Stochastically Constrained Ranking and Selection via SCORE

RAGHU PASUPATHY, Virginia Tech and IBM Research
SUSAN R. HUNTER, Cornell University
NUGROHO A. PUJOWIDIANTO and LOO HAY LEE, National University of Singapore
CHUN-HUNG CHEN, George Mason University

Consider the context of *constrained* simulation optimization (SO), that is, optimization problems where the objective and constraint functions are known through dependent Monte Carlo estimators. For solving such problems on *large finite spaces*, we provide an easily implemented sampling framework called SCORE (Sampling Criteria for Optimization using Rate Estimators) that approximates the optimal simulation budget allocation. We develop a general theory, but like much of the existing literature on ranking and selection, our focus is on SO problems where the distribution of the simulation observations is Gaussian. We first characterize the nature of the optimal simulation budget as a bilevel optimization problem. We then show that under a certain asymptotic limit, the solution to the bilevel optimization problem becomes surprisingly tractable and is expressed through a single intuitive measure, the *score*. We provide an iterative SO algorithm that repeatedly estimates the score and determines how the available simulation budget should be expended across contending systems. Numerical experience with the algorithm resulting from the proposed sampling approximation is very encouraging — in numerous examples of constrained SO problems having 1,000 to 10,000 systems, the optimal allocation is identified to negligible error within a few seconds to one minute on a typical laptop computer. Corresponding times to solve the full bilevel optimization problem range from tens of minutes to several hours.

Additional Key Words and Phrases: constrained simulation optimization; ranking and selection

ACM Reference Format:
Raghu Pasupathy, Susan R. Hunter, Nugroho A. Pujowidianto, Loo Hay Lee, and Chun-Hung Chen. 2014. Stochastically constrained ranking and selection via SCORE. *ACM Trans. Model. Comput. Simul.* 0, 0, Article 0 (2014), 25 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Constrained Simulation Optimization (SO) is a class of nonlinear optimization problems where the objective and constraint functions can be expressed implicitly, e.g., using a stochastic simulation. This implicit definition of the functions is extremely ver-

Raghu Pasupathy was supported in part by Office of Naval Research contracts N000141110419, N00014-11-1-0065 and National Science Foundation (NSF) grant CMMI 478709. He thanks IBM Research, Yorktown Heights, NY where he spent his sabbatical year 2012–2013. Susan R. Hunter was supported in part by NSF Grants CMMI-0800688 and CMMI 1200315. Loo Hay Lee was supported by The Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning. Chun-Hung Chen was supported in part by NSF under Award CMMI-1233376, Department of Energy under Award DE-SC0002223, NIH under Grant 1R21DK088368-01, and National Science Council of Taiwan under Award NSC-100-2218-E-002-027-MY3.

Author's addresses: Raghu Pasupathy, Department of Statistics, Purdue University; S. R. Hunter, School of Industrial Engineering, Purdue University; N. A. Pujowidianto and Loo Hay Lee, Department of Industrial and Systems Engineering, National University of Singapore; Chun-Hung Chen, Systems Engineering and Operations Research, George Mason University.

DOI:http://dx.doi.org/10.1145/0000000.0000000

satile, in contrast to more traditional optimization settings that stipulate an explicit expression of the objective and constraint functions. The versatility of the SO formulation has resulted in its widespread adoption. “Real world” problems in a variety of areas such as transportation systems, financial engineering, logistics, and healthcare now routinely employ SO formulations as a framework to solve complex optimization problems. For overviews and specific examples, see Spall [2003], Fu et al. [2005], Barton and Meckesheimer [2006], April et al. [2001], Andradóttir [2006], Pasupathy and Henderson [2006; 2011].

Analogous to deterministic optimization problems, SO problems are broadly categorized by the nature of the feasible region and the type of solution required. For instance, they are generally considered either categorical, integer-ordered, or continuous, depending on the nature of the feasible region, with problems falling in more than one category called mixed SO problems. Furthermore, SO problems in each of the integer-ordered and continuous categories can either be global or local, depending on the nature of the solution required. For examples in each of these categories, visit the library of SO problems at www.simopt.org [Pasupathy and Henderson 2006; 2011].

In this paper, we consider stochastically constrained SO problems on categorical spaces. This SO variation involves identifying the best system from a finite population of systems, as measured by an estimable objective function, from among those systems that are feasible, as measured by a set of estimable constraint functions. Our particular interest is solving large-scale problems having many thousands of competing systems, possibly using recursive algorithms that are easily implementable and provably near-optimal from the standpoint of computing effort. It is worth noting that since the SO variation we consider here includes stochastic constraints, it subsumes the unconstrained version of the categorical SO problem, broadly known as the ranking and selection (R&S) formulation [Kim and Nelson 2006]. Unlike R&S, which has been heavily studied [Kim and Nelson 2006; Branke et al. 2007], research on the constrained version is still in its infancy. Attempts at solution have been relatively few and very recent; entry points to work in this area include Andradóttir and Kim [2010], Hunter and Pasupathy [2013], and Lee et al. [2012].

1.1. Overview of Questions Answered

To provide a better sense of the specific questions we answer, consider the following general setting for identifying the best feasible system from among a finite set of competing systems. Suppose simulation runs are allocated across the available systems according to a budgeting scheme. After expending a certain amount of the simulation budget, the system with the smallest observed objective function estimate among those systems estimated to be feasible is chosen as the best system. The estimated best system may or may not coincide with the true best system, thereby giving rise to the notion of a false selection event, which is the event that the estimated best system is not the true best system. The probability of false selection ($P\{\text{FS}\}$) is the probability of observing a false selection event.

Our questions in this paper relate to the behavior of the $P\{\text{FS}\}$ as a function of the simulation budget and its allocation across systems, with an emphasis on settings where the design space is large. Further, we make no independence assumptions between the objective and constraint estimators for a system. Specifically, we ask:

Q.1 What is the optimal simulation budget allocation across designs, that is, what is the nature of the budget allocation that maximizes the rate of decay of $P\{\text{FS}\}$ to zero?

Q.2 Can a rigorously obtained but tractable approximation of the optimal allocation be derived, for use in settings where the number of systems is large?

Q.3 Can the answer to Q.2 be used to construct an easily implementable algorithmic scheme to solve large-scale stochastically constrained finite SO problems driven by Gaussian simulation observations?

Our answer to question Q.1 appears in §3 and extends work by Hunter [2011] and Hunter and Pasupathy [2013]. We answer question Q.2 in §4, where we demonstrate that the optimal allocation in the large-scale setting reduces to a form that is remarkably simple in structure and intuition. Specifically, we show that as the number of systems becomes large, the optimal simulation budget allocated to any suboptimal system (henceforth defined as any system other than the optimal system) is inversely proportional to a suboptimality-infeasibility measure that we call the *score*. Not surprisingly, the score for a suboptimal system depends only on the random variables inherent to that system and the optimal system. Furthermore, the score has an expression that seems easily estimable when the distributions driving the observations from each system are known or assumed. For example, when the observations corresponding to the constraint and objective functions from each system are independently normal, the score for a system is the sum of its squared standardized optimality gap and squared standardized infeasibility gaps across violated constraints, where standardization implies measuring the gap in standard deviation units. More generally, calculating the score amounts to minimizing a strictly convex function. See Table I for score expressions in a few other settings.

From the implementation standpoint in Q.3, when solving constrained SO problems on large finite spaces, our insight from answering Q.2 points to a solution scheme with three repeating steps akin to the popular OCBA scheme [Chen et al. 2000]: estimate the score, update the optimal simulation allocation across systems to be in inverse proportion to the estimated scores, and then select designs on which to execute the simulation according to the updated allocation scheme. This procedure results in a simple sequential algorithm that mimics the optimal budget allocation scheme, while reliably solving "large" problems with known or assumed distributions. For instance, as we demonstrate in §7, SO allocation problems with 10,000 systems and driven by Gaussian simulation observations are "solved" within a minute on a typical laptop computer. Without the use of the proposed approximation, the corresponding computing times are over six hours.

1.2. Competitors

Relatively little has been written on the topic of constrained SO on categorical spaces. Among the first papers on this topic are Andradóttir and Kim [2010] and Batur and Kim [2010], which deviate slightly from the question we consider here. Specifically, while we seek implementable allocation schemes that are provably near-optimal in the limit, Andradóttir and Kim [2010] and Batur and Kim [2010] seek finite-time simulation allocation schemes and termination strategies that identify the best system with probability exceeding a stipulated threshold. Given this finite-time objective, the authors propose schemes that can provide a rigorous finite-time probabilistic guarantee while striving to exceed the stipulated threshold by as little as possible.

The more recent paper by Lee et al. [2012] is closer to the current paper with respect to using an infinite-time objective in measuring algorithmic efficiency. However, unlike the current paper, the allocation proposed in Lee et al. [2012], called "OCBA-CO," is not optimal in any rigorous sense. (See Hunter and Pasupathy [2013] for examples where the allocation proposed by Lee et al. [2012] deviates substantially from the optimal allocation, leading to inferior decay rates of false selection.) Also a basis for the allocation proposed in Lee et al. [2012] is that the objective and constraint function estimators are uncorrelated, unlike the current paper.

Another work of particular relevance is Hunter and Pasupathy [2013], which characterizes the optimal sampling plan for contexts where the objective and constraint function estimators are mutually independent. What we propose here deviates from Hunter and Pasupathy [2013] in two ways, the first of which is crucial.

(i) The allocation procedure proposed in Hunter and Pasupathy [2013] involves solving a convex optimization problem at *each* step in a sequential procedure. This renders implementing Hunter and Pasupathy [2013] unsuitable when the number of systems in contention exceeds a few hundred (see Table II). The simulation allocation scheme proposed in this paper is aimed at solving problems with much larger numbers of systems, of the order of a few thousand, with negligible loss in simulation budgeting efficiency. As we shall demonstrate, we achieve this result as a limiting form of expressions generalized from Hunter and Pasupathy [2013]. Thus our proposed allocations are "closed-form" and *do not* require solving a convex optimization problem as in Hunter and Pasupathy [2013].

(ii) The theory in Hunter and Pasupathy [2013] assumes that the objective and constraint function estimators are independent, unlike the current paper. The theory in both the current paper and in Hunter and Pasupathy [2013] treat general distributions, although our implementation focus is on simulations driven by Gaussian observations.

2. PROBLEM SETTING AND FORMULATION

In this section, we outline a formal problem statement, notational conventions, and assumptions.

2.1. Problem Statement

The problem statement we consider is identical to the following problem statement from Hunter and Pasupathy [2013]. We study this problem statement in the case where the number of systems r is large.

> **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 $g_{ij} \leq \gamma_j$. That is, we consider
>
> $$\text{Problem } P: \text{ Find } \arg\min_{i=1,2,\ldots,r} \quad h_i$$
> $$\text{s.t.} \quad g_{ij} \leq \gamma_j, \text{ for all } j = 1, 2, \ldots, s,$$
>
> 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 $\alpha_i \geq 0$ for all $i = 1, 2, \ldots, r$. 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?

2.2. Notational Conventions

Where it is reasonable to do so, we generally use upper case letters for random variables, lower case letters for fixed quantities, bold type for vectors, and script letters for sets. For brevity, we write $i \le r$ and $j \le s$ to indicate $i = 1, 2 \ldots, r$ and $j = 1, 2 \ldots, s$. Throughout the paper, we let system 1 denote the best feasible system, that is, the system with the smallest value of h_i that satisfies the constraints $g_{ij} \le \gamma_j$ for all $j \le s$.

We also adopt the following notation throughout the paper. (i) For vectors $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_m)$, the notation $\boldsymbol{a} \le \boldsymbol{b}$ means $a_i \le b_i$ for all $i \le m$. (ii) $\mathrm{dist}(x, B) = \inf\{\|x - y\| : y \in B\}$ denotes the Euclidean distance between a point $x \in \mathbb{R}^q$ and a set $B \subset \mathbb{R}^q$. (iii) $\mathrm{diam}(B) = \sup\{\|x - y\| : x, y \in B\}$ denotes the diameter of the set $B \subset \mathbb{R}^q$. (iv) For a sequence of real numbers $\{a_n\}$, we say $a_n = o(1)$ if $\lim_{n\to\infty} a_n = 0$; and $a_n = O(1)$ if $\{a_n\}$ is bounded, i.e., $\exists c > 0$ with $|a_n| < c, \forall n$. We say that $a_n = \Theta(1)$ if $0 < \liminf a_n \le \limsup a_n < \infty$. (v) Let $C = \{1, 2, \ldots, k\}$ be a finite set and let $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ be a corresponding random vector having the $k \times k$ covariance matrix Σ. If $\tilde{C} \subset C$, then $\tilde{X}(\tilde{C})$ denotes the random vector comprising the elements of X with indices corresponding to $\tilde{C}$, and $\Sigma(\tilde{C})$ denotes the covariance matrix of $\tilde{X}(\tilde{C})$.

2.3. Assumptions

This paper follows from the general theory for constrained simulation optimization with correlation between the objective and constraint estimators outlined in Hunter [2011]. To this end, we require the same assumptions as those required in Hunter [2011]. First, to estimate the unknown quantities h_i and $\boldsymbol{g}_i = (g_{i1}, g_{i2}, \ldots, g_{is})$, we assume we may obtain replicates of the output random variables $(H_i, \boldsymbol{G}_i) = (H_i, G_{i1}, G_{i2}, \ldots, G_{is})$ from each system, where each system is simulated independently of the others.

ASSUMPTION 1. *The systems are simulated independently of each other, that is, the random vectors $(H_i, \boldsymbol{G}_i)$ are mutually independent for all $i \le r$.*

We also require the assumption that no system lies exactly on a constraint, and that no system has exactly the same objective function value as that of the best feasible system, system 1. This assumption is standard in literature that seeks an optimal sampling allocation since it ensures that two values may be distinguished with a finite simulation budget.

ASSUMPTION 2. *We assume $h_i \ne h_1 \; \forall\, i = 2, \ldots, r$ and $g_{ij} \ne \gamma_j \; \forall\, i \le r, j \le s$.*

Since this paper builds directly from the theory derived in Hunter [2011], the following two assumptions, standard in literature using large deviations theory, are required. Since our focus in this paper is to derive a broad sampling law for a large number of systems, we replicate these assumptions for completeness and refer the reader to Dembo and Zeitouni [1998] for further explanation. We first define the required notation.

