\min_{i\in\Gamma}(\inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x)))}_{\text{beaten by worse feasible system}},\ \overbrace{\underbrace{\min_{i\in\mathcal{S}_b}\alpha_i J_i(\gamma)}_{\text{beaten by better infeasible system}},\ \underbrace{\min_{i\in\mathcal{S}_w}(\inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x)) + \alpha_i J_i(\gamma))}_{\text{beaten by worse infeasible system}}}^{\text{declared feasible}}).$$

4 OPTIMAL ALLOCATION STRATEGY

We now wish to derive an optimal allocation strategy that asymptotically minimizes the probability of false selection. We first present a general allocation framework, followed by a relaxed framework which allows the user to specify the allocation to system 1.

4.1 General Optimal Allocation Framework

From Theorem 1, an asympotically optimal allocation strategy will result from maximizing the rate at which $P\{FS\}$ tends to zero with increasing n. Thus we wish to allocate the α_i's to solve the following optimization problem:

$$\max\quad \min\left(\alpha_1 J_1(\gamma),\ \min_{i\in\Gamma}\left(\inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x))\right),\ \min_{i\in\mathcal{S}_b}\alpha_i J_i(\gamma),\ \min_{i\in\mathcal{S}_w}\left(\inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x)) + \alpha_i J_i(\gamma)\right)\right)$$

$$s.t.\quad \sum_{i=1}^r \alpha_i = 1$$

$$\alpha_i \geq 0.$$

By Glynn and Juneja (2006), $\inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x))$ is a concave, strictly increasing, C^∞ function of α_1 and α_i. Let $x(\alpha_1, \alpha_i) = \arg\inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x))$. Then for $\alpha_1 > 0$ and $\alpha_i > 0$, $x(\alpha_1, \alpha_i)$ is C^∞ since $I(\cdot)$ is (Glynn and Juneja 2006). Likewise, the linear functions $\alpha_1 J_1(\gamma)$ and $\alpha_i J_i(\gamma)$ and the sum $\inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x)) + \alpha_i J_i(\gamma)$ are also concave, strictly increasing, C^∞ functions of α_1 and α_i. Hence the minimum of concave, strictly increasing functions is also concave and strictly increasing, and we have a concave optimization problem. Equivalently, we may rewrite the concave program as,

$$\max\quad z\quad s.t.$$

$$\alpha_i J_i(\gamma) \geq z,\ i \in \{1\}\cup\mathcal{S}_b$$

$$\alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i)) \geq z,\ i \in \Gamma$$

$$\alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i)) + \alpha_i J_i(\gamma) \geq z,\ i \in \mathcal{S}_w$$

$$\sum_{i=1}^r \alpha_i = 1$$

$$\alpha_i \geq 0.$$

Since our problem is concave with differentiable objective function and constraints, the Karush-Kuhn Tucker conditions are sufficient for global optimality. Therefore an allocation $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_r)$ that satisfies the conditions in the following theorem will be asymptotically optimal.

Theorem 2. *If* $\boldsymbol{\alpha}^* > 0, \sum_{i=1}^{r} \alpha_i^* = 1$ *minimizes the asymptotic probability of false selection, then it satisfies*

$$\sum_{i \in \Gamma} \frac{I_1(x(\alpha_1^*, \alpha_i^*))}{I_i(x(\alpha_1^*, \alpha_i^*))} + \sum_{i \in \mathcal{S}_w} \frac{I_1(x(\alpha_1^*, \alpha_i^*))}{I_i(x(\alpha_1^*, \alpha_i^*)) + J_i(\gamma)} = 1; \tag{3}$$

$$\alpha_i^* J_i(\gamma) = \alpha_1^* I_1(x(\alpha_1^*, \alpha_j^*)) + \alpha_j^* I_j(x(\alpha_1^*, \alpha_j^*)) = \alpha_1^* I_1(x(\alpha_1^*, \alpha_k^*)) + \alpha_k^* [I_k(x(\alpha_1^*, \alpha_k^*)) + J_k(\gamma)],$$
$$\text{for all } i \in \mathcal{S}_b, j \in \Gamma, k \in \mathcal{S}_w. \tag{4}$$

