Efficient Simulation Budget Allocation for Selecting an Optimal Subset[1]

Chun-Hung Chen[2] and Donghai He
Department of Systems Engineering & Operations Research
George Mason University
4400 University Drive, MS 4A6
Fairfax, VA 22030

Michael Fu
Robert H. Smith School of Business
and Institute for Systems Research
University of Maryland
College Park, MD 20742-1871

Abstract

We consider a variation of the subset selection problem in ranking and selection, where motivated by recently developed global optimization approaches applied to simulation optimization, our objective is to identify the top-*m* out of *k* designs based on simulated output. Using the optimal computing budget framework, we formulate the problem as that of maximizing the probability of correctly selecting all of the top-*m* designs subject to a constraint on the total number of samples available. For an approximation of this correct selection probability, we derive an asymptotically optimal allocation procedure that is easy to implement. Numerical experiments indicate that the resulting allocations are superior to other methods in the literature, and the relative efficiency increases for larger problems.

[1] This work has been supported in part by the National Science Council of the Republic of China under Grant NSC 95-2811-E-002-009, by NSF under Grants IIS-0325074 and DMI-0323220, by NASA Ames Research Center under Grants NAG-2-1643 and NNA05CV26G, by FAA under Grant 00-G-016, and by AFOSR under Grant FA95500410210.

[2] **Corresponding author:** Professor Chun-Hung Chen, Tel: 703-993-3572; Fax: 703-993-1521; Email: cchen9@gmu.edu; Web: mason.gmu.edu/~cchen9

1. Introduction

We consider the problem of selecting the top *m* out of *k* designs, where the performance of each design is estimated with noise (uncertainty). The primary context is simulation, where the goal is to determine the best allocation of simulation replications among the various designs in order to maximize the probability of selecting all top-*m* designs. This problem setting falls under the well-established branch of statistics known as ranking and selection or multiple comparison procedures (cf. Bechhofer, Santner, and Goldsman 1995). In the context of simulation, Goldsman and Nelson (1998) provide an overview of this field; see also Andradottir et al. (2005).

The primary motivation for the setting considered in this paper comes from some recent developments in global optimization that, when applied to the simulation setting, require the selection of an "elite" subset of good candidate solutions in each iteration of the algorithm. Examples of these include genetic algorithms (Holland 1975, Chambers 1995), the cross entropy method (CE, see Rubinstein and Kroese 2004), the model reference adaptive search method (MRAS, cf. Hu, Fu, and Marcus 2006ab), and more generally, evolutionary population-based algorithms that require the selection of an "elite" population in the evolutionary process (see Fu, Hu, and Marcus 2006). Instead of trying to find a subset that contains the *single best* among a currently generated set of candidate solutions, the objective is to find an *optimal subset* such that *all* members are among the best performers in that candidate set. The reason for this requirement is that this entire subset is used to update the subsequent population or sampling distribution that drives the search for additional candidates. A subset with poor performing solutions will result in an update that leads the search in a possibly misleading direction. The overall efficiency of these types of simulation optimization algorithms depends on how efficiently we simulate the candidates and correctly select the top-*m* designs. The algorithm developed herein is generic enough that it can be integrated with any such simulation-based evolutionary optimization search algorithms.

Most of the ranking-and-selection research has focused on identifying the best design. Typical of these are two-stage or sequential procedures that ultimately return a single choice as the estimated optimum, e.g., Dudewicz and Dalal (1975) and Rinott (1978). Even the traditional "subset selection" procedures aim at identifying a subset that *contains* the best design, dating back to Gupta (1965), who presented a single-stage procedure for producing a subset (of random size) containing the best design with a specified probability. Extensions of this work relevant to the simulation setting include Sullivan and Wilson (1989), who derive a two-stage subset selection procedure that determines a subset of maximum size *m* that, with a specified probability, contains designs that are all within a pre-specified amount of the optimum. This indifference zone procedure approach also results in a subset of random size, and the designs are assumed to follow a normal distribution, with independence between designs assumed and unknown and unequal moments. The primary motivation for such procedures is *screening*, whereby the selected subset can be scrutinized further to find the single optimum. This is in contrast to the motivation for our setting, as alluded to earlier. More recently, these procedures have also been incorporated into simulation optimization, but in a different manner, where the ranking-and-selection procedure is incorporated in order to be able to make statistically valid inferences rather than driving the actual optimization process itself; see, e.g., Buchholz and Thümmler (2005), Boesel, Nelson, and Kim (2003), and Nelson et al. (2001), who also consider the setting of unknown and unequal variances; see the references therein for the cases of known or unknown but equal variances. Swisher, Jacobson, and Yücesan (2003) includes a discussion of subset selection in the context of simulation optimization along this vein. Note that these approaches are still focused on selecting a subset containing the single best. As a result, the selected subset may also contain very poor solutions, which can affect the convergence rate of procedures such as MRAS and the CE method when applied to the simulation optimization setting, where the use of the selection procedures are in the *iterative* updating steps and not in the final determination of the optimum.

To reiterate, instead of selecting the very best design from a given set or finding a subset that is highly likely to contain the best design, the objective in this papers is to find *all* top-*m* designs. About the only substantive work we are aware of addressing this problem is Koenig and Law (1985), who along the lines of the procedure in Dudewicz and Dalal (1975), develop a two-stage procedure for selecting all the *m* best designs (see also Law & Kelton 2000 for an overview of the procedure). The number of additional simulation replications for the second stage is computed based on a least favorable configuration, resulting in very conservative allocations, so that the required computational cost is much higher than actually needed.

To improve the efficiency of allocating simulation replications among competing designs, Chen et al. (1997, 2000), Chen and Kelton (2000), Chick and Inoue (2001ab), Hyden and Schruben (2000), Lee and Chew (2003), Trailovic and Pao (2004), and Fu et al. (2006) have approached the ranking-and-selection problem from the perspective of allocating a fixed number of simulation replications in order to maximize the probability of correct selection, under a framework called "optimal computing budget allocation." Intuitively, to ensure a high probability of correct selection, a larger portion of the computing budget should be allocated to those designs that are critical in the process of identifying the best design. In terms of traditional ranking and selection, for example, this results in the use of both the means and variances in the allocation procedures (for normally distributed design performances), rather than just the variances, as in Dudewciz and Dalal (1975) and Rinott (1978). However, all of this work has focused on selecting the single best, and there has been no research involving subset selection. This paper aims to fill this gap by providing an efficient allocation procedure for selecting the *m* best designs. Note that among the selected *m* designs, there is no further ranking done within the set. Again, this is consistent with the requirements of the CE method and MRAS approach, as well as other evolutionary population-based methods that require an "elite" population of some type.

The paper is organized as follows. In the next section, we formulate the optimal computing budget allocation problem for selecting the top-*m* designs. Section 3 derives an allocation scheme based on approximating the correction selection probability and then carrying out an asymptotic analysis. The performance of the technique is illustrated with a series of numerical examples in Section 4. Section 5 concludes the paper.

2. Problem Statement

We introduce the following notation:

T = total number of simulation replications (budget),

k = total number of designs,

m = number of top designs to be selected in the optimal subset,

S_m = set of m (distinct) indices indicating designs in selected subset,

N_i = number of simulation replications allocated to design i,

X_{ij} = j-th simulation replication for design i,

$\bar{J}_i$ = $\frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}$, sample mean for design i,

J_i = mean for design i,

σ_i^2 = variance for design i,

$\delta_{i,j}$ = $\bar{J}_i$ - $\bar{J}_j$.

The objective is to find a simulation budget allocation that maximizes the probability of selecting the *optimal subset,* defined as the set of m ($< k$) best designs, for m a fixed number. Our approach is developed based on Bayesian setting (e.g., Inoue and Chick 1998). The mean of the simulation output for each design, J_i, is assumed unknown and treated as a random variable, whose posterior distribution is updated as simulation proceeds. Without loss of generality, we will take as the m best designs those designs with the m smallest means (but this is unknown), so that in terms of our notation, the correct selection event is defined by S_m containing all of the m smallest mean designs:

$$\mathrm{CS}_m \equiv \{ \bigcap_{i \in S_m} \bigcap_{j \notin S_m} (J_i \le J_j) \} = \{ \max_{i \in S_m} J_i \le \min_{i \notin S_m} J_i \}. \tag{1}$$

The optimal computing budget allocation (OCBA) problem is given by

$$\max_{N_1,\cdots,N_k} P\{\mathrm{CS}_m\}$$

$$\text{s.t. } N_1 + N_2 + \cdots + N_k = T. \tag{2}$$

Here $N_1 + N_2 + \cdots + N_k$ denotes the total computational cost assuming the simulation execution times for different designs are roughly the same. This formulation implicitly assumes that the computational cost of each replication is constant across designs. The simulation budget allocation problems given in Chen et al. (2000) is actually a special case of (2) with $m = 1$. For notational simplification, we will drop the "*m*" in $P\{\mathrm{CS}_m\}$ in the remaining discussion.