For all systems $i \le r$ and constraints $j \le s$, denote the sample means after t samples as $\bar{H}_i(t) = \frac{1}{t}\sum_{k=1}^{t} H_{ik}$ and $\bar{G}_{ij}(t) = \frac{1}{t}\sum_{k=1}^{t} G_{ijk}$. Define $(\bar{H}_i(t), \bar{\boldsymbol{G}}_i(t)) := (\bar{H}_i(t), \bar{G}_{i1}(t), \ldots, \bar{G}_{is}(t))$. We use $(\hat{H}_i, \hat{\boldsymbol{G}}_i) \equiv (\bar{H}_i(\alpha_i t), \bar{\boldsymbol{G}}_i(\alpha_i t))$ as shorthand for the estimator of $(h_i, \boldsymbol{g}_i)$ when system i receives $\alpha_i > 0$ proportion of the total sampling budget t. For simplicity, we ignore that $\alpha_i t$ is not necessarily an integer. Let the cumulant generating functions of $\bar{H}_i(t)$, $\bar{G}_{ij}(t)$, and $(\bar{H}_i(t), \bar{\boldsymbol{G}}_i(t))$ be $\Lambda^{(t)}_{H_i}(\theta) = \log \mathrm{E}\big[e^{\theta \bar{H}_i(t)}\big]$, $\Lambda^{(t)}_{G_{ij}}(\theta) = \log \mathrm{E}\big[e^{\theta \bar{G}_{ij}(t)}\big]$, and $\Lambda^{(t)}_{(H_i, \boldsymbol{G}_i)}(\boldsymbol{\theta}) = \log \mathrm{E}\big[e^{\langle \boldsymbol{\theta}, (\bar{H}_i(t), \bar{\boldsymbol{G}}_i(t))\rangle}\big]$, respectively, where $\theta \in \mathbb{R}$,

$\boldsymbol{\theta} \in \mathbb{R}^{s+1}$, and $\langle \cdot, \cdot \rangle$ denotes the dot product. Let the effective domain of a function $f(\cdot)$ be denoted by $\mathcal{D}_f = \{x : f(x) < \infty\}$ and its interior by $\mathcal{D}_f^\circ$. Let $\nabla f(\boldsymbol{x})$ be the gradient of f with respect to $\boldsymbol{x}$, and $f'(x)$ the derivative of f with respect to x.

ASSUMPTION 3. *Let the following hold for each $i \leq r$ and $j \leq s$:*

(1) the limit $\Lambda_{(H_i,\boldsymbol{G}_i)}(\boldsymbol{\theta}) = \lim_{t\to\infty} \frac{1}{t}\Lambda^{(t)}_{(H_i,\boldsymbol{G}_i)}(t\boldsymbol{\theta})$ exists as an extended real number $\forall\ \boldsymbol{\theta} \in \mathbb{R}^{s+1}$, where we denote $\Lambda_{H_i}(\theta) = \lim_{t\to\infty} \frac{1}{t}\Lambda^{(t)}_{H_i}(t\theta)$ and $\Lambda_{G_{ij}}(\theta) = \lim_{t\to\infty} \frac{1}{t}\Lambda^{(t)}_{G_{ij}}(t\theta)\ \forall\ \theta \in \mathbb{R}$;
(2) the origin belongs to the interior of $\mathcal{D}_{\Lambda_{(H_i,\boldsymbol{G}_i)}}$, that is, $0 \in \mathcal{D}^\circ_{\Lambda_{(H_i,\boldsymbol{G}_i)}}$;
(3) $\Lambda_{(H_i,\boldsymbol{G}_i)}(\boldsymbol{\theta})$ is strictly convex and C^∞ on $\mathcal{D}^\circ_{\Lambda_{(H_i,\boldsymbol{G}_i)}}$;
(4) $\Lambda_{(H_i,\boldsymbol{G}_i)}(\boldsymbol{\theta})$ is steep, that is, for any sequence $\{\boldsymbol{\theta}(t)\} \in \mathcal{D}_{\Lambda_{(H_i,\boldsymbol{G}_i)}}$ converging to a boundary point of $\mathcal{D}_{\Lambda_{(H_i,\boldsymbol{G}_i)}}$, then $\lim_{t\to\infty} |\nabla\Lambda_{(H_i,\boldsymbol{G}_i)}(\boldsymbol{\theta}(t))| = \infty$.

Under Assumption 3, the large deviations principle (LDP) holds for the estimators $\bar{H}_i(t)$, $\bar{G}_{ij}(t)$, and $(\bar{H}_i(t), \bar{\boldsymbol{G}}_i(t))$ with good, strictly convex rate functions $I_i(x) = \sup_{\theta\in\mathbb{R}}\{\theta x - \Lambda_{H_i}(\theta)\}$, $J_{ij}(y) = \sup_{\theta\in\mathbb{R}}\{\theta y - \Lambda_{G_{ij}}(\theta)\}$, and $I_i(x, \boldsymbol{y}) = \sup_{\boldsymbol{\theta}\in\mathbb{R}^{s+1}}\{\langle\boldsymbol{\theta}, (x, \boldsymbol{y})\rangle - \Lambda_{(H_i,\boldsymbol{G}_i)}(\boldsymbol{\theta})\}$, respectively [Dembo and Zeitouni 1998, p. 44]. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_s)$,

$$(x, \boldsymbol{y}) \in \mathcal{F}^\circ_{(H_i,\boldsymbol{G}_i)} = \text{int}\{\nabla\Lambda_{(H_i,\boldsymbol{G}_i)}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{D}^\circ_{\Lambda_{(H_i,\boldsymbol{G}_i)}}\},$$

and let $\mathcal{F}_d^c$ denote the closure of the convex hull of the set of points $\{(h_i, \boldsymbol{\gamma}) : (h_i, \boldsymbol{\gamma}) \in \mathbb{R}^{s+1}, i \leq r\}$.

ASSUMPTION 4. *The closure of the convex hull of all points $(h_i, \boldsymbol{\gamma}) \in \mathbb{R}^{s+1}$ is a subset of the intersection of the interiors of the effective domains of the rate functions $I_i(x, \boldsymbol{y})\ \forall\ i \leq r$, that is, $\mathcal{F}_d^c \subset \cap_{i=1}^r \mathcal{F}^\circ_{(H_i,\boldsymbol{G}_i)}$.*

Henceforth, for ease of notation, we redefine all vectors as column vectors.

3. CHARACTERIZATION OF THE OPTIMAL BUDGET ALLOCATION

Recall that our problem context is Problem P (see §2.1), and our solution context involves three steps: sample from each of the designs to obtain objective function and constraint function estimators; estimate the feasible set of systems by observing their constraint function estimators; and estimate the optimal system from the estimated feasible set as that system having the smallest estimated objective function value. In this section, we rigorously characterize the optimal allocation as the allocation that minimizes the probability that the system returned as the "solution" at the end of some sampling effort t is not the true best system.

We build upon the characterization of the optimal budget allocation for general distributions in the presence of correlation between the objective and constraint estimators that was formally derived in Hunter [2011]. Hunter [2011] characterizes the optimal allocation as the solution to a concave maximization problem. We replicate the key results here, and then further characterize the solution to the concave maximization problem in terms of its Karush-Kuhn-Tucker (KKT) conditions [Boyd and Vandenberghe 2004]. This will set us up towards developing limiting approximations of the optimal budgeting plan in §4.

Recall that t is the computing budget, $\alpha_i \in [0, 1]$ is the fraction of the simulation budget devoted to system i, $\hat{H}_i = (\alpha_i t)^{-1}\sum_{k=1}^{\alpha_i t} H_{ik}$, and $\hat{G}_{ij} = (\alpha_i t)^{-1}\sum_{k=1}^{\alpha_i t} G_{ijk}$. From Hunter [2011], the probability of incorrectly estimating the best system, henceforth called the probability of false selection $P\{\mathrm{FS}\}$, is

$$P\{\mathrm{FS}\} = P\{[\underbrace{\cup_{j=1}^{s} \hat{G}_{1j} > \gamma_j}_{\substack{\text{system 1}\\ \text{estimated}\\ \text{infeasible}}}] \cup_{i=2}^{r} [(\underbrace{\cap_{j=1}^{s} \hat{G}_{ij} \le \gamma_j}_{\substack{\text{system } i\\ \text{estimated}\\ \text{feasible}}}) \cap (\underbrace{\hat{H}_i \le \hat{H}_1}_{\substack{\text{system } i\\ \text{``beats''}\\ \text{system 1}}})]\} \tag{1}$$

and the rate function of the $P\{\mathrm{FS}\}$ is

$$-\lim_{t\to\infty} \frac{1}{t} \log P\{\mathrm{FS}\} = \min\Big(\min_{j\le s} \alpha_1 J_{1j}(\gamma_j),\ \min_{2\le i\le r} R_i(\alpha_1, \alpha_i)\Big), \tag{2}$$

where $\alpha_1 J_{1j}(\gamma_j)$ is the rate function for the probability that system 1 is classified infeasible on the jth constraint, and $R_i(\alpha_1, \alpha_i) = \inf_{x_i \le x_{1i}, \boldsymbol{y}_i \le \boldsymbol{\gamma}} (\alpha_1 I_1(x_{1i}) + \alpha_i I_i(x_i, \boldsymbol{y}_i))$ is the rate function for the probability that system i is estimated feasible *and* system i has a "better" estimated objective function value than system 1.

We are interested in identifying the allocation $\boldsymbol{\alpha}$ that maximizes the rate of decay in (2). This problem can be formally stated as

$$\text{maximize} \quad \min\Big(\min_{j\le s} \alpha_1 J_{1j}(\gamma_j),\ \min_{2\le i\le r} R_i(\alpha_1, \alpha_i)\Big) \quad \text{s.t.} \quad \textstyle\sum_{i=1}^{r} \alpha_i = 1, \quad \alpha_i \ge 0. \tag{3}$$

We may equivalently write this problem as

$$\begin{aligned} \text{Problem } Q: \quad \text{maximize} \quad & z \quad \text{s.t.} \\ \alpha_1 J_{1j}(\gamma_j) &\ge z, \quad j \le s, \\ R_i(\alpha_1, \alpha_i) &\ge z, \quad i = 2, \dots, r, \\ \textstyle\sum_{i=1}^{r} \alpha_i &= 1,\ \alpha_i \ge 0, \end{aligned}$$

where for each $i = 2, \dots, r$, the values of $R_i(\alpha_1, \alpha_i)$ are obtained by solving

$$\text{Problem } R_i: \quad \text{minimize} \quad \alpha_1 I_1(x_{1i}) + \alpha_i I_i(x_i, \boldsymbol{y}_i) \quad \text{s.t.} \quad x_i \le x_{1i}, \quad \boldsymbol{y}_i \le \boldsymbol{\gamma}.$$

Thus Problem R_i allows us to solve for the rate function of system i for any particular $\boldsymbol{\alpha}$, where $i = 2, \dots, r$. As a matter of notation, we distinguish Problem R_i as an optimization problem in $(x_{1i}, x_i, \boldsymbol{y}_i^T)$, and $R_i(\alpha_1, \alpha_i)$ as the value of its objective function at optimality. By Hunter [2011], Problem R_i is a strictly convex minimization problem with a unique optimal solution. Further, Problem Q is a concave maximization problem to which the optimal solution exists, and the solution is strictly positive, that is, $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_r^*)^T > \mathbf{0}$, and hence all systems receive a nonzero portion of the sampling budget at optimality.

Two other problems that are related to Problem Q are of particular interest to us. The first is a reformulation of Problem Q obtained by converting the inequality constraints associated with $R_i(\cdot, \cdot)$ to equality constraints.

$$\begin{aligned} \text{Reformulated Problem } Q: \quad \text{maximize} \quad & z \quad \text{s.t.} \\ \alpha_1 J_{1j}(\gamma_j) &\ge z, \quad j \le s, \\ R_i(\alpha_1, \alpha_i) &= z, \quad i = 2, \dots, r, \\ \textstyle\sum_{i=1}^{r} \alpha_i &= 1,\ \alpha_i > 0. \end{aligned}$$

The above Reformulated Problem Q is equivalent to Problem Q in the sense that it is also a concave maximization problem with a solution that exists and coincides with $\boldsymbol{\alpha}^*$. (We do not go into the details of demonstrating such equivalence, but a proof closely follows the steps in Hunter and Pasupathy [2013].) Due to this equivalence, for ease of exposition, we henceforth refer to the Reformulated Problem Q as Problem Q without any loss in clarity.

The second related formulation of interest is Problem $\tilde{Q}$, obtained by relaxing the inequality constraints involving $J_{1\cdot}(\cdot)$ in Problem Q.

$$\text{Problem } \tilde{Q}: \qquad \text{maximize} \quad z \quad \text{s.t.}$$
$$R_i(\tilde{\alpha}_1, \tilde{\alpha}_i) = z, \quad i = 2, \ldots, r,$$
$$\textstyle\sum_{i=1}^{r} \tilde{\alpha}_i = 1, \ \tilde{\alpha}_i > 0.$$

Problem $\tilde{Q}$ also happens to be a concave maximization problem with a solution $\tilde{\boldsymbol{\alpha}}^*$ that is guaranteed to exist. Furthermore, since Problem $\tilde{Q}$ satisfies Slater's condition [Boyd and Vandenberghe 2004], the solution $\tilde{\boldsymbol{\alpha}}^*$ is obtained as the solution to the KKT system

$$R_i(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*) = R_k(\tilde{\alpha}_1^*, \tilde{\alpha}_k^*) \text{ for all } i, k \neq 1, \tag{4}$$

$$\sum_{i=2}^{r} \frac{\partial R_i\,(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*)/\partial \tilde{\alpha}_1}{\partial R_i\,(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*)/\partial \tilde{\alpha}_i} = 1. \tag{5}$$

Much of the analyses that follow will pertain to Problem $\tilde{Q}$, particularly through its KKT conditions given by (4) and (5). Our focus on Problem $\tilde{Q}$ (as opposed to Problem Q) will be justified through a result in §5 where we show that Problem $\tilde{Q}$ and Problem Q become equivalent as the number of systems under consideration becomes large in a certain precise sense.

We now present an explicit expression for the summands in equation (5).