Equation (4) tells us that we must allocate sampling within and between the sets $\Gamma, \mathcal{S}_b$, and $\mathcal{S}_w$ to equate the rate functions of the systems in these sets. In other words, we use $\boldsymbol{\alpha}$ as a dial that allows us to equate the rates at which the probabilities of these systems beating system 1 go to zero. Equation (3) tells us how to allocate sampling between system 1 and all of the other systems. We note that in this equation, we sum over $i \in \Gamma$ and $i \in \mathcal{S}_w$, but not $i \in \mathcal{S}_b$. These sets, Γ and $\mathcal{S}_w$, are the sets in which α_1 appears in the rate function. Systems in $\Gamma \cup \mathcal{S}_w$ must beat system 1 to be falsely selected as optimal, while systems in $\mathcal{S}_b$ must only pretend to be feasible when the estimated value of g_i is compared to a constant. We note that when considering only systems in $\{1\} \cup \Gamma$, this result reduces to the result presented in Glynn and Juneja (2004).

4.2 Optimal Allocation for a Specified α_1

Since the optimal allocation presented in Theorem 2 may be difficult to solve, we consider a relaxation in which we fix α_1 and optimize across the remaining α_i's. Let $\boldsymbol{\alpha}'^* = (\alpha_1, \alpha_2^*, \ldots, \alpha_r^*)$. Then the following corollary follows from Theorem 2.

Corollary 1. *If* $\boldsymbol{\alpha}'^* > 0, \sum_{i=2}^{r} \alpha_i^* = 1 - \alpha_1$ *minimizes the asymptotic probability of false selection, then*

$$\alpha_i^* J_i(\gamma) = \alpha_1 I_1(x(\alpha_1, \alpha_j^*)) + \alpha_j^* I_j(x(\alpha_1, \alpha_j^*)) = \alpha_1 I_1(x(\alpha_1, \alpha_k^*)) + \alpha_k^* [I_k(x(\alpha_1, \alpha_k^*)) + J_k(\gamma)],$$
$$\text{for all } i \in \mathcal{S}_b, j \in \Gamma, k \in \mathcal{S}_w. \tag{5}$$

We now consider the case in which the random variables corresponding to both the objective and constraint have normal distributions.

Example 1. *Normal.*

Let $H_i \sim N(h_i, \sigma_{h_i}^2)$ and $G_i \sim N(g_i, \sigma_{g_i}^2)$ for all $i \leq r$. Then

$$I_i(x) = \frac{(x - h_i)^2}{2\sigma_{h_i}^2} \quad \text{and} \quad J_i(\gamma) = \frac{(\gamma - g_i)^2}{2\sigma_{g_i}^2}.$$

Differentiating to find the value of x at which $\inf_x(\alpha_1 I_1(x) + \alpha_i I_i(x))$ is achieved yields

$$x(\alpha_1, \alpha_i) = \frac{h_1}{1 + \left(\dfrac{\alpha_i \sigma_{h_1}^2}{\alpha_1 \sigma_{h_i}^2}\right)} + \frac{h_i}{1 + \left(\dfrac{\alpha_1 \sigma_{h_i}^2}{\alpha_i \sigma_{h_1}^2}\right)}$$