Note that rank order within the subset is not part of the objective. In this paper, we will take S_m to be the m designs with the smallest *sample* means. Let $\bar{J}_{i_r}$ be the *r*-th smallest (order statistic) of $\{\bar{J}_1, \bar{J}_2, ..., \bar{J}_k\}$, i.e., $\bar{J}_{i_1} \leq \bar{J}_{i_2} \leq ... \leq \bar{J}_{i_k}$. Then, the selected subset is given by

$$S_m \equiv \{ i_1, i_2, ..., i_m \}.$$

We assume that the simulation output samples $\{X_{ij}\}$ are normally distributed and independent from replication to replication, i.e., $X_{i1}, X_{i2},..., X_{iN_i}$, are i.i.d. $\mathrm{N}(J_i, \sigma_i^2)$, as well as independent across designs. The normality assumption is typically satisfied in simulation, because the output is obtained from an average performance or batch means, so that Central Limit Theorem effects usually hold.

3. Approximate Asymptotically Optimal Allocation Scheme

To solve the OCBA problem (2), we estimate $P\{\mathrm{CS}\}$ using the Bayesian model presented in Chen et al. (2000) and He et al. (2006). After the simulation is performed, a posterior distribution for the unknown mean J_i, $p(J_i \mid X_{ij}, j=1,...,N_i)$, is constructed based on two pieces of information: (i) prior knowledge of the system's performance, and (ii) current simulation output. Thus, in the Bayesian framework, the probability of correct selection defined by (1) is given by

$$P\{\mathrm{CS}\} = P\{\tilde{J}_i \leq \tilde{J}_j,\ i \in S_m \text{ and } j \notin S_m\}, \tag{3}$$

where $\tilde{J}_i$, $i=1,...,k$, denotes the random variable whose probability distribution is the posterior distribution of design *i*. As in Chen et al. (2000), we assume that the unknown mean J_i has a conjugate normal prior distribution and consider non-informative prior distributions, which implies that no prior knowledge is available about the performance of any design before conducting the simulations, in which case the posterior distribution of J_i is (cf. DeGroot 1970)

$$\tilde{J}_i \sim N(\bar{J}_i, \frac{\sigma_i^2}{N_i}).$$

After the simulation is performed, $\bar{J}_i$ can be calculated, σ_i^2 can be approximated by the sample variance, and the $P\{\mathrm{CS}\}$ given by Equation (3) can then be estimated using Monte Carlo simulation. However, since estimating $P\{\mathrm{CS}\}$ via Monte Carlo simulation is time-consuming and the purpose of budget allocation is to improve simulation efficiency, we adopt an approximation of $P\{\mathrm{CS}\}$ using a lower bound.

3.1 Approximating the Probability of Correct Selection

For a constant *c*,

$$P\{\text{CS}\} = P\{\, \tilde{J}_i \leq \tilde{J}_j \,,\, i \in S_m \text{ and } j \notin S_m\}$$

$$\geq P\{\, \tilde{J}_i \leq c \text{ and } \tilde{J}_j \geq c,\ \ i \in S_m \text{ and } j \notin S_m\}$$

$$= \prod_{i \in S_m} P\{\tilde{J}_i \leq c\} \prod_{i \notin S_m} P\{\tilde{J}_i \geq c\} \equiv APCSm, \tag{4}$$

where the last line is due to independence across designs. We refer to this lower bound for $P\{\text{CS}\}$, which can be computed easily and eliminates the need for extra Monte Carlo simulation, as the *Approximate Probability of Correct Selection for m best* (*APCSm*). Determining an appropriate value for *c* will be deferred to later this section. Using the approximation given by Equation (4), the OCBA problem (2) becomes

$$\max_{N_1,\cdots,N_k} \prod_{i \in S_m} P\{\tilde{J}_i \leq c\} \prod_{i \notin S_m} P\{\tilde{J}_i \geq c\}$$

$$\text{s.t. } N_1 + N_2 + \cdots + N_k = T. \tag{5}$$

Now we solve OCBA problem (5), assuming the variables $\{N_i\}$ are continuous.

3.2 Asymptotically Optimal Solution

For notation simplification, we define the variable $\delta_i = \bar{J}_i$ - c, i=1,2,…,k.

For $i \in S_m$,

$$P(\tilde{J}_i \leq c) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\left(\sigma_i / \sqrt{N_i}\right)} e^{-\frac{(x-\delta_i)^2}{2\left(\sigma_i^2 / N_i\right)}} dx$$

$$= \int_{\frac{\delta_i}{\left(\sigma_i / \sqrt{N_i}\right)}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt,$$

and for $i \notin S_m$,

$$P(\widetilde{J}_i \geq c) = \int_0^{\infty} \frac{1}{\sqrt{2\pi}\,(\sigma_i / \sqrt{N_i})}\, e^{-\frac{(x-\delta_i)^2}{2(\sigma_i^2/N_i)}}\, dx$$

$$= \int_{\frac{-\delta_i}{(\sigma_i/\sqrt{N_i})}}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{t^2}{2}}\, dt.$$

Now let F be the Lagrangian relaxation of (5), with Lagrange multiplier λ:

$$F = \prod_{i \in S_m} P\{\widetilde{J}_i \leq c\} \cdot \prod_{i \notin S_m} P\{\widetilde{J}_i \geq c\} - \lambda\left(\sum_{i=1}^{k} N_i - T\right)$$

$$= \prod_{i \in S_m} \int_{\frac{\delta_i}{(\sigma_i/\sqrt{N_i})}}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{t^2}{2}}\, dt \cdot \prod_{i \notin S_m} \int_{\frac{-\delta_i}{(\sigma_i/\sqrt{N_i})}}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{t^2}{2}}\, dt - \lambda\left(\sum_{i=1}^{k} N_i - T\right).$$

Furthermore, the Karush-Kuhn-Tucker (KKT) (Walker 1999) conditions of this problem can be stated as follows.

For $i \in S_m$,

$$\frac{\partial F}{\partial N_i} = \prod_{\substack{j \in S_m \\ j \neq i}} P\{\widetilde{J}_j \leq c\} \cdot \prod_{j \notin S_m} P\{\widetilde{J}_j \geq c\} \cdot \frac{1}{2\sqrt{2\pi}}\, e^{-\frac{\delta_i^2}{2(\sigma_i^2/N_i)}}\, \frac{\delta_i}{\sigma_i}\, N_i^{-1/2} - \lambda = 0. \qquad (6)$$

For $i \notin S_m$,

$$\frac{\partial F}{\partial N_i} = \prod_{j \in S_m} P\{\widetilde{J}_j \leq c\} \cdot \prod_{\substack{j \notin S_m \\ j \neq i}} P\{\widetilde{J}_j \geq c\} \cdot \frac{-1}{2\sqrt{2\pi}}\, e^{-\frac{\delta_i^2}{2(\sigma_i^2/N_i)}}\, \frac{\delta_i}{\sigma_i}\, N_i^{-1/2} - \lambda = 0. \qquad (7)$$

Also, $\frac{\partial F}{\partial \lambda} = 0$ returns the budget constraint $\sum_{i=1}^{k} N_i - T = 0$.

To examine the relationship between N_i and N_j for $i \neq j$, we consider three cases:

(1) $i \in S_m$, and $j \notin S_m$:

Equating the expressions in Equations (6) and (7),

$$\prod_{\substack{r \in S_m \\ r \neq i}} P\{\widetilde{J}_r \leq c\} \cdot \prod_{r \notin S_m} P\{\widetilde{J}_r \geq c\} \cdot \frac{1}{2\sqrt{2\pi}}\, e^{-\frac{\delta_i^2}{2(\sigma_i^2/N_i)}}\, \frac{\delta_i}{\sigma_i}\, N_i^{-1/2} - \lambda$$

$$= \prod_{r\in S_m} P\{\widetilde{J}_r \le c\} \cdot \prod_{\substack{r\notin S_m \\ r\neq j}} P\{\widetilde{J}_r \ge c\} \cdot \frac{-1}{2\sqrt{2\pi}}\, e^{-\frac{\delta_j^2}{2(\sigma_j^2/N_j)}}\, \frac{\delta_j}{\sigma_j}\, N_j^{-1/2} - \lambda.$$

Simplifying,

$$P\{\widetilde{J}_j \ge c\} \cdot e^{-\frac{\delta_i^2}{2(\sigma_i^2/N_i)}}\, \frac{\delta_i}{\sigma_i}\, N_i^{-1/2} = P\{\widetilde{J}_i \le c\} \cdot e^{-\frac{\delta_j^2}{2(\sigma_j^2/N_j)}}\, \frac{-\delta_j}{\sigma_j}\, N_j^{-1/2}.$$