LEMMA 3.1. *For a system i, the ith term in the summand of equation* (5) *is given by*

$$\frac{\partial R_i(\tilde{\alpha}_1, \tilde{\alpha}_i)/\partial \tilde{\alpha}_1}{\partial R_i(\tilde{\alpha}_1, \tilde{\alpha}_i)/\partial \tilde{\alpha}_i} = \frac{I_1(x_{1i}^*)}{I_i(x_i^*, \boldsymbol{y}_i^*)}, \tag{6}$$

where $(x_{1i}^*, x_i^*, \boldsymbol{y}_i^{*T})$ is the unique optimal solution to Problem R_i.

PROOF. For ease of exposition in the body of the paper, we introduce the following notation required for the proof here. Let $\lambda_{ix} \leq 0$ and $\lambda_{ij} \leq 0, j \leq s$, be Lagrange multipliers associated with the constraints in Problem R_i, where $\boldsymbol{\lambda}_i = (\lambda_{ix}, \lambda_{i1}, \ldots, \lambda_{is})^T$. Also define the following sets in terms of the Lagrange multipliers and the optimal solution to Problem R_i:

$$\mathcal{C}_F^{i*} = \{j : \lambda_{ij} = 0 \text{ and } y_{ij}^* \leq \gamma_j\};$$
$$\mathcal{C}_I^{i*} = \{j : \lambda_{ij} < 0 \text{ and } y_{ij}^* = \gamma_j\};$$
$$\Gamma^* = \{i : \lambda_{ix} < 0, x_{1i}^* = x_i^* \text{ and } \mathcal{C}_I^{i*} \text{ empty}, i \neq 1\};$$
$$\mathcal{S}_b^* = \{i : \lambda_{ix} = 0, x_i^* \leq x_{1i}^* \text{ and } \mathcal{C}_I^{i*} \text{ nonempty}, i \neq 1\};$$
$$\mathcal{S}_w^* = \{i : \lambda_{ix} < 0, x_{1i}^* = x_i^* \text{ and } \mathcal{C}_I^{i*} \text{ nonempty}, i \neq 1\}.$$

See Appendix A for the remainder of the proof. □

The sets $\Gamma^*, \mathcal{S}_b^*$, and $\mathcal{S}_w^*$ form a partition of the design space $\{1, 2, \ldots, r\}$, and the sets $\mathcal{C}_F^{i*}$ and $\mathcal{C}_I^{i*}$ form a partition of the set of constraints $\{1, 2, \ldots, s\}$ for each design i. The rate function corresponding to a system i will depend on its classification into exactly one of the sets $\Gamma^*, \mathcal{S}_b^*$, or $\mathcal{S}_w^*$, and the classification of each of its constraints into exactly one of the sets $\mathcal{C}_F^{i*}$ or $\mathcal{C}_I^{i*}$. The sets are best understood for the case in which the objective function and constraint estimators are mutually independent. In this case, the sets $\Gamma^*, \mathcal{S}_b^*$, and $\mathcal{S}_w^*$ correspond to the set of truly feasible designs, the set of truly infeasible designs that are better than system 1 in objective function value, and the set of truly infeasible designs that are worse than system 1 in objective function value, respectively. Under mutual independence, for each system i, the sets $\mathcal{C}_F^{i*}$ and $\mathcal{C}_I^{i*}$ are the sets of constraints on which system i is feasible and infeasible, respectively.

The terms of the simplified summand in equation (6) of Lemma 3.1 can be further simplified, as noted in the following Lemma 3.2.

LEMMA 3.2. *The KKT conditions for Problem $\tilde{Q}$ in* (4) *and* (5) *may be written as*

$$R_i(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*) = R_k(\tilde{\alpha}_1^*, \tilde{\alpha}_k^*) \textit{ for all } i, k \neq 1, \tag{7}$$

$$\sum_{i \in \Gamma^* \cup \mathcal{S}_w^*} \frac{I_1(x_{1i}^*)}{I_i(x_i^*, \boldsymbol{y}_i^*)} = 1. \tag{8}$$

PROOF. See Appendix B. □

Since the rate functions involved in (7) and (8) are unknown and cumbersome to estimate, solving this KKT system is usually impractical. However as we demonstrate in the sections that follow, the KKT conditions for Problem $\tilde{Q}$ become remarkably easier to solve under certain conditions, most notably when the number of systems r tends to infinity. Furthermore, as noted earlier, Problem Q and Problem $\tilde{Q}$ become equivalent in this asymptotic regime.

4. LIMITING APPROXIMATION TO THE OPTIMAL BUDGET ALLOCATION

With a view toward efficiently solving Problem $\tilde{Q}$, this section proposes a "closed-form" limiting approximation to the solution of the KKT system presented in Lemma 3.2, obtained as a certain asymptotic limit. Under the same limit, it is shown in §5 that the inequality constraint set that differentiates Problem Q and Problem $\tilde{Q}$ becomes redundant, thereby rendering them equivalent.

4.1. Allocations to Suboptimal Systems

Recall that the total number of systems $r = |\Gamma^* \cup S_w^* \cup S_b^*|$. In what follows, we denote $\tilde{r} = |\Gamma^* \cup S_w^*|$ and let $\tilde{r} \to \infty$, that is, we progressively include more systems into the set $\Gamma^* \cup S_w^*$ to obtain the closed-form sample allocation results. Our results require that $|\Gamma^* \cup S_w^*| \to \infty$ because our limiting argument requires an increasing number of systems that compete in objective function value with the best feasible system. Therefore $|S_b^*|$ may remain fixed or tend to infinity, as long as $|\Gamma^* \cup S_w^*| \to \infty$.

To further understand what sending $|\Gamma^* \cup \mathcal{S}_w^*| \to \infty$ means, consider the context where the objective function and constraint estimators for each system are mutually independent. In this context, $|\Gamma^* \cup \mathcal{S}_w^*| \to \infty$ implies that the collective cardinality of systems inferior to system 1, as measured by the objective function $h(\cdot)$, tends to infinity. In the more general context, the interpretation becomes slightly more nuanced. A sufficient condition guaranteeing that $\tilde{r} \to \infty$ in the general context is that the cardinality of the set of truly feasible systems tends to infinity.

The following regularity assumptions are made about the nature of systems added as $\tilde{r} \to \infty$. They are assumed to hold throughout the rest of the paper.

ASSUMPTION 5. *The means $(h_i, \boldsymbol{g}_i)$ satisfy* $\inf\{|h_i - h_1|, |g_{ij} - \gamma_j| : 1 < i \leq r, j \leq s\} > \epsilon$ *for some $\epsilon > 0$.*

ASSUMPTION 6. *There exists a compact set $\mathcal{C} \subset \mathbb{R}^{s+1}$ such that $(h_i, \boldsymbol{g}_i) \in \mathcal{C}$ and $(h_i, \gamma) \in \mathcal{C}$ for all $i \leq r$, and such that $\mathcal{C} \subset \cap_{i=1}^{r} \mathcal{F}^{o}_{(H_i, \boldsymbol{G}_i)}$*

ASSUMPTION 7. *There exist quadratic forms* $I_i^\ell(x,\boldsymbol{y}) = \frac{(x-h_i)^2}{2\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^2}(\boldsymbol{y}-\boldsymbol{g}_i)^T(\boldsymbol{y}-\boldsymbol{g}_i)$ *and* $I_i^u(x,\boldsymbol{y}) = \frac{(x-h_i)^2}{2\underset{\sim}{\sigma}^2} + \frac{1}{2\underset{\sim}{\sigma}^2}(\boldsymbol{y}-\boldsymbol{g}_i)^T(\boldsymbol{y}-\boldsymbol{g}_i)$ *such that* $0 < \underset{\sim}{\sigma}^2 \le \tilde{\sigma}^2 < \infty$ *and* $0 < I_i^\ell(x,\boldsymbol{y}) \le I_i(x,\boldsymbol{y}) \le I_i^u(x,\boldsymbol{y}) < \infty$ *for all* $\{(x,\boldsymbol{y}) \in \mathcal{C} : (x,\boldsymbol{y}) \neq (h_i,\boldsymbol{g}_i)\}$ *and for all* $i \in \{1\} \cup \Gamma^* \cup S_w^*$*, where we emphasize that* $\mathcal{C}$ *is a* compact *set.*

Assumptions 5 and 6 are regularity assumptions intended to limit the manner in which systems are introduced into the analysis. For example, Assumption 5 stipulates that systems that are included in the analysis do not progressively "approach" the best system or the feasibility threshold. This requirement ensures that the probability of correctly identifying the better of two systems, or the probability of correctly checking the feasibility of a given system, can be driven to one by increased sampling. In the absence of this requirement, it may be impossible to say with certainty as to whether a system is feasible or whether a system is better than another. Assumption 6 essentially does the opposite, that is, prevents addition of systems that become irrelevant in the limit because their performance measures become unbounded. Assumption 7 is essentially a structural assumption imposing limits on the "steepness" and "shallowness" of the rate functions on the compact set $\mathcal{C}$, expressed using multivariate Gaussian rate functions as envelopes. We emphasize that *Assumption 7 is on the compact set* $\mathcal{C}$*, and some algebra reveals that it holds for several common distribution families, e.g., Gaussian and exponential.*

As noted, our analysis in this section will involve sending the cardinality $\tilde{r} = |\Gamma^* \cup S_w^*|$ to infinity. To this extent, most quantities of interest, for example, $\tilde{\boldsymbol{\alpha}}^*$ and x_{1i}^*, are actually functions of $r = \tilde{r} + |S_b^*|$ and should be denoted $\tilde{\boldsymbol{\alpha}}^*(r), x_{1i}^*(r)$. However for notational convenience, we do not explicitly indicate such dependence.

We are now ready to state a result that lists key properties of the solution to Problem $\tilde{Q}$ under the stated asymptotic regime.

THEOREM 4.1. *Suppose Assumptions 5–7 hold, and recall that* $\tilde{r} = |\Gamma^* \cup \mathcal{S}_w^*|$*. Then the following statements are true.*

(i) There exists $\kappa_1 > 0$ *such that* $\tilde{\alpha}_1^*/\tilde{\alpha}_k^* > \kappa_1$ *for all* $k \in \Gamma^* \cup S_w^*$ *and all* r*.*
(ii) There exists $\kappa_2 < \infty$ *such that* $\tilde{\alpha}_i^*/\tilde{\alpha}_k^* < \kappa_2$ *for all* $i,k \in \Gamma^* \cup S_w^*$ *and all* r*.*
(iii) As $\tilde{r} \to \infty$*,* $\tilde{\alpha}_i^* = O(1/\tilde{r})$ *for all* $i \in \Gamma^* \cup S_w^*$*.*
(iv) As $\tilde{r} \to \infty$*,* $x_i^* = x_{1i}^* \to h_1$ *for all* $i \in \Gamma^* \cup S_w^*$*.*
(v) As $\tilde{r} \to \infty$*,* $\tilde{\alpha}_i^*/\tilde{\alpha}_1^* \to 0$ *for all* $i \in \Gamma^* \cup S_w^*$*.*

PROOF. *Proof of part (i).* We do not provide a proof for part (i) but note that it follows from equation (7) and Assumptions 6 and 7.

Proof of part (ii). For $i \in \Gamma^* \cup \mathcal{S}_w^*$, $R_i(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*) = \inf_{x_i = x_{1i}, \boldsymbol{y}_i \le \boldsymbol{\gamma}} (\tilde{\alpha}_1^* I_1(x_{1i}) + \tilde{\alpha}_i^* I_i(x_i, \boldsymbol{y}_i))$. Using Assumption 7, after some algebra, we see that for $i \in \Gamma^* \cup S_w^*$,

$$R_i(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*) \ge \frac{(h_i - h_1)^2}{2(\tilde{\sigma}^2/\tilde{\alpha}_1^* + \tilde{\sigma}^2/\tilde{\alpha}_i^*)} = \frac{\tilde{\alpha}_i^*(h_i - h_1)^2}{2(\tilde{\sigma}^2(\tilde{\alpha}_i^*/\tilde{\alpha}_1^*) + \tilde{\sigma}^2)}. \tag{9}$$

Similarly, we also have for $k \in \Gamma^* \cup S_w^*$,

$$R_k(\tilde{\alpha}_1^*, \tilde{\alpha}_k^*) \le \frac{\tilde{\alpha}_k^*(h_k - h_1)^2}{2(\underset{\sim}{\sigma}^2(\tilde{\alpha}_k^*/\tilde{\alpha}_1^*) + \underset{\sim}{\sigma}^2)} + \tilde{\alpha}_k^* \sum_{j=1}^{s} \frac{(g_{kj} - \gamma_j)^2}{2\underset{\sim}{\sigma}^2} \mathbb{I}\{g_{kj} > \gamma_j\}, \tag{10}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. Using (9) and (10), and since $R_i(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*) = R_k(\tilde{\alpha}_1^*, \tilde{\alpha}_k^*)$ by Lemma 3.2, we have that for $i, k \in \Gamma^* \cup S_w^*$,

$$\begin{aligned}\frac{\tilde{\alpha}_i^*}{\tilde{\alpha}_k^*} &\leq \left(\frac{(h_k - h_1)^2}{(\underset{\sim}{\sigma}^2(\tilde{\alpha}_k^*/\tilde{\alpha}_1^*) + \underset{\sim}{\sigma}^2)} + \sum_{j=1}^{s} \frac{(g_{kj} - \gamma_j)^2}{\underset{\sim}{\sigma}^2} \mathbb{I}\{g_{kj} > \gamma_j\} \right) \Big/ \left(\frac{(h_i - h_1)^2}{(\tilde{\sigma}^2(\tilde{\alpha}_i^*/\tilde{\alpha}_1^*) + \tilde{\sigma}^2)} \right) \\ &= \frac{\tilde{\sigma}^2}{\underset{\sim}{\sigma}^2} \frac{(h_k - h_1)^2((\tilde{\alpha}_i^*/\tilde{\alpha}_1^*) + 1)}{(h_i - h_1)^2((\tilde{\alpha}_k^*/\tilde{\alpha}_1^*) + 1)} + \frac{((\tilde{\alpha}_i^*/\tilde{\alpha}_1^*) + 1) \sum_{j=1}^{s} (g_{kj} - \gamma_j)^2 \mathbb{I}\{g_{kj} > \gamma_j\}}{(h_i - h_1)^2} \\ &\leq \frac{\tilde{\sigma}^2}{\underset{\sim}{\sigma}^2} \frac{(h_k - h_1)^2(\kappa_1^{-1} + 1)}{(h_i - h_1)^2} + \frac{(\kappa_1^{-1} + 1) \sum_{j=1}^{s} (g_{kj} - \gamma_j)^2 \mathbb{I}\{g_{kj} > \gamma_j\}}{(h_i - h_1)^2}.\end{aligned}$$

Now use part (i) and Assumption 5 to conclude that the assertion in part (ii) holds.