Then the rate functions are,

$$\alpha_i J_i(\gamma) = \frac{\alpha_i(\gamma - g_i)^2}{2\sigma_{g_i}^2}, \, i \in \{1\} \cup \mathcal{S}_b, \tag{6}$$

$$\alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i I_i(x(\alpha_1, \alpha_i)) = \frac{\alpha_1 \alpha_i (h_1 - h_i)^2}{2(\alpha_i \sigma_{h_1}^2 + \alpha_1 \sigma_{h_i}^2)}, \, i \in \Gamma, \tag{7}$$

$$\alpha_1 I_1(x(\alpha_1, \alpha_i)) + \alpha_i [I_i(x(\alpha_1, \alpha_i)) + J_i(\gamma)] = \frac{\alpha_1 \alpha_i (h_1 - h_i)^2}{2(\alpha_i \sigma_{h_1}^2 + \alpha_1 \sigma_{h_i}^2)} + \frac{\alpha_i(\gamma - g_i)^2}{2\sigma_{g_i}^2}, \, i \in \mathcal{S}_w. \tag{8}$$

Taking partial derivatives with respect to α_i, we find

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

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

The general solution for the optimal allocation in the normal case does not have a closed-form expression. In the unconstrained normal case, Chen et al. (2000) and Glynn and Juneja (2004) simplify the optimal allocation by allowing $\alpha_1^* \gg \alpha_i^*$. However it is not clear what this approximation means in terms of problem parameters. For example, to take the limit as $\alpha_1^* \to 1$ and maintain the KKT conditions for optimality, problem parameters such as σ_{h_1} must change to maintain the optimality and feasibility of the new allocation. It may be shown that the relative rates of change of these parameters force (3) not to hold. We circumvent this problem by removing the optimal allocation requirement between system 1 and all other systems, that is, fixing α_1 and optimizing with respect to $\alpha_i, i \in \{2,\ldots,r\}$. We thus fall within the purview of Corollary 1. Using (5), we then have,

$$
\begin{aligned}
\frac{\alpha_i^*}{\alpha_j^*} &= \frac{\alpha_1 I_1(x(\alpha_1,\alpha_j^*))\mathbb{I}_{(h_1<h_j)} - \alpha_1 I_1(x(\alpha_1,\alpha_i^*))\mathbb{I}_{(h_1<h_i)} + \alpha_j^* I_j(x(\alpha_1,\alpha_j^*))\mathbb{I}_{(h_1<h_j)} + \alpha_j^* J_j(\gamma)\mathbb{I}_{(j \notin \Gamma)}}{\alpha_j^* I_i(x(\alpha_1,\alpha_i^*))\mathbb{I}_{(h_1<h_i)} + \alpha_j^* J_i(\gamma)\mathbb{I}_{(i \notin \Gamma)}} \\[2mm]
&= \frac{\dfrac{\alpha_1 \alpha_j^* \sigma_{h_1}^2 (h_1-h_j)^2}{(\alpha_j^* \sigma_{h_1}^2 + \alpha_1 \sigma_{h_j}^2)^2}\mathbb{I}_{(h_1<h_j)} + \dfrac{\alpha_1^2 \sigma_{h_j}^2 (h_1-h_j)^2}{(\alpha_j^* \sigma_{h_1}^2 + \alpha_1 \sigma_{h_j}^2)^2}\mathbb{I}_{(h_1<h_j)} + \dfrac{(\gamma-g_j)^2}{\sigma_{g_j}^2}\mathbb{I}_{(j \notin \Gamma)}}{\dfrac{\alpha_1 \alpha_i^* \sigma_{h_1}^2 (h_1-h_i)^2}{(\alpha_i^* \sigma_{h_1}^2 + \alpha_1 \sigma_{h_i}^2)^2}\mathbb{I}_{(h_1<h_i)} + \dfrac{\alpha_1^2 \sigma_{h_i}^2 (h_1-h_i)^2}{(\alpha_i^* \sigma_{h_1}^2 + \alpha_1 \sigma_{h_i}^2)^2}\mathbb{I}_{(h_1<h_i)} + \dfrac{(\gamma-g_i)^2}{\sigma_{g_i}^2}\mathbb{I}_{(i \notin \Gamma)}}.
\end{aligned}
\tag{9}
$$