Taking the log on both sides,

$$\log(P\{\widetilde{J}_j \ge c\}) - \frac{\delta_i^2 N_i}{2\sigma_i^2} + \log(\frac{\delta_i}{\sigma_i}) - \frac{1}{2}\log(N_i) = \log(P\{\widetilde{J}_i \le c\}) - \frac{\delta_j^2 N_j}{2\sigma_j^2} + \log(\frac{-\delta_j}{\sigma_j}) - \frac{1}{2}\log(N_j). \quad (8)$$

Now, we consider the asymptotic limit $T \to \infty$ with $N_i = \alpha_i T$, $\sum_{i=1}^{k} \alpha_i = 1$. Substituting for N_i, Equation (8) becomes

$$\log(P\{\widetilde{J}_j \ge c\}) - \frac{\delta_i^2 \alpha_i}{2\sigma_i^2} T + \log(\frac{\delta_i}{\sigma_i}) - \frac{1}{2}\log(\alpha_i T)$$

$$= \log(P\{\widetilde{J}_i \le c\}) - \frac{\delta_j^2 \alpha_j}{2\sigma_j^2} T + \log(\frac{-\delta_j}{\sigma_j}) - \frac{1}{2}\log(\alpha_j T).$$

Dividing by T,

$$\frac{1}{T}\log(P\{\widetilde{J}_j \ge c\}) - \frac{\delta_i^2}{2\sigma_i^2}\alpha_i + \frac{1}{T}\log(\frac{\delta_i}{\sigma_i}) - \frac{1}{2T}\log(\alpha_i T)$$

$$= \frac{1}{T}\log(P\{\widetilde{J}_i \le c\}) - \frac{\delta_j^2}{2\sigma_j^2}\alpha_j + \frac{1}{T}\log(\frac{-\delta_j}{\sigma_j}) - \frac{1}{2T}\log(\alpha_j T).$$

and then taking $T \to \infty$ yields

$$\frac{\delta_i^2}{\sigma_i^2}\alpha_i = \frac{\delta_j^2}{\sigma_j^2}\alpha_j.$$

Therefore, we obtain the ratio between α_i and α_j or between N_i and N_j as:

$$\frac{N_i}{N_j} = \frac{\alpha_i}{\alpha_j} = \left(\frac{\sigma_i/\delta_i}{\sigma_j/\delta_j}\right)^2 \text{ for } i \in S_m \text{, and } j \notin S_m. \quad (9)$$

(2) Both $i, j \in S_m$ and $i \neq j$.:

From Equation (6),

$$\frac{\partial F}{\partial N_i} = \frac{\partial F}{\partial N_j} = 0 \text{ yields}$$

$$\prod_{\substack{r\in S_m\\ r\neq i}} P\{\tilde{J}_r \le c\}\cdot \prod_{r\notin S_m} P\{\tilde{J}_r \ge c\}\cdot \frac{1}{2\sqrt{2\pi}}\, e^{-\frac{\delta_i^2}{2(\sigma_i^2/N_i)}}\, \frac{\delta_i}{(\sigma_i^2/N_i)}\, \frac{\sigma_i}{N_i^{3/2}} - \lambda$$

$$= \prod_{\substack{r\in S_m\\ r\neq j}} P\{\tilde{J}_r \le c\}\cdot \prod_{r\notin S_m} P\{\tilde{J}_r \ge c\}\cdot \frac{1}{2\sqrt{2\pi}}\, e^{-\frac{\delta_j^2}{2(\sigma_j^2/N_j)}}\, \frac{\delta_j}{(\sigma_j^2/N_j)}\, \frac{\sigma_j}{N_j^{3/2}} - \lambda.$$

Then,

$$P\{\tilde{J}_j \le c\}\cdot e^{-\frac{\delta_i^2}{2(\sigma_i^2/N_i)}}\, \frac{\delta_i}{\sigma_i}\, N_i^{-1/2} = P\{\tilde{J}_i \le c\}\cdot e^{-\frac{\delta_j^2}{2(\sigma_j^2/N_j)}}\, \frac{\delta_j}{\sigma_j}\, N_j^{-1/2}.$$

Following the analogous derivation that led to Equation (9) yields the same result

$$\frac{N_i}{N_j} = \frac{\alpha_i}{\alpha_j} = \left(\frac{\sigma_i/\delta_i}{\sigma_j/\delta_j}\right)^2 \text{ for } i, j \in S_m \text{ and } i \neq j. \quad (10)$$

(3) $i, j \notin S_m$, and $i \neq j$:

Again, following the same derivation procedures that led to Equations (9) and (10) yields

$$\frac{N_i}{N_j} = \frac{\alpha_i}{\alpha_j} = \left(\frac{\sigma_i/\delta_i}{\sigma_j/\delta_j}\right)^2 \text{ for } i, j \notin S_m \text{ and } i \neq j. \quad (11)$$

Thus, since Equations (9), (10), and (11) are identical, we write

$$\frac{N_i}{N_j} = \frac{\alpha_i}{\alpha_j} = \left(\frac{\sigma_i/\delta_i}{\sigma_j/\delta_j}\right)^2, \; i, j \in \{1, 2, ..., k\}, \text{ and } i \neq j. \quad (12)$$

In conclusion, if a solution satisfies Equation (12), then the KKT sufficient conditions must hold asymptotically, so that the corresponding solution is a locally optimal solution to the Lagrangian relaxation of the OCBA problem (5). We therefore have the following result.

Theorem 1. The allocation given by (12) is asymptotically (as $T \to \infty$) a locally optimal solution for OCBA problem (5), where $\delta_i = \bar{J}_i - c$, for c a constant, and the variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$ are finite, i.e., *APCSm* is asymptotically maximized by the allocation given by (12).

3.3 Determination of *c* Value

The parameter c impacts the quality of the approximation *APCSm* to $P\{\text{CS}\}$. Since *APCSm* is a lower bound of $P\{\text{CS}\}$, choosing c to make *APCSm* as large as possible is likely to provide a better approximation of *APCSm* to $P\{\text{CS}\}$. Figure 1 is provided to help explain our choice of c, by giving an example of probability density functions for $\tilde{J}_i$, $i = 1, 2, ..., k$.

Note that *APCSm* is a product of $P\{\tilde{J}_i \leq c\}$ for $i \in S_m$ and $P\{\tilde{J}_i \geq c\}$ for $i \notin S_m$. Consider the case Var($\tilde{J}_{i_1}$) = Var($\tilde{J}_{i_2}$) = ... = Var($\tilde{J}_{i_k}$). For any c, $P\{\tilde{J}_{i_1} \leq c\} > P\{\tilde{J}_{i_2} \leq c\} > ... > P\{\tilde{J}_{i_m} \leq c\}$, and $P\{\tilde{J}_{i_{m+1}} \geq c\} < P\{\tilde{J}_{i_{m+2}} \geq c\} < ... < P\{\tilde{J}_{i_k} \geq c\}$. To prevent *APCSm* from being small, we want to choose c to avoid any of the product terms being too small, particularly for $P\{\tilde{J}_{i_m} \leq c\}$ and $P\{\tilde{J}_{i_{m+1}} \geq c\}$, because one of these two terms is the smallest one in the product no matter what c is. A good choice of c is a number between $\bar{J}_{i_m}$ and $\bar{J}_{i_{m+1}}$, because

(i) if $c = c' < \bar{J}_{i_m}$, then $P\{\tilde{J}_{i_m} < c'\} < 0.5$. The smaller ($c' - \bar{J}_{i_m}$), the smaller $P\{\tilde{J}_{i_m} < c\}$, resulting in a negative impact on *APCSm*;

(ii) if $c = c'' > \bar{J}_{i_{m+1}}$, then $P\{\tilde{J}_{i_{m+1}} > c''\}$ becomes small and so does *APCSm*.

With these considerations, one would like to maximize both ($c - \bar{J}_{i_m}$) and ($\bar{J}_{i_{m+1}} - c$), or to maximize both $P\{\tilde{J}_{i_m} \leq c\}$ and $P\{\tilde{J}_{i_{m+1}} \geq c\}$. We choose $c = (\bar{J}_{i_m} + \bar{J}_{i_{m+1}}) / 2$, which in theory maximizes $\{(c - \bar{J}_{i_m})^2 + (\bar{J}_{i_{m+1}} - c)^2\}$, and in numerical testing results in good performance while requiring negligible computation cost.