Proof of part (iii). From part (ii), we see that $\tilde{\alpha}_i^* \kappa_2^{-1} \leq \tilde{\alpha}_k^*$ for $i, k \in \Gamma^* \cup S_w^*$. Then for $i \in \Gamma^* \cup S_w^*$, we have $\tilde{\alpha}_1^* + \tilde{r}\tilde{\alpha}_i^* \kappa_2^{-1} \leq \sum_{k=1}^{\tilde{r}} \tilde{\alpha}_k^*$ and hence $\tilde{\alpha}_i^* = O(1/\tilde{r})$ as $\tilde{r} \to \infty$.

Proof of part (iv). Let $(x_i^\ell, \boldsymbol{y}_i^\ell)$ and $(x_i^u, \boldsymbol{y}_i^u)$ denote the solutions to $\inf_{x_{1i}=x_i, \boldsymbol{y}_i} \tilde{\alpha}_1^* I_1^u(x_{1i}) + \tilde{\alpha}_i^* I_i^\ell(x_i, \boldsymbol{y}_i)$ and $\inf_{x_{1i}=x_i, \boldsymbol{y}_i} \tilde{\alpha}_1^* I_1^\ell(x_{1i}) + \tilde{\alpha}_i^* I_i^u(x_i, \boldsymbol{y}_i)$, respectively. From the explicit forms of $I_1^\ell(\cdot), I_1^u(\cdot), I_i^\ell(\cdot, \cdot)$, and $I_i^u(\cdot, \cdot)$, the stationarity conditions imply that $\dfrac{x_i^\ell - h_1}{h_i - x_i^\ell} = \dfrac{\tilde{\alpha}_i^*}{\tilde{\alpha}_1^*} \dfrac{\underset{\sim}{\sigma}^2}{\tilde{\sigma}^2}$ and $\dfrac{x_i^u - h_1}{h_i - x_i^u} = \dfrac{\tilde{\alpha}_i^*}{\tilde{\alpha}_1^*} \dfrac{\tilde{\sigma}^2}{\underset{\sim}{\sigma}^2}$. Solving for x_i^ℓ and x_i^u, along with the assertion in part (i) and Assumptions 5 and 6, implies that there exist constants $\kappa, \kappa' \in (0, \infty)$ such that for all $i \in \Gamma^* \cup S_w^*$ and all r,

$$x_i^\ell - h_1 = \frac{\tilde{\alpha}_i^*}{\tilde{\alpha}_1^*} \left(\frac{h_i - h_1}{\frac{\tilde{\alpha}_i^*}{\tilde{\alpha}_1^*} + \frac{\tilde{\sigma}^2}{\underset{\sim}{\sigma}^2}} \right) \geq \frac{\tilde{\alpha}_i^*}{\tilde{\alpha}_1^*} \kappa \quad \text{and} \quad x_i^u - h_1 = \frac{\tilde{\alpha}_i^*}{\tilde{\alpha}_1^*} \left(\frac{h_i - h_1}{\frac{\tilde{\alpha}_i^*}{\tilde{\alpha}_1^*} + \frac{\underset{\sim}{\sigma}^2}{\tilde{\sigma}^2}} \right) \leq \frac{\tilde{\alpha}_i^*}{\tilde{\alpha}_1^*} \kappa'. \quad (11)$$

We know from Assumption 7 that $x_{1i}^* \in (x_i^\ell, x_i^u)$, and hence

$$\frac{I_1^\ell(x_i^\ell)}{I_i^u(x_i^u, \boldsymbol{y}_i^*)} \leq \frac{I_1(x_{1i}^*)}{I_i(x_{1i}^*, \boldsymbol{y}_i^*)} \leq \frac{I_1^u(x_i^u)}{I_i^\ell(x_i^\ell, \boldsymbol{y}_i^*)}.$$

Writing a corresponding inequality for system k, (dividing) and then using the explicit forms for $I_1^\ell(\cdot)$ and $I_1^u(\cdot)$, we can write

$$\frac{\underset{\sim}{\sigma}^2}{\tilde{\sigma}^2} \frac{(x_i^\ell - h_1)^2}{(x_k^u - h_1)^2} \frac{I_k^\ell(x_{1k}^*, \boldsymbol{y}_k^*)}{I_i^u(x_{1i}^*, \boldsymbol{y}_i^*)} \leq \frac{I_1(x_{1i}^*)/I_i(x_{1i}^*, \boldsymbol{y}_i^*)}{I_1(x_{1k}^*)/I_k(x_{1k}^*, \boldsymbol{y}_k^*)} \leq \frac{\tilde{\sigma}^2}{\underset{\sim}{\sigma}^2} \frac{(x_i^u - h_1)^2}{(x_k^\ell - h_1)^2} \frac{I_k^u(x_{1k}^*, \boldsymbol{y}_k^*)}{I_i^\ell(x_{1i}^*, \boldsymbol{y}_i^*)}. \quad (12)$$

Using (11) and (12), and noting from part (ii) that $\tilde{\alpha}_i^*/\tilde{\alpha}_k^* < \kappa_2 \in (0, \infty)$ for all $i, k \in \Gamma^* \cup S_w^*$ and all r, we see that there exists $\kappa_3 \in (0, \infty)$ such that $\dfrac{I_1(x_{1i}^*)}{I_i(x_i^*, \boldsymbol{y}_i^*)} < \kappa_3 \dfrac{I_1(x_{1k}^*)}{I_k(x_k^*, \boldsymbol{y}_k^*)}$ for all $i, k \in \Gamma^* \cup S_w^*$ and all r.

Let $\ell = \arg \min_{i \in \Gamma^* \cup \mathcal{S}_w^*} \dfrac{I_1(x_{1i}^*)}{I_i(x_i^*, \boldsymbol{y}_i^*)}$. Then from the KKT condition for Problem Q in (8),

$$\frac{I_1(x_{1\ell}^*)}{I_\ell(x_\ell^*, \boldsymbol{y}_\ell^*)} \leq \frac{1}{\tilde{r}}.$$

Further, from previous arguments, we see that

$$\frac{I_1(x_{1i}^*)}{I_i(x_i^*, \boldsymbol{y}_i^*)} \leq \kappa_3 \frac{I_1(x_{1\ell}^*)}{I_\ell(x_\ell^*, \boldsymbol{y}_\ell^*)} \leq \frac{\kappa_3}{\tilde{r}} \text{ for all } i \in \Gamma^* \cup \mathcal{S}_w^*. \quad (13)$$

Conclude from (13) that $x_{1i}^* = x_i^* \to h_1$ as $r \to \infty$ for all $i \in \Gamma^* \cup S_w^*$.

Proof of part (v). Notice from the first-order conditions for stationarity in Problem R_i that $\tilde{\alpha}_i^*/\tilde{\alpha}_1^* = -\dfrac{\partial I_1(x_{1i}^*)/\partial x_{1i}}{\partial I_i(x_i^*, \boldsymbol{y}_i^*)/\partial x_{1i}}$, and then use the assertion in part (iv). □

The main assertions of Theorem 4.1 are contained in parts (iii), (iv), and (v). In view of Assumptions 5 and 6, the assertion in part (iii) of Theorem 4.1 should be intuitively clear. Specifically, if the systems being included in the analysis are "similar" in the sense described by Assumptions 5 and 6, and we increase the number of "contending" systems without bound, the fraction of the budget that should be devoted to any particular suboptimal system should tend to zero. What is less intuitive is part (v) of Theorem 4.1, which states that as the number of systems in $\Gamma^* \cup \mathcal{S}_w^*$ tends to infinity, optimality dictates that the fraction of the budget given to the optimal design 1 far exceed the fraction given to any of the suboptimal designs in $\Gamma^* \cup \mathcal{S}_w^*$. This result makes sense if one thinks of each of the suboptimal systems as individually attempting to "beat the best design" by inducing a false selection event. Optimality dictates that the best design receive far more samples than these competitors in a bid to minimize the probability of occurrence of the most likely of the numerous false selection events, made possible by the assumption $|\Gamma^* \cup \mathcal{S}_w^*| \to \infty$. Part (iv) of Theorem 4.1 is algebraic and notes that the optimal solution to Problem R_i tends to the primary performance measure h_1 of system 1 as the number of systems in $\Gamma^* \cup \mathcal{S}_w^*$ tends to infinity.

We are now ready to present Theorem 4.2 — the main result of the paper. Theorem 4.2 asserts that as $\tilde{r} \to \infty$, the ratio of the rate $R_i(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*)$ to the optimal fraction $\tilde{\alpha}_i^*$ for the ith system tends to the minimum value attained by the rate function $I_i(x_i, \boldsymbol{y}_i)$ in the rectangular region $x_i \leq h_1, \boldsymbol{y}_i \leq \gamma$. This result combined with the fact that the KKT system for Problem Q in (7) dictates equating $R_i(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*) = R_k(\tilde{\alpha}_1^*, \tilde{\alpha}_k^*)$ for $i, k \in \{2, 3, \ldots, r\}$ will form the basis of our proposed allocation.

THEOREM 4.2. *Suppose Assumptions 5–7 hold. As* $\tilde{r} = |\Gamma^* \cup \mathcal{S}_w^*| \to \infty$,

$$\frac{R_i(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*)}{\tilde{\alpha}_i^*} = \frac{R_i(\tilde{\alpha}_i^*)}{\tilde{\alpha}_i^*} = \inf_{x_i \leq h_1, \boldsymbol{y}_i \leq \boldsymbol{\gamma}} I_i(x_i, \boldsymbol{y}_i) \textit{ for all } i = 2, \ldots, r.$$

PROOF. We know that

$$\frac{R_i(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*)}{\tilde{\alpha}_i} = \inf_{x_i \leq x_{1i}, \boldsymbol{y}_i \leq \boldsymbol{\gamma}} \frac{\tilde{\alpha}_1^*}{\tilde{\alpha}_i^*} I_1(x_{1i}) + I_i(x_i, \boldsymbol{y}_i). \tag{14}$$

Since $I_1(x_{1i}^*) = 0$ and $x_{1i}^* = h_1$ for all $i \in \mathcal{S}_b^*$, equation (14) implies that the result immediately holds for $i \in \mathcal{S}_b^*$. Consider $i \in \Gamma^* \cup \mathcal{S}_w^*$. Recall that the KKT conditions for Problem R_i imply

$$\frac{\tilde{\alpha}_1^*}{\tilde{\alpha}_i^*} = -\frac{\partial I_i(x_i^*, \boldsymbol{y}_i^*)/\partial x_i}{\partial I_1(x_{1i}^*)/\partial x_{1i}}.$$

Therefore

$$\frac{\tilde{\alpha}_1^*}{\tilde{\alpha}_i^*} I_1(x_{1i}^*) = -\frac{I_1(x_{1i}^*)}{\partial I_1(x_{1i}^*)/\partial x_{1i}} \partial I_i(x_i^*, \boldsymbol{y}_i^*)/\partial x_i.$$

Since $I_1(\cdot)$ is a smooth function attaining its minimum at h_1, we see that

$$\frac{I_1(x_{1i}^*)}{\partial I_1(x_{1i}^*)/\partial x_{1i}} = O(x_{1i}^* - h_1).$$

Using this along with the facts that from part (iv) of Theorem 4.1, $x_{1i}^* = x_i^* \to h_1$, and $\partial I_i(x_i^*, \boldsymbol{y}_i^*)/\partial x_i$ is bounded away from infinity, we obtain the result. □

To see why Theorem 4.2 has important consequences, define the *score* $\mathbb{S}_i$ for any sub-optimal system i as

$$\mathbb{S}_i := \inf_{x_i \le h_1, \boldsymbol{y}_i \le \boldsymbol{\gamma}} I_i(x_i, \boldsymbol{y}_i) \text{ for all } i = 2, \ldots, r.$$

Then, we notice from Theorem 4.2 that the ratio $R_i(\tilde{\alpha}_i^*)/\tilde{\alpha}_i^*$ tends to the score $\mathbb{S}_i$ in the limit $\tilde{r} \to \infty$. Since the KKT conditions of Problem $\tilde{Q}$ dictate that $R_i(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*) = R_i(\tilde{\alpha}_1^*, \tilde{\alpha}_j^*)$ for all i, this suggests that the optimal budget fraction for the ith system is inversely proportional to its score. This result is formally stated in Theorem 4.3.

THEOREM 4.3. *As $\tilde{r} = |\Gamma^* \cup \mathcal{S}_w^*| \to \infty$, the optimal allocation vector $\tilde{\boldsymbol{\alpha}}^*$ satisfies*

$$\frac{\tilde{\alpha}_i^*}{\tilde{\alpha}_k^*} = \frac{\mathbb{S}_k}{\mathbb{S}_i} = \frac{\inf\limits_{x_k \le h_1, \boldsymbol{y}_k \le \boldsymbol{\gamma}} I_k(x_k, \boldsymbol{y}_k)}{\inf\limits_{x_i \le h_1, \boldsymbol{y}_i \le \boldsymbol{\gamma}} I_i(x_i, \boldsymbol{y}_i)}.$$

Theorem 4.3 is important in that it provides a potentially simple budget allocation rule when the number of systems in contention is “large.” Computing the optimal allocation vector through an exact calculation as provided through convex optimization of Problem Q is comparatively burdensome. Optimal allocation using Theorem 4.3 becomes especially tractable in contexts where the score $\mathbb{S}_i$ is easily estimated. Table I presents expressions for the score in a number of such contexts. Notice that the score expression in the fifth row of Table I (independent normal context) is particularly intuitive — it involves a penalty term for suboptimality and a penalty term for infeasibility, each measured in standard deviation units.

4.2. Allocation to the Best Feasible System

The following Theorem 4.4 provides a sense of the fraction of the budget allocated to the best feasible system when optimally apportioning the simulation budget.