Now let $\alpha_1 \to 1$ in (9). Since $1 - \alpha_1 = \sum_{i=2}^r \alpha_i^*$, then $\alpha_i^* \to 0$ for all $i \in \{2\ldots,r\}$. Then

$$\frac{\alpha_i^*}{\alpha_j^*} \approx \frac{\left(\dfrac{h_1-h_j}{\sigma_{h_j}}\right)^2 \mathbb{I}(h_1<h_j) + \left(\dfrac{\gamma-g_j}{\sigma_{g_j}}\right)^2 \mathbb{I}(j \notin \Gamma)}{\left(\dfrac{h_1-h_i}{\sigma_{h_i}}\right)^2 \mathbb{I}(h_1<h_i) + \left(\dfrac{\gamma-g_i}{\sigma_{g_i}}\right)^2 \mathbb{I}(i \notin \Gamma)},$$

where equality holds for all $i,j \in \mathcal{S}_b$. When considering only feasible systems, these results reduces to those presented by Glynn and Juneja (2004).

5   CONCLUDING REMARKS

The constrained SO problem on small 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 correct solution, and minimum sample size rules that probabilistically guarantee attainment of the correct solution remain largely unexplored. Following recent work by Glynn and Juneja (2004) for the unconstrained SO context and Szechtman and Yücesan (2008) for the context of detecting feasibility, we take the first steps toward answering these questions.

To identify the relationship between sampling and error-probability decay, we strategically divide the competing systems into four sets: best feasible, feasible and worse, infeasible and better, and infeasible and worse. Such strategic division facilitates expressing the rate function of the probability of false selection as the minimum of rate functions over these four sets. Finding the optimal sampling allocation then reduces to a concave maximization problem, akin to Glynn and Juneja (2004). Interestingly, at least for the specific context considered, identifying the optimal sampling allocation, i.e., solving the concave maximization problem, is not "harder" than in the unconstrained context.

We are currently pursuing various directions as part of ongoing research. For example, is there an agreeable relaxation of the concave maximization problem that allows the identification of a general closed-form allocation? In addition to serving as an approximate solution for implementation, such allocation may also be useful as a starting solution for an algorithm that yields a consistent estimator for the optimal solution of the concave maximization problem. Other unanswered questions of interest include removing the independence requirement between the objective function and constraint, and extending our optimization framework to more than one constraint.

REFERENCES

Andradóttir, S., and S.-H. Kim. 2010. Fully sequential procedures for comparing constrained systems via simulation. To appear in *Naval Research Logistics*.

Batur, D., and S.-H. Kim. 2010. Finding feasible systems in the presence of constraints on multiple performance measures. To appear in *ACM Transactions on Modeling and Computer Simulation*.

Bucklew, J. A. 2003. *Introduction to rare event simulation*. New York: Springer.

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

Dembo, A., and O. Zeitouni. 1998. *Large deviations techniques and applications*. 2nd ed. New York: Springer.

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

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

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

Henderson, S. G., and B. L. Nelson. (Eds.) 2006. *Simulation*. Handbooks in Operations Research and Management Science, Volume 13. Elsevier.

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

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

AUTHOR BIOGRAPHIES

SUSAN R. HUNTER is a Ph.D. candidate in the Industrial & Systems Engineering department at Virginia Tech. She holds Masters and Bachelors degrees in Statistics from North Carolina State University. Her research interests include simulation modeling and analysis. Her email address is <srhunter@vt.edu>.

RAGHU PASUPATHY is an assistant professor in the Industrial and Systems Engineering department at Virginia Tech. His research interests lie broadly in Monte Carlo methods with a specific focus on simulation optimization and stochastic root finding. He is a member of INFORMS, IIE, and ASA, and serves as an Associate Editor for ACM TOMACS and INFORMS Journal on Computing. His e-mail address is ⟨pasupath@vt.edu⟩, and his web page is ⟨https://filebox.vt.edu/users/pasupath/pasupath.htm⟩.