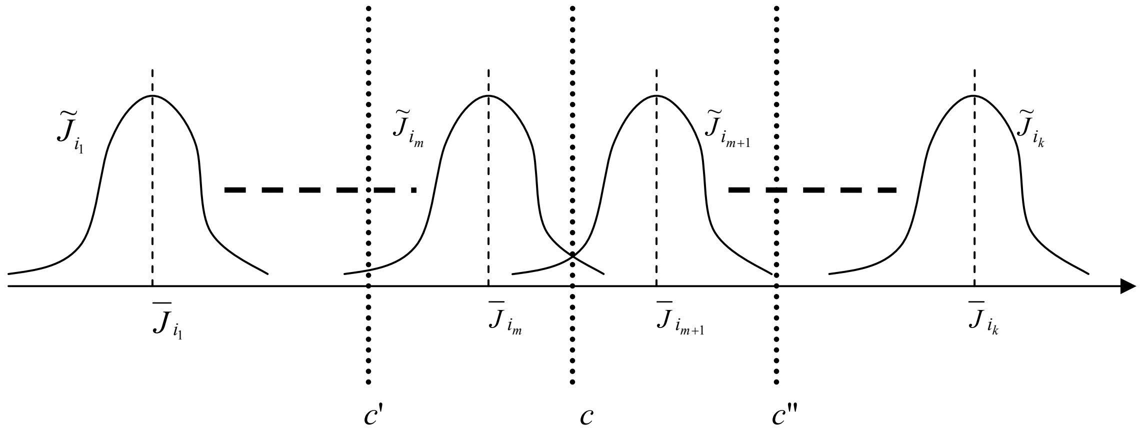

Figure 1. An example of probability density functions for $\tilde{J}_i$, $i = 1, 2,..., k$, $c' < \bar{J}_{i_m} < c < \bar{J}_{i_{m+1}} < c''$.

3.4 Sequential Allocation Scheme

The allocation given by (12) assumes known variances. In practice, a sequential algorithm is used to estimate these quantities using the updated sample variances. Furthermore, the "constant" c and sample means are also updated during each iteration. Each design is initially simulated with n_0 replications in the first stage, and additional replications are allocated incrementally with Δ replications to be allocated in each iteration. In summary, we have the following algorithm (assuming T-kn_0 is a multiple of Δ).

OCBA-m Allocation Procedure

```
INITIALIZE    $l \leftarrow 0$;
              Perform $n_0$ simulation replications for all designs; $N_1^l = N_2^l = \cdots = N_k^l = n_0$.

LOOP          WHILE $\sum_{i=1}^{k} N_i^l < T$ DO

    UPDATE    Calculate sample means and sample variance using the new simulation output;
              Compute $c = (\bar{J}_{i_m} + \bar{J}_{i_{m+1}}) / 2$.

    ALLOCATE  Increase the computing budget by $\Delta$ and calculate the new budget allocation,
              $N_1^{l+1}, N_2^{l+1}, \ldots, N_k^{l+1}$, according to (12).

    SIMULATE  Perform additional max( $N_i^{l+1}$ - $N_i^l$ ,0) simulations for design $i$, $i = 1,\ldots,k$;
              $l \leftarrow l+1$.

END OF LOOP
```

4. Numerical Testing and Comparison with Other Allocation Procedures

In this section, we test the OCBA-m algorithm by comparing it on several numerical experiments with different allocation procedures: Equal Allocation, which simulates all design alternatives equally; the Koenig and Law (1985) procedure denoted by KL; Proportional To Variance (PTV), which is a modification of KL that allocates replications proportional to the estimated variances; and the OCBA allocation algorithm for selecting only the best design (Chen et al. 2000), denoted by OCBA-1. For notational simplicity, we assume $J_{[1]} < J_{[2]} < \ldots < J_{[k]}$, so design [1] is the best and correct selection would be $S_m = \{[1], [2], ..., [m]\}$ (but this is unknown a priori).

4.1 Computing Budget Allocation Procedures

Equal Allocation

The simulation budget is allocated equally to all designs, i.e., $N_i = T/k$ for each i. The performance of equal allocation will serve as a benchmark for comparison.

KL (Koenig and Law 1985)

The two-stage procedure of Koenig and Law (1985) selects a subset of specified size m, with probability at least P^*, so that the selected subset is exactly the actual subset with the best (smallest) expected values, provided that $J_{[m+1]} - J_{[m]}$ is no less than an indifference zone, d. As in our setting, the ordering within the selected subset does not matter.

In the first stage, all designs are simulated for n_0 samples. Based on the sample variance estimate (S_i^2) obtained from the first stage and given the minimum correct selection probability P^*, the number of additional simulation samples for each design in the second stage is determined by:

$$N_i = \max(n_0+1, \lceil h_3^2 S_i^2(n_0) / d^2 \rceil), \text{ for } i = 1, 2, \ldots, k, \qquad (13)$$

where $\lceil \bullet \rceil$ is the integer "round-up" function, and h_3 is a constant that depends on k, P^*, and n_0.

Proportional To Variance (PTV)

This is a sequential modified version of the KL procedure, based on the observation that (13) implies that N_i is proportional to the estimated sample variances S_i^2. Thus, the PTV procedure sequentially determines $\{N_i\}$ based on the newly updated sample variances by replacing the **ALLOCATE** step in the OCBA-m algorithm by

$$N_i^{l+1} = C^l S_i^2, \qquad \text{for } i = 1, 2, \ldots, k,$$

where C^l is determined based on the total simulation samples equals to the given computing budget at iteration *l*. Note that the indifference-zone parameter has been removed in this modification in order to make it comparable to the other procedures.

OCBA-1 (Chen et al. 2000)

The sequential OCBA procedure of Chen et al. (2000) allocates the computing budget with the objective of selecting the best design, i.e., $m = 1$, for which extensive numerical testing has demonstrated its efficiency (e.g., Branke et al. 2006). While it is not designed for $m > 1$, we test this procedure here for benchmarking purposes, and denote it be OCBA-1.

4.2 Numerical Results

To compare the performance of the procedures, we carried out numerical experiments for several typical selection problems. In comparing the procedures, the measurement of effectiveness used is the $P\{\text{CS}\}$ estimated by the fraction of times the procedure successfully finds *all* the true *m*-best designs out of 100,000 independent experiments. Because this penalizes incorrect selections equally – e.g., a subset containing the top-1, top-2, ..., and top-(*m*-1) designs and missing only the top-*m* design is treated no differently than a subset containing not a single of the top-m designs – in our numerical experiments, we also include a second measure of selection quality, the so-called expected opportunity cost E[OC], where

$$\text{OC} \equiv \sum_{j=1}^{m} (J_{i_j} - J_{[j]}) .$$

This measure penalizes particularly bad choices more than mildly bad choices. For example, when $m = 3$, a selection of {top-1, top-2, top-4} is better than {top-1, top-2, top-5}, and both are better than {top-1, top-3, top-5}. Note that OC returns a minimum value of 0 when all the top-*m* designs area correctly selected. The estimated E[OC] is the average of OC estimate over the 100,000 independent experiments.

Each of the procedures simulates each of the *k* designs for n_0= 20 replications initially (following recommendations in Koenig and Law 1985 and Law and Kelton 2000). KL allocates additional replications in a second stage (so the total number is not fixed a priori), whereas the other procedures allocate replications incrementally by Δ=50 each time until the total budget, *T*, is consumed. For each level of computing budget, we estimate the achieved $P\{\text{CS}\}$ and E[OC].

Since KL is a two-stage indifference-zone procedure, we must specify the values for the desired probability of correct selection, *P**, and the indifference zone *d* to satisfy the condition that $J_{[m+1]}$ - $J_{[m]} \geq d$, where a smaller *d* implies a higher required computation cost based on Equation (13). In practice, the value of $J_{[m+1]}$ or $J_{[m]}$ is unknown beforehand, but for benchmarking purposes, we set $d = J_{[m+1]}$ - $J_{[m]}$, which leads to the *minimum* computational requirement (or maximum efficiency) for the procedure. As is done for the other procedures, the resulting $P\{\text{CS}\}$ and E[OC] can be estimated over the 100,000 independent experiments. Since the required computation cost also varies from one experiment to another, we will indicate the average number of total replications based on the 100,000 independent experiments.

Example 1. Equal variance

There are 10 alternative designs, with distribution $N(i, 6^2)$ for design $i = 1, 2, \ldots, 10$. The goal is to identify the top-3 designs via simulation samples, i.e., m=3 in this example.

To characterize the performance of different procedures as a function of T, we vary T between 200 and 8000 for all of the procedures other than KL, and the estimated achieved P\{CS\} and E[OC] as a function of T is shown in Figure 2. For KL, we test two cases $P^* = 0.9$ and $P^* = 0.95$, and the corresponding estimated P\{CS\} and E[OC] vs. the average total simulation replications are shown as two single points (the triangle and circle) in Figure 2.

We see that all procedures obtain a higher P\{CS\} and lower E[OC] as the available computing budget increases. However, OCBA-m achieves the highest P\{CS\} and lowest E[OC] for the same amount of computing budget. It is interesting to observe that OCBA-1, which performs significantly better than Equal Allocation and PTV when the objective is to find the single best design, fares worse in this example than these two allocations when the objective is changed to finding all the top-3 designs. Equal allocation performs almost identically to PTV, which makes sense, since the variance is constant across designs. Specifically, the computation costs to attain P\{CS\} = 0.95 for OCBA-m, OCBA-1, Equal, and PTV are 800, 3200, 1950, 2000, respectively.

Not surprisingly, the performance of KL is along the performance curve of PTV, since KL basically allocate the computing budget based on designs' variance. However, KL achieves a substantially higher P\{CS\} than the desired level (e.g., exceeding 0.99 for the target minimum of P* = 0.9) by spending a much higher computing budget than actually needed, consistent with the fact that typical two-stage indifference-zone procedures are conservative.