THEOREM 4.4. *Suppose Assumptions 5–7 hold. Then the following statements are true.*

(i) $\left(\sqrt{\frac{1}{4+\frac{8M}{\epsilon}}}\right)\frac{\utilde{\sigma}^3}{\tilde{\sigma}^3}\sqrt{\sum_{i\in\Gamma^*\cup S_w^*}\tilde{\alpha}_i^{*2}} \;\le\; \tilde{\alpha}_1^* \le 2\frac{\tilde{\sigma}^3}{\utilde{\sigma}^3}\sqrt{\sum_{i\in\Gamma^*\cup S_w^*}\tilde{\alpha}_i^{*2}}$, *where $M = diam(\mathcal{C})$ and ϵ are constants implicit in Assumptions* 6 *and* 5, *respectively.*

(ii) As $\tilde{r} \to \infty$, $\tilde{\alpha}_1^* = \Theta(1/\sqrt{\tilde{r}})$.

PROOF. *Proof of assertion (i).* We first prove the upper bound. We see that

$$\frac{I_1(x_{1i}^*)}{I_i(x_i^*, \boldsymbol{y}_i^*)} \le \frac{I_1^u(x_i^u)}{I_i^\ell(x_i^\ell, \boldsymbol{y}_i^\ell)} = \frac{(x_i^u - h_1)^2/2\utilde{\sigma}^2}{\frac{(x_i^\ell - h_i)^2}{2\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^2}(\boldsymbol{y}_i^\ell - \boldsymbol{g}_i)^T(\boldsymbol{y}_i^\ell - \boldsymbol{g}_i)}. \tag{15}$$

(Recall that $(x_i^\ell, \boldsymbol{y}_i^\ell)$ and $(x_i^u, \boldsymbol{y}_i^u)$ are the solutions to $\inf_{x_{1i}=x_i,\boldsymbol{y}_i} \tilde{\alpha}_1^* I_1^u(x_{1i}) + \tilde{\alpha}_i^* I_i^\ell(x_i, \boldsymbol{y}_i)$ and $\inf_{x_{1i}=x_i,\boldsymbol{y}_i} \tilde{\alpha}_1^* I_1^\ell(x_{1i}) + \tilde{\alpha}_i^* I_i^u(x_i, \boldsymbol{y}_i)$, respectively.) Also,

$$x_i^u - h_1 = \frac{c}{1+c}(h_i - h_1), \quad h_i - x_i^\ell = \frac{1}{1+c'}(h_i - h_1) \tag{16}$$

where $c = \frac{\tilde{\alpha}_i^*}{\tilde{\alpha}_1^*}\frac{\tilde{\sigma}^2}{\utilde{\sigma}^2}, c' = \frac{\tilde{\alpha}_i^*}{\tilde{\alpha}_1^*}\frac{\utilde{\sigma}^2}{\tilde{\sigma}^2}$. Plugging (16) in (15), we have

$$\frac{I_1(x_{1i}^*)}{I_i(x_i^*, \boldsymbol{y}_i^*)} \le \frac{\tilde{\alpha}_i^{*2}}{\tilde{\alpha}_1^{*2}}\frac{\tilde{\sigma}^6}{\utilde{\sigma}^6}(\frac{1+c'}{1+c})^2. \tag{17}$$

Table I. Score expressions for some special contexts. “Optimal” convergence dictates that the fraction of the simulation budget to a system be inversely proportional to its score.

Context	Correlation	Score Expression
General unconstrained[1]	NA	$\mathbb{S}_i = I_i(h_1),\ \forall\, i \neq 1.$
$\mathcal{N}(h_i, \sigma_i^2)$ objective, unconstrained[2]	×	$\mathbb{S}_i = \dfrac{(h_i - h_1)^2}{2\sigma_i^2},\ \forall\, i \neq 1.$
Fixed objective, general constraints[3]	✓	$\mathbb{S}_i = \inf_{\boldsymbol{y}_i \leq \boldsymbol{\gamma}} I_i(\boldsymbol{y}_i)\mathbb{I}\{i \in \mathcal{S}_b\} + \infty\mathbb{I}\{i \in \Gamma \cup \mathcal{S}_w\} + \min_{j \leq s} J_{ij}(\gamma_j)\mathbb{I}\{i = 1\}$, for all $i \leq r$.
Bernoulli(h_i) objective, Bernoulli(g_{ij}) constraints[4]	×	$\mathbb{S}_i = B(h_1, h_i)\mathbb{I}\{h_i > h_1\} + \sum_{j=1}^{s} B(g_{ij}, \gamma_j)\mathbb{I}\{g_{ij} > \gamma_j\},\ \forall\, i \neq 1,$ where $B(a,b) = a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}$.
$\mathcal{N}(h_i, \sigma_i^2)$ objective, $\mathcal{N}(g_{ij}, \sigma_{ij}^2)$ constraints[4]	×	$\mathbb{S}_i = \dfrac{(h_i - h_1)^2}{2\sigma_i^2}\mathbb{I}\{h_i > h_1\} + \sum_{j=1}^{s} \dfrac{(\gamma_j - g_{ij})^2}{2\sigma_{ij}^2}\mathbb{I}\{g_{ij} > \gamma_j\},\ \forall\, i \neq 1.$
$\mathcal{N}(\left[\begin{smallmatrix} h_i \\ \boldsymbol{g}_i \end{smallmatrix}\right], \Sigma_i)$ objective and constraints[5]	✓	$\mathbb{S}_i = \frac{1}{2}\Bigg(\dfrac{\Delta h_i^2}{\sigma_i^2}\mathbb{I}\{i \in \Gamma^*\}$ $+ \Delta\boldsymbol{g}_i(\mathcal{C}_I^{i*})^T \Sigma_i(\boldsymbol{G}_i(\mathcal{C}_I^{i*}))^{-1}\left[\Delta\boldsymbol{g}_i(\mathcal{C}_I^{i*})\right]\mathbb{I}\{i \in \mathcal{S}_b^*\}$ $+ \left[\begin{smallmatrix} \Delta h_i \\ \Delta\boldsymbol{g}_i(\mathcal{C}_I^{i*}) \end{smallmatrix}\right]^T \Sigma_i(\left[\begin{smallmatrix} H_i \\ \boldsymbol{G}_i(\mathcal{C}_I^{i*}) \end{smallmatrix}\right])^{-1}\left[\begin{smallmatrix} \Delta h_i \\ \Delta\boldsymbol{g}_i(\mathcal{C}_I^{i*}) \end{smallmatrix}\right]\mathbb{I}\{i \in \mathcal{S}_w^*\}\Bigg),\ \forall\, i \neq 1,$ where $\Delta\boldsymbol{g}_i = \boldsymbol{\gamma} - \boldsymbol{g}_i$, $\Delta h_i = h_1 - h_i$. (See §2.2 for notation.)

Related Settings: [1]Glynn and Juneja [2004]; [2]Chen et al. [2000]; [3]Szechtman and Yücesan [2008]; [4]Hunter and Pasupathy [2013]; [5]Hunter [2011].

Notice that $c, c' \to 0$ as $\tilde{r} \to 0$ from part (v) of Theorem 4.1. Using this in (17), and since (6) holds, we obtain the upper bound in assertion (i).

To obtain the lower bound in assertion (i), we use

$$x_i^{\ell} - h_1 = \frac{c'}{1+c'}(h_i - h_1), \quad h_i - x_i^u = \frac{1}{1+c}(h_i - h_1) \tag{18}$$

to write

$$\begin{aligned} \frac{I_1(x_{1i}^*)}{I_i(x_i^*, \boldsymbol{y}_i^*)} \geq \frac{I_i^{\ell}(x_i^{\ell})}{I_i^u(x_i^u, \boldsymbol{y}_i^u)} &= \frac{(x_i^{\ell} - h_1)^2/2\tilde{\sigma}^2}{\dfrac{(x_i^u - h_i)^2}{2\underset{\sim}{\sigma}^2} + \dfrac{1}{2\underset{\sim}{\sigma}^2}(\boldsymbol{y}_i^u - \boldsymbol{g}_i)^T(\boldsymbol{y}_i^u - \boldsymbol{g}_i)} \\ &= (\frac{1+c}{1+c'})^2 \frac{\tilde{\alpha}_i^{*2}}{\tilde{\alpha}_1^{*2}} \frac{\underset{\sim}{\sigma}^6}{\tilde{\sigma}^6} \frac{1}{1 + \dfrac{(\boldsymbol{y}_i^u - g_i)^T(\boldsymbol{y}_i^u - g_i)}{(h_i - h_1)^2}(1+c')^2}. \end{aligned} \tag{19}$$

Again noticing that $c, c' \to 0$ as $\tilde{r} \to 0$, and using Assumptions 5 and 6 to bound the quadratic form appearing in the denominator of the last term on the right-hand side of (19), we have that

$$\frac{I_1(x_{1i}^*)}{I_i(x_i^*, \boldsymbol{y}_i^*)} \geq 0.25 \frac{\tilde{\alpha}_i^{*2}}{\tilde{\alpha}_1^{*2}} \frac{\underset{\sim}{\sigma}^6}{\tilde{\sigma}^6} \frac{1}{1 + \frac{2M}{\epsilon}}.$$

Using this in (17), we obtain the lower bound in assertion (i).

Proof of assertion (ii). The proof of assertion (ii) follows trivially from the definition of $\Theta(\cdot)$ and from assertion (i). □

We emphasize that Theorem 4.4 provides only a sense of the optimal allocation to the best system as $\tilde{r} \to \infty$. Theorem 4.4 asserts that, like the suboptimal systems, the allocation to the best system also tends to zero, albeit at a much slower rate. Given that Theorem 4.4 only determines the optimal allocation to the best system to within a constant, implementation usually involves choosing this constant heuristically, or making distributional assumptions on the observations obtained from the various systems and then using the KKT condition in equation (8).

5. THE SCORE ALLOCATIONS

We now remind the reader that all the results in the previous section, Theorems 4.1, 4.2, and 4.3, pertain not to the real problem of interest, Problem Q, but to its relaxation, Problem $\tilde{Q}$. A natural question that follows is how the insight derived from Theorems 4.1, 4.2, and 4.3 applies to Problem Q. To address this question, we demonstrate through the following result that Problem $\tilde{Q}$ and Problem Q are equivalent as $\tilde{r} \to \infty$. The implication is that all insight we have derived thus far about the solution $\tilde{\boldsymbol{\alpha}}^*$ transfers over to the solution $\boldsymbol{\alpha}^*$ as $\tilde{r} \to \infty$.

THEOREM 5.1. *For large enough $\tilde{r}$, $\tilde{\boldsymbol{\alpha}}^* = \boldsymbol{\alpha}^*$.*

PROOF. Since the only difference between Problem $\tilde{Q}$ and Problem Q is the constraint set $\alpha_1 J_{1j}(\gamma_j) \geq z$ for all $j \leq s$, the assertion will follow if we show that $\tilde{\boldsymbol{\alpha}}^*$ satisfies the inequality $\tilde{\alpha}_1^* J_{1j}(\gamma_j) \geq z^*$ for large enough $\tilde{r}$ for all $j \leq s$, where z^* is the optimal value of Problem $\tilde{Q}$.

To see this, consider any $i \in |\Gamma^* \cup S_w^*|$. Notice from part (v) of Theorem 4.1 that $\tilde{\alpha}_1^*/\tilde{\alpha}_i^* \to \infty$ as $\tilde{r} \to \infty$. This implies that for $j \leq s$, $\lim_{\tilde{r}\to\infty}(\tilde{\alpha}_1^*/\tilde{\alpha}_i^*)J_{1j}(\gamma_j) = \infty$. However, we know from Theorem 4.2 that $\lim_{\tilde{r}\to\infty} z^*/\tilde{\alpha}_i^* = \lim_{\tilde{r}\to\infty} R_i(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*)/\tilde{\alpha}_i^* < \infty$. Conclude that $\tilde{\alpha}_1^* J_{1j}(\gamma_j) \geq z^*$ for large enough $\tilde{r}$ for all $j \leq s$. □

We now outline what we call the SCORE (Sampling Criteria for Optimization using Rate Estimators) allocations. Specifically, we outline two versions of SCORE allocations, called SCORE and SCORE-c (SCORE – correlated). In both allocations, we allocate to the suboptimal systems using the result in Theorem 4.3 and the scores provided in Table I, which are optimal solutions to Problem Q in the sense described by Theorem 5.1. In the SCORE-c allocation, we implement the KKT condition in equation (8) to determine the allocation to the best feasible system, system 1. To implement the KKT condition, we calculate the scores for the suboptimal systems and set $c_i := \mathbb{S}_i^{-1}/(\sum_{k=2}^{r} \mathbb{S}_k^{-1})$. We then solve

$$\sum_{i \in \Gamma^* \cup \mathcal{S}_w^*} \frac{I_1(x_{1i}^*(\alpha_1^*, c_i(1-\alpha_1^*)))}{I_i(x_i^*(\alpha_1^*, c_i(1-\alpha_1^*)), y_i^*(\alpha_1^*, c_i(1-\alpha_1^*)))} = 1 \tag{20}$$

for α_1^*. Then the SCORE-c allocation for the suboptimal systems is $\alpha_i^* = c_i(1-\alpha_1^*)$ for all $i = 2, \ldots, r$.

The SCORE allocation is identical to the SCORE-c allocation, except that when solving the root-finding problem in equation (20), the sets Γ^* and S_w^* and the rate functions are evaluated under the assumption of independence. That is, we instead solve

$$\sum_{i \in \Gamma} \frac{I_1(x_i^*(\alpha_1, c_i(1-\alpha_1)))}{I_i(x_i^*(\alpha_1, c_i(1-\alpha_1)))} + \sum_{i \in \mathcal{S}_w} \frac{I_1(x_i^*(\alpha_1, c_i(1-\alpha_1)))}{I_i(x_i^*(\alpha_1, c_i(1-\alpha_1))) + \sum_{j \in \mathcal{C}_I^i} J_{ij}(\gamma_j)} = 1, \tag{21}$$

where $x_i^*(\alpha_1, c_i(1-\alpha_1))$ is the solution to $\inf_{x_i}(\alpha_1 I_1(x_i) + c_i(1-\alpha_1)I_i(x_i))$, Γ is the set of truly feasible systems, $\mathcal{S}_w$ is the set of truly infeasible and worse systems, and $\mathcal{C}_I^i$ is the set of constraints on which system i is truly infeasible (see also Hunter and Pasupathy [2013]). While there are many ways to choose an allocation to system 1, we propose these two allocations. As shown in the numerical section, at least under the multivariate normal assumption, the SCORE allocation is fast and performs well for large numbers of systems.

Example 5.2. Since our implementation focus is on the Gaussian case, we now present the SCORE allocation for the case of multivariate normal random variables with correlation, assuming all distribution parameters are known. Suppose the random observations of the objective and constraints from system i have a normal distribution with mean $(h_i, \boldsymbol{g}_i)^T$ and covariance matrix Σ_i for all $i \leq r$. From Table I, letting $\Delta\boldsymbol{g}_i = \boldsymbol{\gamma} - \boldsymbol{g}_i$, $\Delta h_i = h_1 - h_i$ (see §2.2 for notation), we first calculate

$$\begin{aligned} \mathbb{S}_i &= \inf_{x_i \leq h_1, \boldsymbol{y}_i \leq \boldsymbol{\gamma}} I_i(x_i, \boldsymbol{y}_i) = \inf_{x_i \leq h_1, \boldsymbol{y}_i \leq \boldsymbol{\gamma}} \frac{1}{2} \begin{bmatrix} x - h_i \\ \boldsymbol{y} - \boldsymbol{g}_i \end{bmatrix}^T \Sigma_i^{-1} \begin{bmatrix} x - h_i \\ \boldsymbol{y} - \boldsymbol{g}_i \end{bmatrix} \\ &= \frac{1}{2}\Bigg(\frac{\Delta h_i^2}{\sigma_i^2} \mathbb{I}\{i \in \Gamma^*\} + \Delta\boldsymbol{g}_i(\mathcal{C}_I^{i*})^T \Sigma_i (\boldsymbol{G}_i(\mathcal{C}_I^{i*}))^{-1} \left[\Delta\boldsymbol{g}_i(\mathcal{C}_I^{i*})\right] \mathbb{I}\{i \in \mathcal{S}_b^*\} \\ &\qquad + \left[\begin{smallmatrix} \Delta h_i \\ \Delta\boldsymbol{g}_i(\mathcal{C}_I^{i*}) \end{smallmatrix}\right]^T \Sigma_i(\left[\begin{smallmatrix} H_i \\ \boldsymbol{G}_i(\mathcal{C}_I^{i*}) \end{smallmatrix}\right])^{-1} \left[\begin{smallmatrix} \Delta h_i \\ \Delta\boldsymbol{g}_i(\mathcal{C}_I^{i*}) \end{smallmatrix}\right] \mathbb{I}\{i \in \mathcal{S}_w^*\} \Bigg) \end{aligned} \tag{22}$$

for all suboptimal systems $i \neq 1$, where we note that (22) is a quadratic program with box constraints. For all $i \in \Gamma \cup \mathcal{S}_w$, set $c_i = \mathbb{S}_i^{-1}/(\sum_{k=2}^r \mathbb{S}_k^{-1})$.

Now we are ready to compute the allocation to system 1 using equation (21). Since all rate functions are normal, $x_i^*(\alpha_1, c_i(1-\alpha_1))$ from equation (21) can be written in closed form as

$$x_i^*(\alpha_1, c_i(1-\alpha_1)) = \frac{(\alpha_1/\sigma_1^2)h_1 + (c_i(1-\alpha_1)/\sigma_i^2)h_i}{\alpha_1/\sigma_1^2 + c_i(1-\alpha_1)/\sigma_i^2},$$

which implies equation (21) is equivalent to

$$\sum_{i\in\Gamma} \frac{\sigma_1^2/\alpha_1^2}{\sigma_i^2/[c_i(1-\alpha_1)]^2} + \sum_{i\in\mathcal{S}_w} \frac{\frac{(\sigma_1^2/\alpha_1^2)(h_1-h_i)^2}{[\sigma_1^2/\alpha_1+\sigma_i^2/(c_i(1-\alpha_1))]^2}}{\frac{(\sigma_i^2/[c_i(1-\alpha_1)]^2)(h_1-h_i)^2}{[\sigma_1^2/\alpha_1+\sigma_i^2/(c_i(1-\alpha_1))]^2} + \sum_{j\in\mathcal{C}_I^i} \frac{(\gamma_j-g_{ij})^2}{\sigma_{ij}^2}} = 1.$$

This one-dimensional root finding problem can be solved numerically to find α_1^*. Setting $\alpha_i^* = (\mathbb{S}_i^{-1}(1-\alpha_1^*))/(\sum_{k=2}^r \mathbb{S}_k^{-1})$ for all $i \neq 1$ yields the proposed SCORE allocation.

We note here that the steps to compute the SCORE-c allocation under the multivariate normal assumption are identical to the steps for the SCORE allocation, except that at each step in solving the root-finding problem in equation (20), $r-1$ quadratic programs must be solved to obtain the values of the rate functions.

6. A SEQUENTIAL ALGORITHM FOR IMPLEMENTATION

To implement the proposed optimal allocation sequentially, we use the following Algorithm 1. Algorithm 1 evolves in stages by collecting a fixed number of simulation observations from systems chosen strategically at the beginning of each stage, updating the relevant estimators, and then proceeding to the next stage to begin the process over again. Specifically, at the beginning of each stage, $\delta > 0$ observations are obtained from systems chosen with probabilities in accordance with the prevailing estimated optimal fractions $\hat{\boldsymbol{\alpha}}_n^* = \{\hat{\alpha}_{1,n}^*, \hat{\alpha}_{2,n}^*, \ldots, \hat{\alpha}_{r,n}^*\}$, where n represents the expended number of simulation calls. The observations are then used to update the estimated scores $\hat{\mathbb{S}}_{i,n}$

Algorithm 1 A Sequential Algorithm

```
Require: Number of pilot samples $\delta_0 > 0$; number of samples between allocation vector updates
    $\delta > 0$; and a minimum proportional allocation vector $\boldsymbol{\varepsilon} > 0$.
 1: Initialize: collect $\delta_0$ samples from each system $i \leq r$.
 2: Initialize: total simulation effort $n = r\delta_0$, effort for each system $n_i = \delta_0$.
 3: Update the objective and constraint estimators $(\hat{H}_i(n_i), \hat{\mathbf{G}}_i(n_i))$, the rate function estimator
    $\hat{I}_{i,n}(x_i, \boldsymbol{y_i})$, and the score estimator $\hat{\mathbb{S}}_{i,n}$ for all $i \leq r$.
 4: if no systems are estimated feasible then
 5:     Set $\hat{\boldsymbol{\alpha}}^*_n = (1/r, 1/r, \ldots, 1/r)$.
 6: else
 7:     Update $\hat{1}(n)$, the estimated system 1, and its allocation $\hat{\alpha}^*_{1,n}$.
 8:     Set $\hat{\alpha}^*_{i,n} = (\sum_{k=2}^{r} \hat{\mathbb{S}}_k^{-1})^{-1} \times \hat{\mathbb{S}}_i^{-1} \times (1 - \hat{\alpha}^*_{1,n})$ for all systems $i \geq 2$.
 9: end if
10: Collect one sample at each system $X_k, k = 1, 2, \ldots, \delta$, where the $X_k$'s are iid random variates
    with probability mass function $\hat{\boldsymbol{\alpha}}^*_n$ and support $\{1, 2, \ldots, r\}$. Update $n_{X_k} = n_{X_k} + 1$.
11: Set $n = n + \delta$ and update $\bar{\boldsymbol{\alpha}}_n = \{n_1/n, n_2/n, \ldots, n_r/n\}$.
12: if $\bar{\boldsymbol{\alpha}}_n > \boldsymbol{\varepsilon}$ then
13:     Set $\delta^+ = 0$.
14: else
15:     Collect one sample from each system in the set of systems receiving insufficient sample $\mathcal{I}_n$.
16:     Update $n_i = n_i + 1$ for all $i \in \mathcal{I}_n$. Let $\delta^+ = |\mathcal{I}_n|$, the cardinality of $\mathcal{I}_n$.
17: end if
18: Set $n = n + \delta^+$ and go to step 3.
```

for systems $i \geq 2$, and the estimated best solution $\hat{1}(n)$. The iterative process continues by using the updated scores to modify the estimated optimal fractions $\hat{\boldsymbol{\alpha}}^*_n$, which will in turn be used as the system choice probabilities in the subsequent stage. In the event that the score involves inverting an estimated covariance matrix, we suggest that an independent model be used until the matrix can reliably be inverted numerically.

7. NUMERICAL EXAMPLES

In this section, we conduct numerical experiments to evaluate the performance of the proposed SCORE allocations. The SCORE allocations are extremely general; therefore we make simplifying assumptions for implementation in this section. Consistent with the vast majority of the current ranking and selection literature, we provide insights in the context of SO problems with multivariate-normal simulation observations. Our hope is that the sense conveyed by these experiments will hold in a broader context.

To evaluate the SCORE and SCORE-c allocations in the multivariate normal context, we consider a number of competing allocations. The first competitor is the "MVN True" allocation, which is the asymptotically optimal allocation in the multivariate normal context, provided by Hunter [2011]. The MVN True allocation results from solving a bi-level optimization problem — the outer concave maximization problem corresponds to solving Problem Q, and at each step in solving Problem Q, Problems $R_i(\alpha_1, \alpha_i)$ must be solved for $i = 2, \ldots, r$. In the multivariate normal case, each Problem $R_i(\alpha_1, \alpha_i)$ is a quadratic program. The MVN True allocation becomes computationally burdensome for large numbers of systems, but it allows us to evaluate the optimality gap of the SCORE allocations when the number of systems is finite. The second competing allocation is the "MVN Independent" allocation, provided by Hunter and Pasupathy [2013]. This allocation assumes mutual independence between the objective and constraint estimators, and requires solving a single convex optimization problem, corresponding to Problem Q, to find the optimal allocation (the solution to each Problem $R_i(\alpha_1, \alpha_i)$ is known in this case). Finally, we compare the SCORE allocations to equal allocation. An extensive comparison of OCBA-CO [Lee et al. 2012] with the MVN Independent allocation proposed by Hunter and Pasupathy [2013] appears in Hunter and Pasupathy [2013]; hence we omit comparisons here.

In what follows, we evaluate the performance of each competing allocation as the number of systems grows across a variety of different problems. To create a flexible testbed, we fix the number of constraints at $s = 5$ and randomly generate instances of Problem P (see §2.1) as follows. To ensure system 1 and "worse"-but-feasible systems exist, we set $h_1 = 0$ and generate approximately one-third of the systems with objective and constraint values uniformly distributed in $[0, 3]$ and $[-3, 0]$, respectively. The remaining systems were created by independently generating uniformly distributed objective and constraint values in $[-3, 3]$. The covariance matrices were also randomly generated and scaled so that all variances equal one. Further, for any given instance of Problem P, the covariance matrices are equal to each other across all systems. For numerical distinction, we ensured $|h_1 - h_i| > 0.05$ and $|g_{ij} - \gamma_j| > 0.05$ for $i \leq r, j \leq 5$. We require this numerical distinction since otherwise, calculating the MVN True and MVN Independent allocations would be computationally burdensome due to shallow rate functions. Generating the systems this way ensures that, as the total number of systems increases, the systems become increasingly "dense" in the considered region.

In the following sections, we evaluate the competitors on three key metrics: (i) the time to solve for the allocation during one update of the sequential allocation, (ii) the optimality gap of the allocation from the MVN True allocation (when it can be computed), and (iii) the achieved probability of correct selection when the allocation scheme is implemented sequentially. Metrics (i) and (ii), the time to solve and the optimality gap, are evaluated in §7.1, and metric (iii), the achieved probability of correct selection upon sequential implementation, is evaluated in §7.2. In §7.3, we numerically evaluate the effect of estimating the parameters of the assumed distributional family on the performance of the sequential SCORE algorithm, and in §7.4, we numerically evaluate the robustness of the SCORE allocation to violation of our assumptions. Due to the similar performance of the SCORE and SCORE-c allocations, and given the speed of the SCORE allocation, we evaluate only SCORE in §7.3 and §7.4.

7.1. Time to Solve versus Optimality Gap

In Table II, we evaluate the proposed allocations and all competing allocations in terms of their computation time and optimality gap from the "MVN True" allocation. In Table II, all parameters of the distributional family are assumed to be known; that is, no quantities are estimated. Since some computations are intense and our goal is only to show the order of magnitude of the time required to solve for each allocation, we present results averaged over ten randomly-generated Problems P for each r in the table. The reported times in Table II give a sense of the computing effort required to perform a single update of each allocation scheme.

The primary message we wish to convey in Table II is that, with relatively negligible computing effort, when the number of systems is large, the proposed allocations exhibit performance on par with the truly optimal allocation. Further, with the SCORE allocation, we are able to solve for the nearly-optimal allocation in under a minute (on average) when the number of systems is 10,000. The time to solve for the SCORE allocation increases linearly as the number of systems increases.

In the final rows of Table II, we are unable to compute the true allocation and therefore unable to assess the optimality gap of the proposed allocations. However the differences in the rate of decay of $P\{\text{FS}\}$ between the proposed allocations and equal allocation give some sense of the improvement the SCORE allocations achieve.

Table II. For ten randomly-generated constrained SO problems with r systems and $s = 5$ constraints under the multivariate normal assumption with correlation, the table reports the average wall-clock time to solve for the optimal allocation under each specified allocation model, as well as the average rate of decay of the probability of false selection ($P\{\text{FS}\}$) and the average optimality gap from the "True" optimal allocation.

Number of Systems (r)	Metric of Interest (all rates $\times 10^{-4}$)	MVN True	MVN Indep.	SCORE-c	SCORE	Equal
20	Ave. Time to Solve	2.31 sec	0.05 sec	1.89 sec	0.12 sec	0 sec
	Ave. Rate z	222.99	218.86	154.82	155.17	25.42
	Ave. Opt. Gap of z	0 [a]	4.13	68.17	67.82	197.57
100	Ave. Time to Solve	17.45 sec	0.34 sec	8.77 sec	0.57 sec	0 sec
	Ave. Rate z	11.75	11.72	9.53	10.17	0.37
	Ave. Opt. Gap of z	0	0.03	2.23	1.58	11.39
500	Ave. Time to Solve	3.71 min	1.50 min	40.04 sec	2.72 sec	0 sec
	Ave. Rate z	1.85	1.84	1.53	1.53	0.02
	Ave. Opt. Gap of z	0	0.01	0.32	0.32	1.83
1,000	Ave. Time to Solve	53.87 min	34.76 min	1.54 min	5.39 sec	0 sec
	Ave. Rate z	1.37	1.36	0.98	1.10	0.01
	Ave. Opt. Gap of z	0	0.01	0.39	0.27	1.37
2,000	Ave. Time to Solve	> 6 hr	> 6 hr	3.10 min	11.33 sec	0 sec
	Ave. Rate z	— [b]	—	0.52	0.58	0.003
	Ave. Opt. Gap of z	0	—	—	—	—
5,000	Ave. Time to Solve	> 6 hr	> 6 hr	7.97 min	27.65 sec	0 sec
	Ave. Rate z	—	—	0.26	0.29	0.002
	Ave. Opt. Gap of z	0	—	—	—	—
10,000	Ave. Time to Solve	> 6 hr	> 6 hr	16.37 min	54.43 sec	0 sec
	Ave. Rate z	—	—	0.14	0.16	0.0008
	Ave. Opt. Gap of z	0	—	—	—	—

Note: All computing performed in MATLAB R2011a on a 1.8 GHz Intel Core i7 processor with 4GB 1333 MHz DDR3 memory.