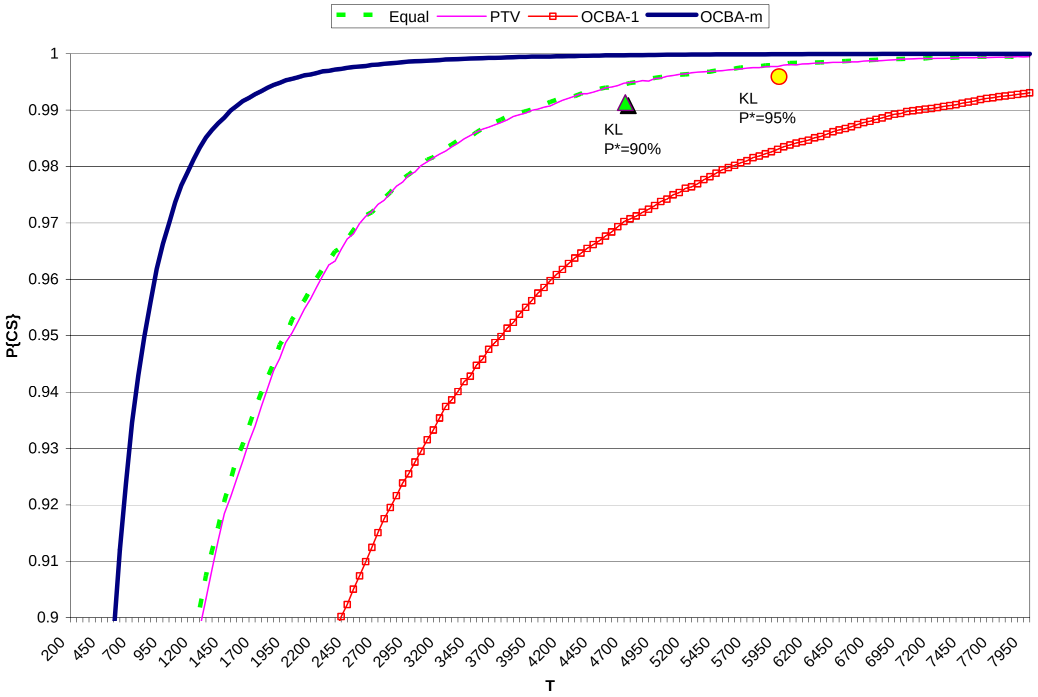

Figure 2a. $P\{\text{CS}\}$ vs. T using four sequential allocation procedures and KL (triangle for P^*=90% and circle for P^*=95%) for Example 1.

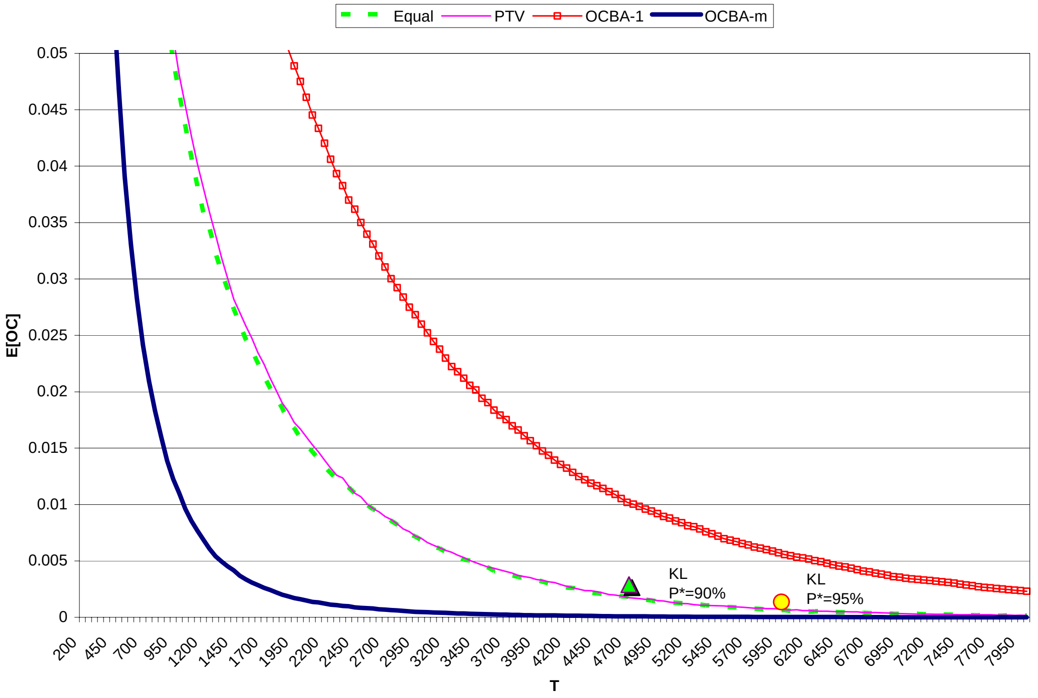

Figure 2b. E[OC] vs. T using four sequential allocation procedures and KL (triangle for P^*=90% and circle for P^*=95%) for Example 1.

Example 2. Variance increasing in value of mean

This is a variant of Example 1. All settings are preserved except that the variance is increasing in the design index, so good designs have smaller variances. Specifically, the designs are distributed $N(i\ ,\ i^2)$ for design i = 1, 2, …, 10. Again, m = 3.

The test results shown in Figure 3 are qualitatively similar to those in Example 1. OCBA-m achieves the highest P\{CS\} for the same amount of computing budget. However, PTV (and KL) performs poorly in this example because good designs receive relatively less computing budget due to their smaller variances, which tend to slow down the process of distinguishing good designs. Specifically, the computation costs to attain P\{CS\} = 0.95 for OCBA-m, OCBA-1, Equal, and PTV are 350, 750, 700, 2250, respectively.

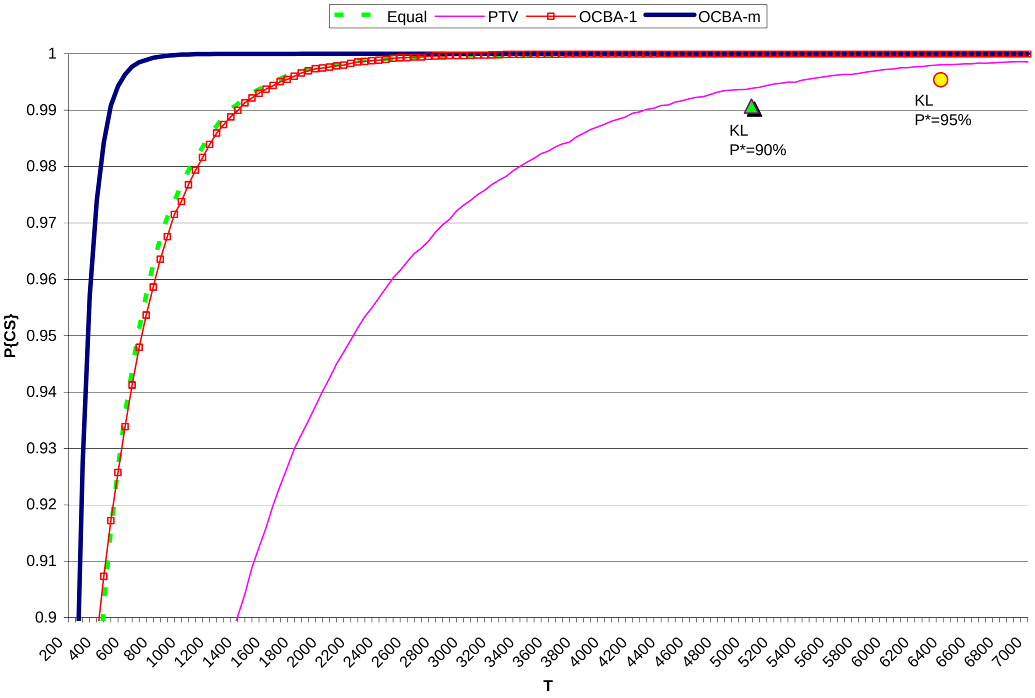

Figure 3a. P\{CS\} vs. T using four sequential allocation procedures and KL (triangle for P^*=90% and circle for P^*=95%) for Example 2.