[a] The optimality gap of the true allocation is to the precision of the solver.

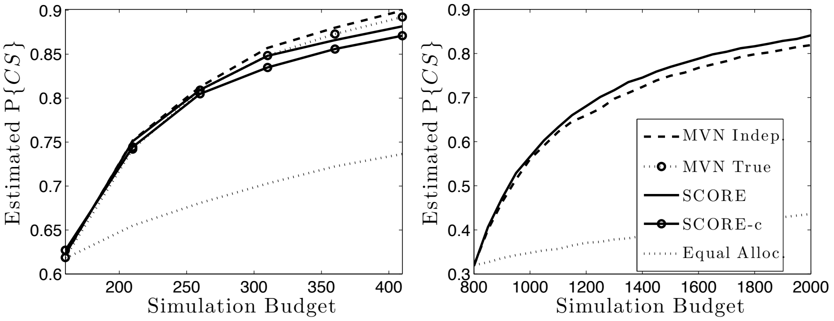
[b] The symbol '—' indicates that the data is unavailable due to the large computational time.

7.2. Finite-time Performance of the Sequential Algorithm

To assess the finite-time performance of the SCORE allocation in the context of the sequential Algorithm 1, we implemented sequential versions of each of the five competing allocations in two different scenarios: $r = 20$ systems, $s = 5$ constraints, and $r = 100$ systems, $s = 5$ constraints. For each sequential algorithm and number of systems, ten thousand randomly-generated Problems P were created and solved.

Figure 1 shows the resulting estimated probability of correct selection $P\{\text{CS}\}$ for each algorithm, where we note that the estimated $P\{\text{CS}\}$ values are correlated across the simulation budget values. The parameters for Algorithm 1 in Figure 1 are $\delta_0 = 8$, $\delta = 50$, $\varepsilon_i = 1 \times 10^{-8}$ for all $i \leq r$, and an eigenvalue tolerance of $\xi = 1 \times 10^{-5}$. The parameter δ_0 was chosen as 8 since the covariance matrix is 6-by-6. For each estimated covariance matrix, the largest eigenvalue was required to be larger than ξ for the sampling model to account for correlation. Otherwise, the corresponding independent model was used. The percent of systems requiring an independent model was less than one percent. The sequential algorithms for the MVN True and MVN Independent allocations used the same parameters as those used in Algorithm 1 for the SCORE allocations. The computation times of the sequential versions of MVN True and SCORE-c were prohibitively long for estimating $P\{\text{CS}\}$ in the case of $r = 100, s = 5$, and therefore these allocations are omitted from Figure 1(b).

In Figure 1, all allocations perform well relative to equal allocation. Among the non-equal allocations, the estimated $P\{\text{CS}\}$ values are close to each other; however it appears that the SCORE allocations do not perform as well as the MVN Independent and MVN True allocations. For a lower number of systems ($r = 20$), this result is



(a) $r = 20$ (b) $r = 100$

Fig. 1. The $P\{\text{CS}\}$ for each allocation was calculated across 10,000 runs of the sequential algorithm on different randomly-generated Problems P, each with five constraints and (a) $r = 20$ or (b) $r = 100$. Due to large computational time, (b) excludes the MVN True and SCORE-c algorithms.

expected since the approximations in the SCORE allocations rely on the existence of a large number of systems. It also appears that the independent versions of the MVN and SCORE allocations slightly outperform the corresponding allocations that account for correlation. When the number of systems is increased to $r = 100$ in Figure 1(b), the SCORE allocations outperform the MVN Independent allocations in finite time. Further research is needed to determine the problem-related boundary conditions to indicate when each allocation should be used.

7.3. Effect of Parameter Estimation and Robustness to Violation of Assumption 5

In the unconstrained R&S context, the effect of estimation of the distributional parameters on the achieved $P\{\text{CS}\}$ of OCBA was explored by Chen et al. [2006]. The results of Chen et al. [2006] seem to suggest that allocations that adapt to the sample path outperform static-but-optimal allocations. In this section, we explore similar experiments in the constrained R&S context — that is, we compare the SCORE allocation using estimated values for the optimal allocation, as outlined in Algorithm 1, with the performance of the SCORE allocation assuming that the optimal SCORE allocation α^* is magically known in advance. Since we implement only the SCORE allocation in this section, we also remove the "numerical distinction" constraint in the generation of random Problems P. That is, we allow Problems P to be generated with systems arbitrarily close to system 1 and to the constraints. All parameters of Algorithm 1 are identical to the implementation from §7.2, which renders Figure 2 directly comparable with Figure 1(b).

Figure 2 shows that, consistent with the results of Chen et al. [2006], SCORE with estimated distributional parameters outperforms SCORE with known distributional parameters in terms of the achieved $P\{\text{CS}\}$. Therefore the estimation of distributional parameters is actually *helpful* toward increasing the $P\{\text{CS}\}$. Toward understanding this effect, we make two observations. First, the SCORE allocation with *true* distributional parameters is *asymptotically* optimal. Consequently, there is no reason to expect that at any specific finite simulation budget in Figure 2, the SCORE allocation with the true distributional parameters should provide the highest probability of correct selection. Second, it is indeed surprising that the SCORE allocation implemented with *estimated* distributional parameters seems to systematically outperform the SCORE allocation with *true* distributional parameters across finite simulation budgets. The reasons behind such behavior are still unclear to us; specifically, we are unsure as to whether this observation is the result of a systematic advantage provided by estima-

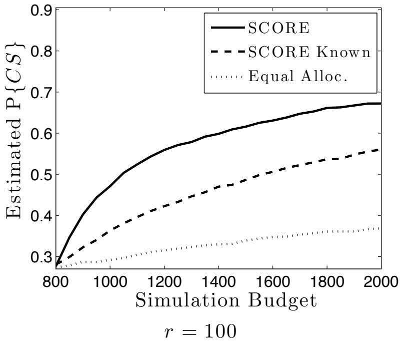


Fig. 2. The $P\{\text{CS}\}$ for each allocation was calculated across 10,000 runs of the sequential algorithm on different randomly-generated Problems P, each with five constraints and $r = 100$. This figure shows the effect of estimating distributional parameters on the achieved $P\{\text{CS}\}$.

tion or simply an artifact of the way we calculate the probability of correct selection. Given that the SCORE allocations using estimated parameters are essentially locations of sample-path maxima, we speculate that this seemingly counterintutitive effect may have connections to a well-known phenomenon in the context of sample-average approximation. Namely, under quite general conditions, the expected maximum value of sample-path functions is larger than the maximum value attained by the expected sample-path function [Mak et al. 1999].

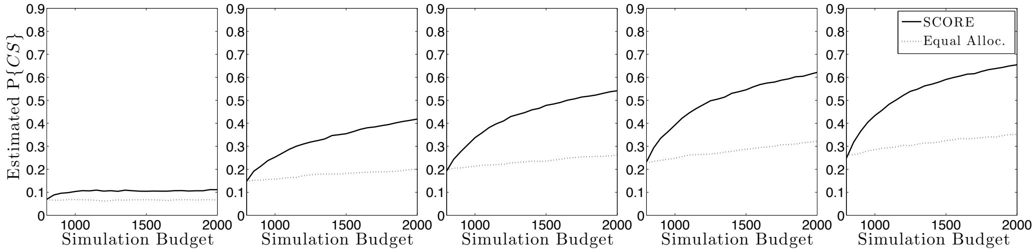
We also note that the gap in performance between the SCORE allocation and equal allocation in Figure 2 is similar to the gap shown by Figure 1(b), which implies that the relative performance of the SCORE allocation is not affected by allowing systems to be arbitrarily close to system 1 and to the constraints, despite our mathematical requirement of Assumption 5. We note that the overall $P\{\text{CS}\}$ estimates in Figure 2 are lower than those in Figure 1(b), which is intuitive because the problems solved in Figure 2 are in some sense “harder.” The SCORE allocation overcomes a severe limitation of the MVN True allocation when the number of systems is large — the need to solve a bi-level optimization problem that is increasingly difficult when systems are close to each other — without reduction in performance.

7.4. Robustness to Violation of Assumptions 5 and 7

We now evaluate the SCORE allocation in terms of its robustness to violation of Assumption 7 and the assumption that, when implementing the SCORE allocation, one might know the distributional family in advance. In this section we assume the true distribution is a multivariate t distribution, so that the distributional family is misspecified *and* the true distribution is heavy-tailed. Further, since we implement only the SCORE allocation in this section, we retain the removal of the “numerical distinction” constraint in the generation of random Problems P. Therefore we allow systems to be generated arbitrarily close to system 1 and to the constraints, so that the figures in this section are comparable with Figure 2.

In the stochastically constrained case, we do not know the optimal allocation for the multivariate t distribution. Therefore we compare the performance of the SCORE allocation with equal allocation. (In the unconstrained case, optimal allocation in the case of heavy-tailed distributions was explored in Broadie et al. [2007] and Blanchet et al. [2008].) As in §7.3, we retain all parameters of Algorithm 1 used in previous numerical examples. Figure 3 shows that SCORE is competitive when the true distribution is multivariate t. After 2,000 samples per sample path for 10,000 sample paths

(20 samples per system per sample path), SCORE selects the best system at nearly double the rate of equal allocation, as shown in Table III. These experiments indicate that SCORE is robust to violations of normality, even when systems may be arbitrarily close to system 1 and the constraints.



(a) $r = 100, df = 1$ (b) $r = 100, df = 2$ (c) $r = 100, df = 3$ (d) $r = 100, df = 5$ (e) $r = 100, df = 10$

Fig. 3. The $P\{\text{CS}\}$ for each allocation was calculated across 10,000 runs of the sequential algorithm on different randomly-generated Problems P, each with five constraints and $r = 100$. This figure shows the achieved $P\{\text{CS}\}$ when the true distributional family is a multivariate t, but the SCORE allocation is based on a multivariate normal family.

Table III. The table reports the estimated percent of the time that system 1 was correctly selected as the best after 2,000 samples in Figure 3.

	$df = 1$	$df = 2$	$df = 3$	$df = 5$	$df = 10$
Equal Allocation	6.6%	20.0%	26.2%	32.2%	35.2%
SCORE	11.1%	41.8%	54.2%	62.2%	65.4%

Note: All standard errors less than 0.5%.

8. CONCLUDING REMARKS

As we have demonstrated, the asymptotic theory in the context of stochastically constrained SO on large finite sets points to a remarkably simple "closed-form" allocation scheme that is asymptotically optimal and based on a single intuitive measure that we have called the score. During implementation, the score can be estimated progressively as simulation observations become available, leading to a scheme that is consistent with respect to the optimal allocation. Furthermore, under distributional assumptions on the random variables comprising the simulation, the proposed scheme becomes highly tractable and opens avenues for solving large-scale stochastically constrained SO problems efficiently. This is borne out by our numerical experiments where the optimal budget allocation for constrained SO problems involving thousands of systems has been approximated remarkably well by the closed-form allocation, and with computing effort that is a few orders of magnitude less than what would be required to identify the true allocation.

Do we expect our proposed scheme to perform as well in broader practical contexts? More generally, are iterative schemes based on estimated rate functions useful in practice, particularly in light of recent work by Glynn and Juneja [2011] arguing that large-deviation rate function estimators tend to be heavy-tailed in many settings, thereby increasing the possibility of large errors during implementation? We do not as yet have a conclusive response to this question, but two comments are relevant. First, we emphasize that we do not recommend estimating the rate functions directly. Instead, we recommend making distribution-family assumptions on the random variables underlying the simulation. Specifically, assume that the simulation observations fall within a convenient but flexible distributional family, thereby circumventing a difficult nonparametric analysis and allowing the parametric estimation of the rate function, akin to the limited examples discussed through Table I. This is also consistent with our broader view of only looking for a model that forms a rigorous basis for optimal budget allocation within an SO context, while expending little effort on model estimation itself. Second, we are aware of pathological examples where the heavy tails of rate function estimators have deleterious effects during implementation. We are also aware that in cases where the underlying random variables have bounded support, e.g., Bernoulli, beta, this problem does not arise. What is unclear is the extent to which heavy tails become relevant during implementation. This is a topic of ongoing research, but our numerical experience has been overwhelmingly in favor of the usefulness of the proposed scheme.

APPENDIX

A. PROOF OF LEMMA 3.1

Before we proceed with the proof of Lemma 3.1, we first solve for the KKT conditions of Problem R_i.

A.1. The KKT conditions for Problem R_i

Since Problem R_i is strictly convex and its unique solution exists [Hunter 2011], and Slater's condition holds, the KKT conditions for Problem R_i are necessary and sufficient for optimality. Recall that $\lambda_{ix} \leq 0$ and $\lambda_{ij} \leq 0, j \leq s$ are the Lagrange multipliers associated with the constraints, where $\boldsymbol{\lambda}_i = (\lambda_{ix}, \lambda_{i1}, \ldots, \lambda_{is})^T$. In addition to the primal feasibility conditions $(x_i^* - x_{1i}^*) \leq 0$ and $(\boldsymbol{y}_i^* - \boldsymbol{\gamma}) \leq \mathbf{0}$, we have complementary slackness conditions $\lambda_{ix}(x_i^* - x_{1i}^*) = 0$ and $\lambda_{ij}(y_{ij}^* - \gamma_j) = 0$ for all $j \leq s$, and the stationarity conditions

$$\alpha_1 \frac{\partial I_1(x_{1i}^*)}{\partial x_{1i}} + \lambda_{ix} = 0, \tag{23}$$

$$\alpha_i \nabla I_i(x_i^*, \boldsymbol{y}_i^*) - \boldsymbol{\lambda}_i = 0. \tag{24}$$

Case: $i \in \Gamma^*$. Then $x_{1i}^* = x_i^*$ and $\lambda_{ij} = 0$ for all $j \leq s$, which implies

$$\alpha_1 \frac{\partial I_1(x_{1i}^*)}{\partial x_{1i}} + \alpha_i \frac{\partial I_i(x_i^*, \boldsymbol{y}_i^*)}{\partial x_i} = 0 \quad \text{and} \quad \alpha_i \frac{\partial I_i(x_i^*, \boldsymbol{y}_i^*)}{\partial y_{ij}} = 0 \text{ for all } j \leq s. \tag{25}$$

Case: $i \in \mathcal{S}_b^*$. Then $\lambda_{ix} = 0$, $y_{ij}^* = \gamma_j$ for all $j \in \mathcal{C}_I^{i*}$, and $\lambda_{ij} = 0$ for all $j \in \mathcal{C}_F^{i*}$ which implies

$$\alpha_1 \frac{\partial I_1(x_{1i}^*)}{\partial x_{1i}} = \alpha_i \frac{\partial I_i(x_i^*, \boldsymbol{y}_i^*)}{\partial x_i} = 0 \quad \text{and} \quad \alpha_i \frac{\partial I_i(x_i^*, \boldsymbol{y}_i^*)}{\partial y_{ij}} = 0 \text{ for all } j \in \mathcal{C}_F^{i*}. \tag{26}$$

Case: $i \in \mathcal{S}_w^*$. Then $x_{1i}^* = x_i^*$, $y_{ij}^* = \gamma_j$ for all $j \in \mathcal{C}_I^{i*}$, and $\lambda_{ij} = 0$ for all $j \in \mathcal{C}_F^{i*}$ which implies

$$\alpha_1 \frac{\partial I_1(x_{1i}^*)}{\partial x_{1i}} + \alpha_i \frac{\partial I_i(x_i^*, \boldsymbol{y}_i^*)}{\partial x_i} = 0 \quad \text{and} \quad \alpha_i \frac{\partial I_i(x_i^*, \boldsymbol{y}_i^*)}{\partial y_{ij}} = 0 \text{ for all } j \in \mathcal{C}_F^{i*}. \tag{27}$$

A.2. Proof of Lemma 3.1

Suppose we are given the optimal value $(x_{1i}^*, x_i^*, \boldsymbol{y}_i^{*T})$ to Problem R_i, which was derived in §A.1. Then $R_i(\tilde{\alpha}_1, \tilde{\alpha}_i) = \tilde{\alpha}_1 I_1(x_{1i}^*) + \tilde{\alpha}_i I_i(x_i^*, \boldsymbol{y}_i^*)$. For $k \in \{1, i\}$, let

$$\frac{\partial(x_i^*, \boldsymbol{y}_i^*)}{\partial \tilde{\alpha}_k} = \left(\frac{\partial x_i^*(\tilde{\alpha}_1, \tilde{\alpha}_i)}{\partial \tilde{\alpha}_k}, \frac{\partial y_{i1}^*(\tilde{\alpha}_1, \tilde{\alpha}_i)}{\partial \tilde{\alpha}_k}, \ldots, \frac{\partial y_{is}^*(\tilde{\alpha}_1, \tilde{\alpha}_i)}{\partial \tilde{\alpha}_k} \right)^T = \left(\frac{\partial x_i^*}{\partial \tilde{\alpha}_k}, \frac{\partial y_{i1}^*}{\partial \tilde{\alpha}_k}, \ldots, \frac{\partial y_{is}^*}{\partial \tilde{\alpha}_k} \right)^T.$$

Then from equations (23) and (24), we find that

$$\begin{aligned}\frac{\partial R_i(\tilde{\alpha}_1, \tilde{\alpha}_i)}{\partial \tilde{\alpha}_1} &= I_1(x_{1i}^*) + \tilde{\alpha}_1 \frac{\partial I_1(x_{1i}^*)}{\partial x_{1i}^*} \frac{\partial x_{1i}^*}{\partial \tilde{\alpha}_1} + \tilde{\alpha}_i \nabla I_i(x_i^*, \boldsymbol{y}_i^*)^T \frac{\partial (x_i^*, \boldsymbol{y}_i^*)}{\partial \tilde{\alpha}_1} \\ &= I_1(x_{1i}^*) - \lambda_{ix} \frac{\partial x_{1i}^*}{\partial \tilde{\alpha}_1} + \boldsymbol{\lambda}_i^T \frac{\partial (x_i^*, \boldsymbol{y}_i^*)}{\partial \tilde{\alpha}_1}, \qquad (28) \\ \frac{\partial R_i(\tilde{\alpha}_1, \tilde{\alpha}_i)}{\partial \tilde{\alpha}_i} &= I_i(x_i^*, \boldsymbol{y}_i^*) + \tilde{\alpha}_1 \frac{\partial I_1(x_{1i}^*)}{\partial x_{1i}^*} \frac{\partial x_{1i}^*}{\partial \tilde{\alpha}_i} + \tilde{\alpha}_i \nabla I_i(x_i^*, \boldsymbol{y}_i^*)^T \frac{\partial (x_i^*, \boldsymbol{y}_i^*)}{\partial \tilde{\alpha}_i} \\ &= I_i(x_i^*, \boldsymbol{y}_i^*) - \lambda_{ix} \frac{\partial x_{1i}^*}{\partial \tilde{\alpha}_i} + \boldsymbol{\lambda}_i^T \frac{\partial (x_i^*, \boldsymbol{y}_i^*)}{\partial \tilde{\alpha}_i}. \qquad (29)\end{aligned}$$

We now use equations (25)–(29) to complete the proof of Lemma 3.1. First, note that for systems $i \in \Gamma^*$, $\lambda_{ij} = 0$ for all $j \leq s$, so that $\boldsymbol{\lambda}_i^T \frac{\partial (x_i^*, \boldsymbol{y}_i^*)}{\partial \tilde{\alpha}_1} = \lambda_{ix} \frac{\partial x_i^*}{\partial \tilde{\alpha}_1}$ and $\boldsymbol{\lambda}_i^T \frac{\partial (x_i^*, \boldsymbol{y}_i^*)}{\partial \tilde{\alpha}_i} = \lambda_{ix} \frac{\partial x_i^*}{\partial \tilde{\alpha}_i}$. However this result also holds for all $i \in \mathcal{S}_b^* \cup \mathcal{S}_w^*$ since in both cases, $\lambda_{ij} = 0$ for all $j \in \mathcal{C}_F^{i*}$ while $y_{ij}^* = \gamma_j$ for all $j \in \mathcal{C}_I^{i*}$, which implies $\frac{\partial y_{ij}^*}{\partial \tilde{\alpha}_1} = \frac{\partial y_{ij}^*}{\partial \tilde{\alpha}_i} = 0$ for all $j \in \mathcal{C}_I^{i*}$. Therefore it generally holds that

$$\frac{\partial R_i(\tilde{\alpha}_1, \tilde{\alpha}_i)}{\partial \tilde{\alpha}_1} = I_1(x_{1i}^*) - \lambda_{ix} \frac{\partial x_{1i}^*}{\partial \tilde{\alpha}_1} + \lambda_{ix} \frac{\partial x_i^*}{\partial \tilde{\alpha}_1}$$

and

$$\frac{\partial R_i(\tilde{\alpha}_1, \tilde{\alpha}_i)}{\partial \tilde{\alpha}_i} = I_i(x_i^*, \boldsymbol{y}_i^*) - \lambda_{ix} \frac{\partial x_{1i}^*}{\partial \tilde{\alpha}_i} + \lambda_{ix} \frac{\partial x_i^*}{\partial \tilde{\alpha}_i}.$$

Then for all $i \in \Gamma^* \cup \mathcal{S}_w^*$, since $x_{1i}^* = x_i^*$, then $\frac{\partial x_{1i}^*}{\partial \tilde{\alpha}_1} = \frac{\partial x_i^*}{\partial \tilde{\alpha}_1}$ and $\frac{\partial x_{1i}^*}{\partial \tilde{\alpha}_i} = \frac{\partial x_i^*}{\partial \tilde{\alpha}_i}$, and the result follows. For all $i \in \mathcal{S}_b^*$, the result follows immediately upon noting that $\lambda_{ix} = 0$.

B. PROOF OF LEMMA 3.2

The result follows by noting that all terms in the sum in equation (5) corresponding to the set $\mathcal{S}_b^*$ are equal to zero, shown as follows. From equation (26) in the proof of Lemma 3.1, for all systems $i \in \mathcal{S}_b^*$, it holds that $\tilde{\alpha}_1 \frac{\partial I_1(x_{1i}^*)}{\partial x_{1i}} = 0$. Since we know $\tilde{\alpha}_1^* > 0$, it follows that $\frac{\partial I_1(x_{1i}^*)}{\partial x_{1i}} = 0$ which implies $x_{1i}^* = h_1$. Therefore $I_1(x_{1i}^*) = I_1(h_1) = 0$. Under Assumptions 3 and 4, $I_i(x_i^*, \boldsymbol{y}_i^*) < \infty$, and it remains only to show that $I_i(x_i^*, \boldsymbol{y}_i^*) > 0$ for all $i \in \mathcal{S}_b^*$. Consider that for systems $i \in \mathcal{S}_b^*$, we have

$$R_i(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*) = \inf_{x_i \leq h_1, \boldsymbol{y}_i \leq \gamma} \tilde{\alpha}_i^* I_i(x_i, \boldsymbol{y}_i).$$

Since $\tilde{\alpha}_i^* > 0$, if $R_i(\tilde{\alpha}_1^*, \tilde{\alpha}_i^*) = 0$, then $I_i(x_i^*, \boldsymbol{y}_i^*) = I_i(h_i, \boldsymbol{g}_i)$, which is a contradiction since $\boldsymbol{g}_i \neq \boldsymbol{\gamma}$ and $\boldsymbol{y}_i^* = \boldsymbol{\gamma}$.

REFERENCES

S. Andradóttir. 2006. Simulation Optimization. Wiley, 307–333.

S. Andradóttir and S.-H. Kim. 2010. Fully Sequential Procedures for Comparing Constrained Systems via Simulation. *Naval Research Logistics* 57, 5 (2010), 403–421.

J. April, J. Glover, J. Kelly, and M. Laguna. 2001. Simulation optimization using "real-world" applications. In *Proceedings of the 2001 Winter Simulation Conference*, B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer (Eds.). Institute of Electrical and Electronics Engineers: Piscataway, New Jersey, 134–138.

R. R. Barton and M. Meckesheimer. 2006. Metamodel-based simulation optimization. In *Simulation*, S. G. Henderson and B. L. Nelson (Eds.). Elsevier, 535–574.

D. Batur and S.-H. Kim. 2010. Finding Feasible Systems in the Presence of Constraints on Multiple Performance Measures. *ACM Transactions on Modeling and Computer Simulation* 20, 3, Article 13 (2010), 26 pages.

J. Blanchet, J. Liu, and B. Zwart. 2008. Large Deviations Perspective on Ordinal Optimization of Heavy-Tailed Systems. In *Proc. of the 2008 Winter Simulation Conference*, S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler (Eds.). Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ, 489–494.

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

J. Branke, S. E. Chick, and C. Schmidt. 2007. Selecting a Selection Procedure. *Management Science* 53, 12 (2007), 1916–1932.

M. Broadie, M. Han, and A. Zeevi. 2007. Implications of heavy tails on simulation-based ordinal optimization. In *Proc. of the 2007 Winter Simulation Conference*, S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton (Eds.). Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ, 439–447.

C.-H. Chen, D. He, and M. Fu. 2006. Efficient Dynamic Simulation Allocation in Ordinal Optimization. *IEEE Trans. Automat. Control* 51, 12 (2006), 2005–2009.

C.-H. Chen, 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 (2000), 251–270.

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

M. C. Fu, F. W. Glover, and J. April. 2005. Simulation optimization: a review, new developments, and applications. In *Proc. of the 2005 Winter Simulation Conference*, M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines (Eds.). Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ, 83–95.

P. W. Glynn and S. Juneja. 2004. A large deviations perspective on ordinal optimization. In *Proc. of the 2004 Winter Simulation Conference*, R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters (Eds.). Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ, 577–585.

P. W. Glynn and S. Juneja. 2011. Ordinal Optimization: A Nonparametric Framework. In *Proc. of the 2011 Winter Simulation Conference*, S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu (Eds.). Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ.

S. R. Hunter. 2011. *Sampling laws for stochastically constrained simulation optimization on finite sets*. Ph.D. Dissertation. Virginia Polytechnic Institute and State University.

S. R. Hunter and R. Pasupathy. 2013. Optimal sampling laws for stochastically constrained simulation optimization on finite sets. *INFORMS Journal on Computing* 25, 3 (2013), 527–542. DOI:http://dx.doi.org/10.1287/ijoc.1120.0519

S.-H. Kim and B. L. Nelson. 2006. Selecting the best system. In *Simulation*, S. G. Henderson and B. L. Nelson (Eds.). Elsevier, 501–534.

L. H. Lee, N. A. Pujowidianto, L.-W. Li, C.-H. Chen, and C. M. Yap. 2012. Approximate Simulation Budget Allocation for Selecting the Best Design in the Presence of Stochastic Constraints. *IEEE Trans. Automat. Control* 57, 11 (2012), 2940–2945.

W.-K. Mak, D. P. Morton, and R. K. Wood. 1999. Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters* 24 (1999), 47–56.

R. Pasupathy and S. G. Henderson. 2006. A Testbed of Simulation-Optimization Problems. In *Proc. of the 2006 Winter Simulation Conference*, L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto (Eds.). Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ.

R. Pasupathy and S. G. Henderson. 2011. SimOpt: A Library of Simulation Optimization Problems. In *Proc. of the 2011 Winter Simulation Conference*, S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu (Eds.). Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ.

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

R. Szechtman and E. Yücesan. 2008. A New Perspective on Feasibility Determination. In *Proc. of the 2008 Winter Simulation Conference*, S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler (Eds.). Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ, 273–280.

Received July 2013; revised January 2014; accepted May 2014