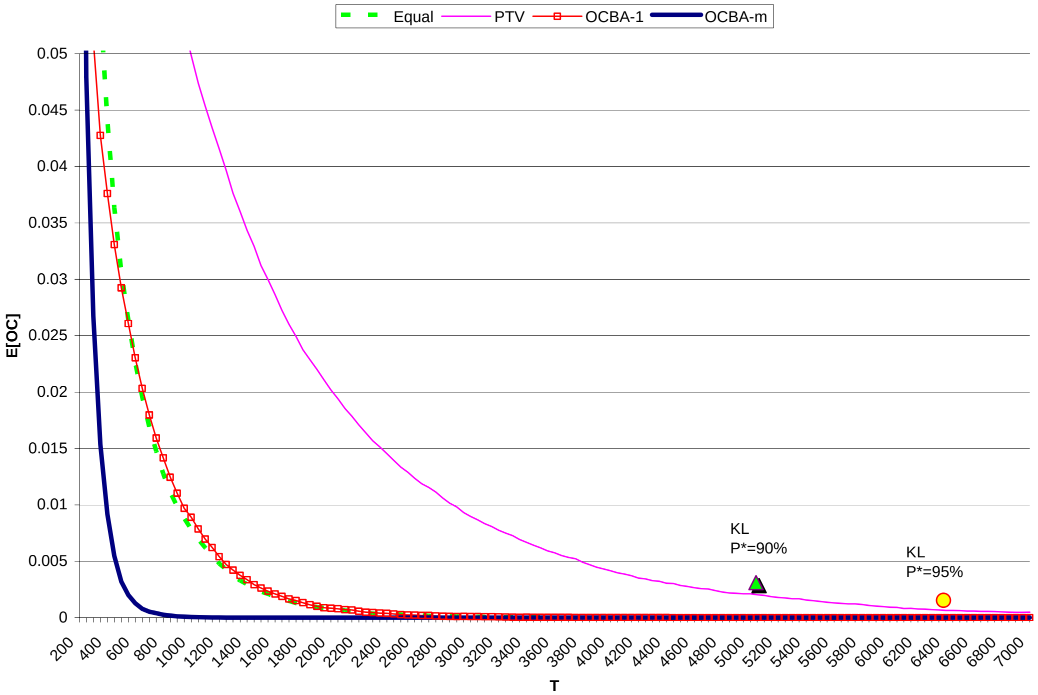

Figure 3b. E[OC] vs. *T* using four sequential allocation procedures and KL (triangle for *P**=90% and circle for *P**=95%) for example 2.

Example 3. Variance decreasing in value of mean

The third example is another variant of Examples 1 and 2, but this time the variance is decreasing in the design index, i.e., the distribution is $N(i\ ,\ (11\text{-}i)^2)$ for design i = 1, 2, …, 10. Under this setting, good designs have larger variance. Again, $m = 3$.

The test results shown in Figure 4 are similar to those in the previous examples, with again OCBA-m performing the best. However, in contrast to Example 2, PTV (and KL) performs relatively well in this example, because good designs receive much more computing budget due to their higher variances. On the other hand, OCBA-1 performs poorly, because it spends an excess amount of the computing budget to distinguish between the very top designs, since its objective is to find the best. In this example, the computation costs to attain $P\{\text{CS}\}$ = 0.95 for OCBA-m, OCBA-1, Equal, and PTV are 1400, 7900, 3050, 2200, respectively.

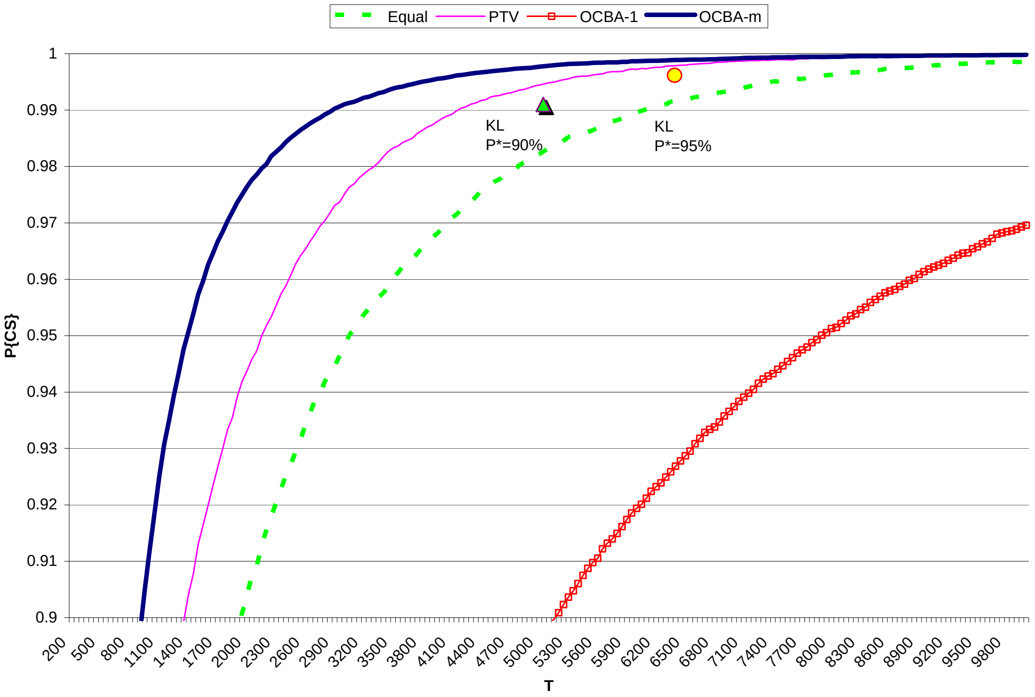

Figure 4a. $P\{\mathrm{CS}\}$ vs. T using four sequential allocation procedures and KL (triangle for P^*=90% and circle for P^*=95%) for Example 3.

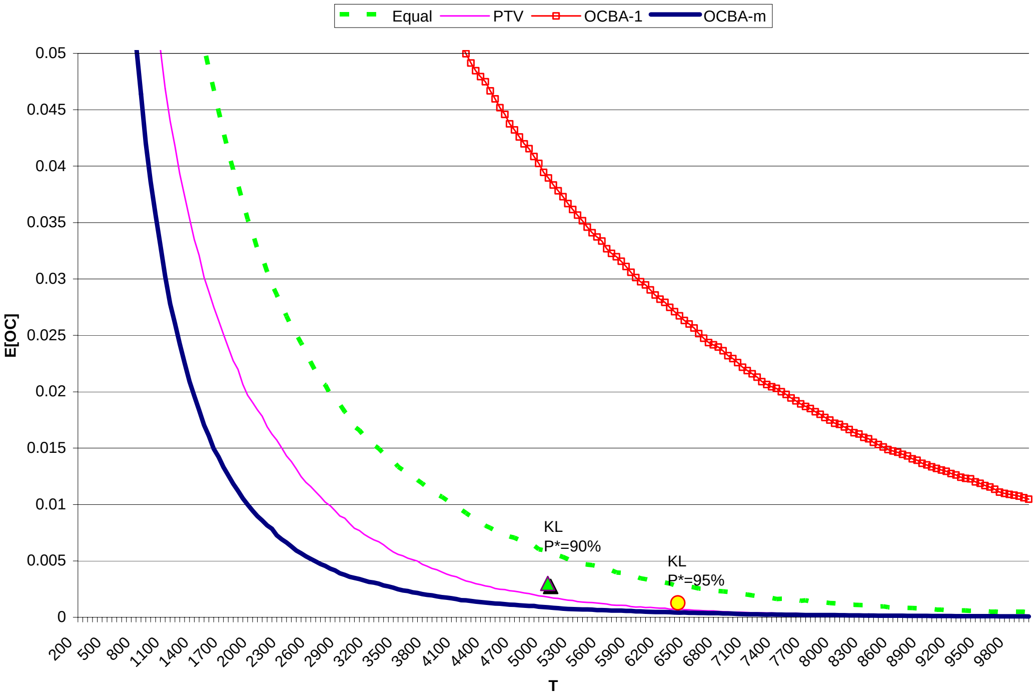

Figure 4b. E[OC] vs. T using four sequential allocation procedures and KL (triangle for P^*=90% and circle for P^*=95%) for Example 3.

Example 4. (s,S) inventory problem

The fourth example is a typical (s, S) inventory policy problem based on the example that was introduced by Koenig and Law (1985) and later analyzed by Nelson and Matejcik (1995). We extend the size of that original example from 5 designs to 10 designs. When random demand brings the inventory of system i on hand down to s_i units, the inventory is reordered to level S_i., for i = 1, 2, …, 10. The 10 systems are defined by the parameters $(s_1, s_2, \ldots, s_{10}) = (20, 20, 20, 40, 40, 40, 60, 60, 60, 80)$ and $(S_1, S_2, \ldots, S_{10}) = (30, 40, 50, 50, 60, 70, 70, 80, 90, 90)$, respectively. The systems with policy (s_3, S_3), (s_6, S_6) and (s_2, S_2) are the top-3 designs (m = 3).

The test results shown in Figure 5 are similar to those in previous examples, in that OCBA-m is clearly the top performer again; however, this time OCBA-1 is the runner up by a slight margin. The computation costs to attain $P\{\text{CS}\} = 0.95$ for OCBA-m, OCBA-1, Equal, and PTV are 500, 1200, 1650, 1350, respectively.

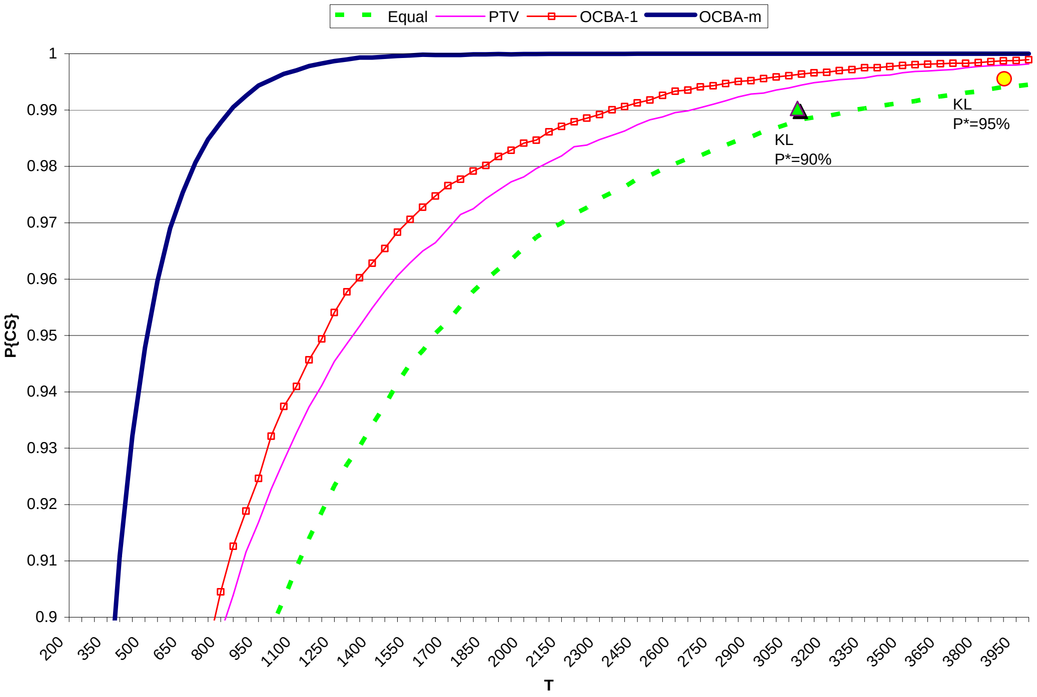

Figure 5a. $P\{\text{CS}\}$ vs. T using four sequential allocation procedures and KL (triangle for P^*=90% and circle for P^*=95%) for Example 4.

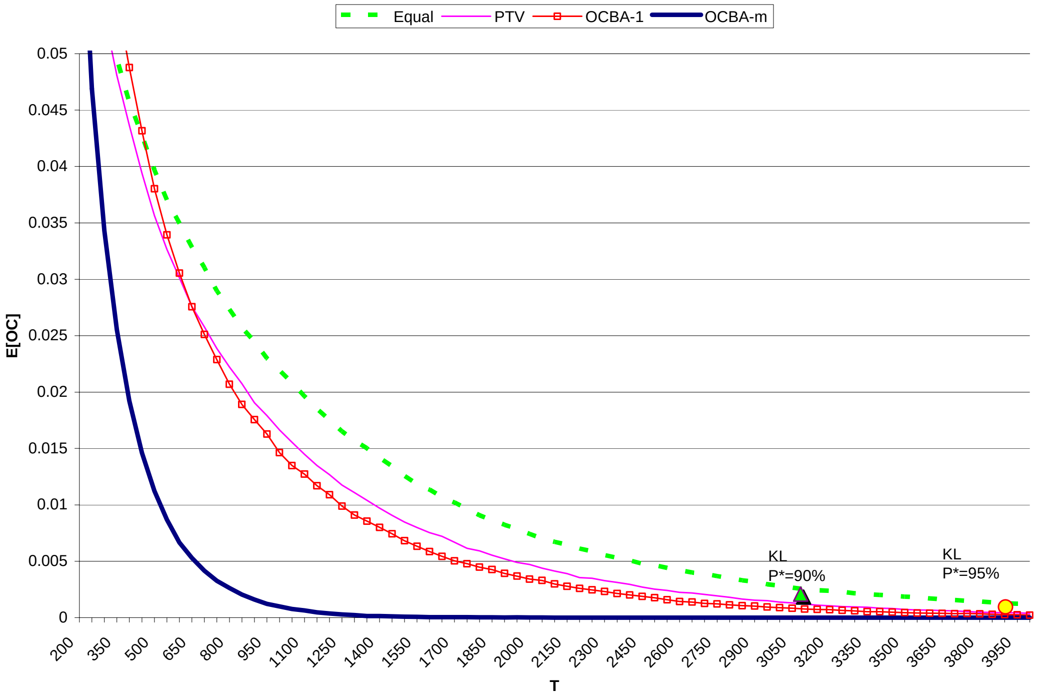

Figure 5b. E[OC] vs. T using four sequential allocation procedures and KL (triangle for P^*=90% and circle for P^*=95%) for Example 4.

Example 5. Larger-scale problem

This is a variant of Example 1 (constant variance), with the number of designs increased to 50. The alternatives have distribution $N(i\ ,\ 10^2)$ for design $i = 1, 2, \ldots, 50$, and $m = 5$. Since KL's performance basically follows that of PTV, but its required computing budget is far beyond the range we are considering here, we exclude KL from the numerical testing.

Figure 6 depicts the simulation results. As in earlier examples, OCBA-m achieves the highest $P\{\text{CS}\}$ and the lowest E[OC] with the same amount of computing budget; however, the performance gap between OCBA-m and other procedures is substantially greater. This is because a larger design space allows the OCBA-m algorithm more flexibility in allocating the computing budget, resulting in even better performance. On the other hand, OCBA-1 performs poorly, because it spends a lot of computing budget on distinguishing the very top designs, a tendency that is penalized even more for larger m. Again, since the variance is constant across designs, the performance of Equal and PTV are nearly indistinguishable. In this example, the computation costs to attain $P\{\text{CS}\}$ = 0.95 for OCBA-m, OCBA-1, Equal, and PTV are 4050, 31050, 27050, 27200, respectively.

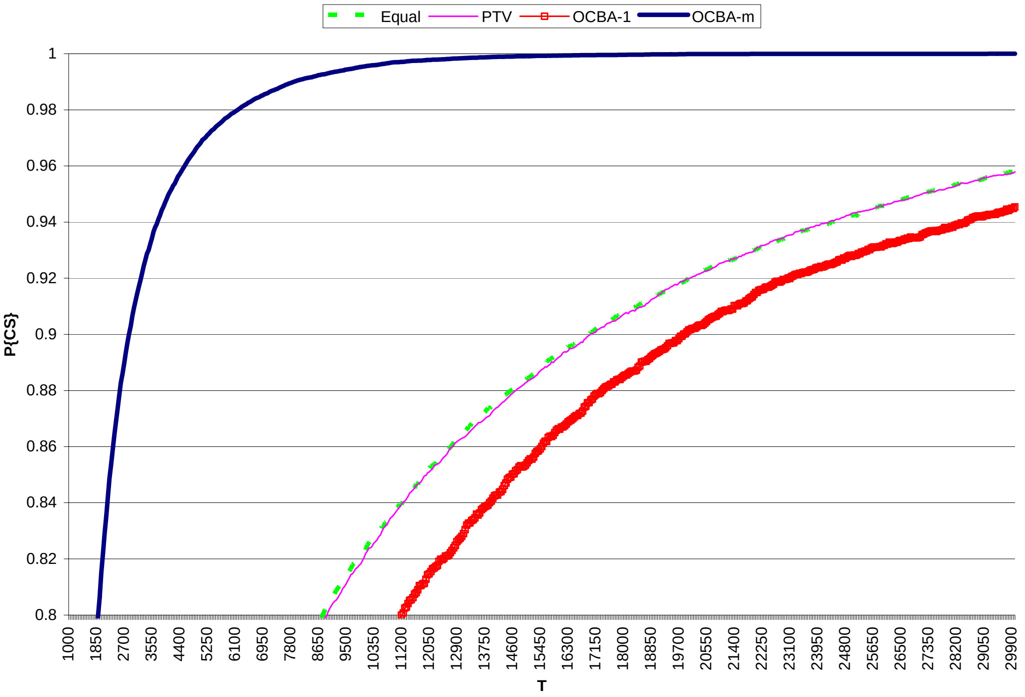

Figure 6a. $P\{\text{CS}\}$ vs. T using four sequential allocation procedures for Example 5.

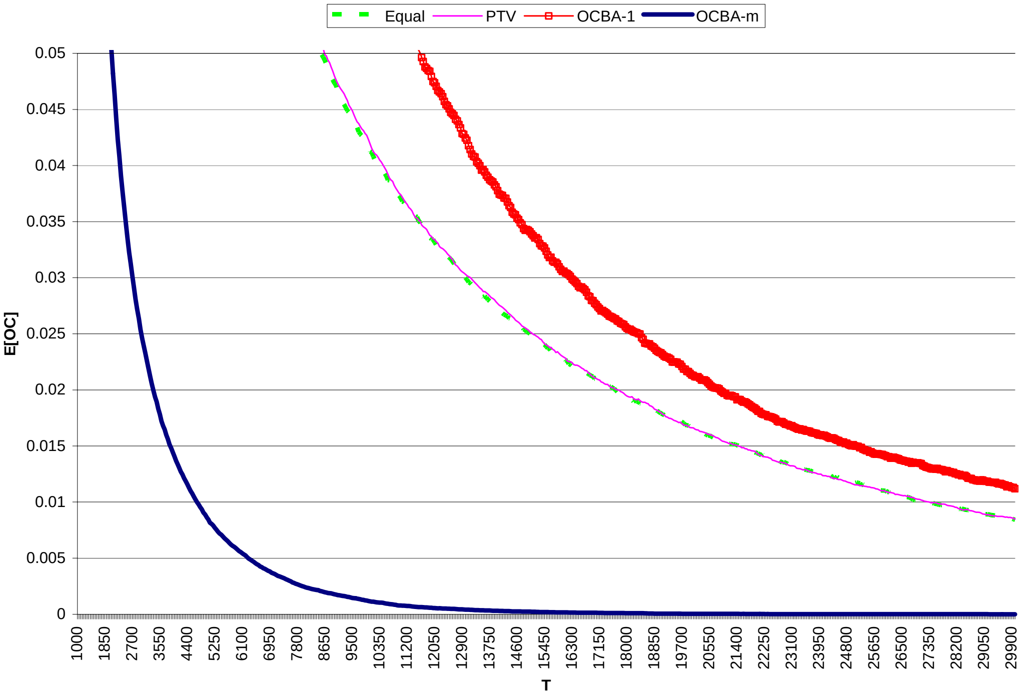

Figure 6b. E[OC] vs. T using four sequential allocation procedures for Example 5.

5. Conclusions

We present an efficient allocation procedure for identifying the top-*m* designs out of *k* (simulated) competing designs. The objective is to maximize the simulation efficiency, expressed as the probability of correct selection within a given computing budget. We propose a heuristic to approximate the associated correct selection probability, and then derive an asymptotically optimal allocation procedure for this approximate probability. Numerical testing indicates that the allocation procedure is significantly more efficient and robust than other methods in the literature, with the relative efficiency increasing in problem size. Furthermore, although the procedure was derived based on an asymptotic derivation, the numerical results indicate that the procedure is effective even when the computing budget is small. Finally, the numerical results illustrate that the allocation specified by the original OCBA algorithm (Chen et al. 2000), designed for selecting the single best design, does not perform well in selecting the top-*m* designs, providing another motivation for the need of a new methodology when the objective is extended beyond selecting just the best design.

References

1. Andradottir, S., D. Goldsman, B. W. Schmeiser, L. W. Schruben, and E. Yücesan, "Analysis Methodology: Are We Done?" *Proceedings of the 2005 Winter Simulation Conference,* pp. 790-796, December 2005.
2. Bechhofer, R.E., T.J. Santner, and D.M. Goldsman, *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*, John Wiley & Sons, 1995.
3. Boesel, J., B.L. Nelson, and S.H. Kim, "Using Ranking and Selection to 'Clean up' After Simulation Optimization," *Operations Research* 51, 814-825, 2003.
4. Branke, J., S. E. Chick, and C. Schmidt, "Selecting a Selection Procedure," submitted to *Management Science*, 2006.
5. Buchholz, P. and A. Thümmler, "Enhancing Evolutionary Algorithms with Statistical Selection Procedures for Simulation Optimization," *Proceedings of the Winter Simulation Conference*, 842-852, 2005.
6. Chambers, L., *Practical Handbook of Genetic Algorithms*, CRC Press, 1995.
7. Chen, C. H., J. Lin, E. Yücesan, and S. E. Chick, "Simulation Budget Allocation for Further Enhancing the Efficiency of Ordinal Optimization," *Journal of Discrete Event Dynamic Systems: Theory and Applications*, Vol. 10, pp. 251-270, 2000.
8. Chen, E. J. and W. D. Kelton, "An Enhanced Two-Stage Selection Procedure," *Proceedings of the Winter Simulation Conference*, pp. 727-735, 2000.
9. Chen, H. C., C. H. Chen, L. Dai, and E. Yucesan, "New Development of Optimal Computing Budget Allocation For Discrete Event Simulation," *Proceedings of the 1997 Winter Simulation Conference,* pp. 334-341, December 1997.

10. Chick, S. and K. Inoue, “New Two-Stage and Sequential Procedures for Selecting the Best Simulated System,” *Operations Research*, Vol. 49, pp. 1609–1624, 2001.

11. Chick, S. and K. Inoue, “New Procedures to Select the Best Simulated System Using Common Random Numbers,” *Management Science*, 47(8), pp. 1133-1149, 2001.

12. DeGroot, M. H., *Optimal Statistical Decisions*. McGraw-Hill, Inc., 1970.

13. Dudewicz, E. J. and S. R. Dalal, “Allocation of Observations in Ranking and Selection with Unequal Variances,” *Sankhya*, B37, pp. 28-78, 1975.

14. Fu, M. C., J. Q. Hu, C. H. Chen, and X. Xiong, “Simulation Allocation for Determining the Best Design in the Presence of Correlated Sampling,” *INFORMS Journal on Computing*, accepted for publication, 2006.

15. Fu, M. C., J. Hu, and S. I. Marcus, “Model-Based Randomized Methods for Global Optimization,” *Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems*, 355-363, 2006.

16. Goldsman, D. and B. L. Nelson, “Comparing Systems via Simulation,” J. Banks, ed. *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, John Wiley & Sons, New York. pp. 273-306, 1998.

17. Gupta, S. S., “On Some Multiple Decision (Selection and Ranking) Rules,” Technometrics 7: 225–245, 1965.

18. He, D., S. E. Chick, C. H. Chen, “The Opportunity Cost and OCBA Selection Procedures in Ordinal Optimization,” to appear in *IEEE Transactions on Systems, Man, and Cybernetics--Part C*, 2006.

19. Holland, J. H., *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, 1975.

20. Hu, J., M. C. Fu, and S. I. Marcus, “A Model Reference Adaptive Search Algorithm for Global Optimization,” *Operations Research*, accepted for publication, 2006a.

21. Hu, J., M. C. Fu, and S. I. Marcus. “A Model Reference Adaptive Search Algorithm for Stochastic Global Optimization,” working paper, 2006b.

22. Hyden, P. and L. Schruben, “Improved Decision Processes Through Simultaneous Simulation and Time Dilation,” *Proceedings of the 2000 Winter Simulation Conference*, pp. 743-748, 2000.

23. Inoue, K., and S. E. Chick, "Comparison of Bayesian and Frequentist Assessments of Uncertainty for Selecting the Best System," *Proceedings of the 1998 Winter Simulation Conference*, pp. 727-734. December, 1998.

24. Koenig, L. W. and A. M. Law, “A Procedure for Selecting a Subset of Size *m* Containing the *l* Best of *k* Independent Normal Populations,” *Communication in Statistics - Simulation and Communication*, B14, pp. 719–734, 1985.

25. Kim, S.-H. and Nelson, B.L. 2006. Selecting the best system. Chapter 18 in *Handbooks in Operations Research and Management Science: Simulation*, S.G. Henderson and B.L. Nelson, eds., Elsevier.

26. Law, A. M. and W. D. Kelton, *Simulation Modeling & Analysis*. McGraw-Hill, Inc., 2000.

27. Lee, L. H. and E. P. Chew, “A Simulation Study on Sampling and Selecting under Fixed Computing Budget,” *Proceedings of 2003 Winter Simulation Conference*, pp. 535-542, December 2003.

28. Nelson, B. L. and F. J. Matejcik, “Using Common Random Numbers for Indifference-Zone Selection and Multiple Comparisons in Simulation,” *Management Science*, 41, pp. 1935-1945, 1995.

29. Rubinstein, R.Y. and D.P. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning.* Springer, 2004.

30. Sullivan, D. W. and J. R. Wilson, “Restricted Subset Selection Procedures for Simulation,” *Operations Research*, 37:52–71, 1989.

31. Swisher, J.R., S.H. Jacobson, and E. Yücesan, “Discrete-Event Simulation Optimization Using Ranking, Selection, and Multiple Comparison Procedures: A Survey,” *ACM Transactions on Modeling and Computer Simulation* 13, 134-154, 2003.

32. Walker, R. C, *Introduction to Mathematical Programming*, Prentice Hall, Upper Saddle River, NJ, 1999

33. Trailovic, L. and L. Y. Pao, “Computing Budget Allocation for Efficient Ranking and Selection of Variances with Application to Target Tracking Algorithms,” to appear in *IEEE Transactions on Automatic Control*, 2004.