

An Alternative Simulation Budget Allocation Scheme for Efficient Simulation

Chun-Hung Chen⁺

Department of Systems Engineering & Operations Research
George Mason University
Fairfax, VA 22030, U.S.A.

Enver Yücesan

Technology Management Area
INSEAD
77305 Fontainebleau, France

Abstract

We present an alternative simulation run allocation scheme for maximizing efficiency in simulation experiments for decision making under uncertainty. The issue of simulation efficiency is addressed from two perspectives: i) We want to minimize the total computation cost, with a constraint that the overall simulation quality must be higher than a desired level; ii) we would like to maximize the simulation quality with the constraint that the total computation cost can not exceed a given budget. While these two problems look different, we show that the solutions to these two problems are identical. Comparisons with other procedures show that our approach can achieve a speedup factor of 3~4 for a 10-design example. The speedup factor is even higher for problems having a larger number of designs.

+ **Corresponding author:** Professor Chun-Hung Chen, Dept. of Systems Engineering & Operations Research, George Mason University, 4400 University Drive, MS 4A6, Fairfax, VA 22030; Tel: 703-993-3572; Fax: 703-993-1521; Email: cchen9@gmu.edu.

++ This work has been supported in part by NSF under Grants DMI-0002900, DMI-0049062 and IIS-0325074, by NASA Ames Research Center under Grants NAG-2-1565 and NAG-2-1643, by FAA under Grant 00-G-016, and by George Mason University Research Foundation.

1. Introduction

This paper is concerned with the efficiency issue in simulation-based decision making. Simulation is a popular tool for analyzing systems and evaluating decision problems since real situations rarely satisfy the assumptions of analytical models. Stochastic simulation technology, such as discrete-event simulation and Monte Carlo simulation, has matured over the past decade and is now commonly used to evaluate large-scale real systems with complex stochastic behavior. Simulation allows one to more accurately specify a system through the use of logically complex, and often non-algebraic, variables and constraints. This capability relaxes the inherent limitation of traditional optimization. However, the added flexibility often creates models that are computationally intensive. To obtain a good statistical estimate for a design decision, a large number of simulation runs is usually required for each design alternative. This is due to the stochastic features and the slow convergence of a performance measure estimator relative to the number of runs. Furthermore, many alternative designs must be simulated in order to find a good design. The total computation cost for simulation-based decision-making approaches may be too expensive.

There exists a large literature on innovative methods for improving the efficiency of simulation experiments. Bratley et al. (1987) and Fishman (1996) provide a comprehensive presentation of recent developments in simulation methodologies. Some methods exploit the fact that the required number of simulation runs decreases when the simulation variance is reduced. Experimental as well as theoretical results for some special cases have shown that inducing some sort of dependence among experiments for different designs can increase the chance of selecting the true best design. In other words, the so-called variance reduction techniques exploit the fact that the required number of simulation runs decreases when the simulation variance is reduced. Glasserman and Yao (1992) show that the schemes of common random numbers and control variates are helpful in obtaining better confidence intervals for various selection procedures when the performance measure is obtained by averaging i.i.d. random variables with normal distributions. Heidelberger (1993) deals with rare event problems by developing an importance sampling scheme. Antithetic Variates induce negative correlation between separate runs (e.g., Cheng 1982, 1984). The selection of an adequate variance reduction or correlated sampling technique usually depends on the particular model of interest. Therefore, a thorough understanding of the inner workings of the models is required for proper use of those techniques. Another major limitation of variance reduction techniques is that only the information of each design is locally used to reduce its variance for improving simulation efficiency. The total simulation cost for a design problem may still be high.

On the other hand, several researchers have shown that allocating simulation runs in an uneven manner can significantly enhance simulation efficiency by reducing the total number of runs required to identify the best candidate design. Previous work such as Rinott (1978) develops two-stage procedures for allocating runs to designs and several papers (as listed in Bechhofer et al. 1995) extend the technique to general ranking and selection problems. The major disadvantage of existing approaches is that they utilize only the information of variance to

control simulation experiments. Chen et al. (1998, 1999) address this limitation and demonstrate that the use of additional information on relative means among different alternatives can dramatically improve simulation efficiency even if only a simple heuristic is applied. Hyden and Schruben (2000), Chick and Inoue (2001), Lee and Chew (2003), Trailovic and Pao (2001, 2004) also demonstrate that simulation efficiency can be improved by utilizing more simulation information that is readily available.

The efficient simulation technique introduced in this paper optimally allocates a computing budget to the designs under evaluation. Intuitively, to ensure a high probability of correctly selecting an optimal design, a larger portion of the computing budget should be allocated to those designs that are critical in the process of identifying good designs. In other words, a larger number of simulations must be conducted with those critical designs in order to reduce estimator variance. Similarly, limited computational effort should be expended on non-critical designs that have little effect on identifying the good designs even if they have large variances. Overall simulation efficiency is improved as less computational effort is spent on simulating non-critical designs and more is spent on critical designs. The ideas are explained using the following simple example. Suppose we are performing simulations for 5 design alternatives in order to determine a design with minimum mean performance measure. First, we conduct some preliminary simulation runs for all 5 designs. Figure 1 gives an example of their 99% confidence intervals obtained from these preliminary simulations. Note that the uncertainty of estimation is due to the system's stochastic features.

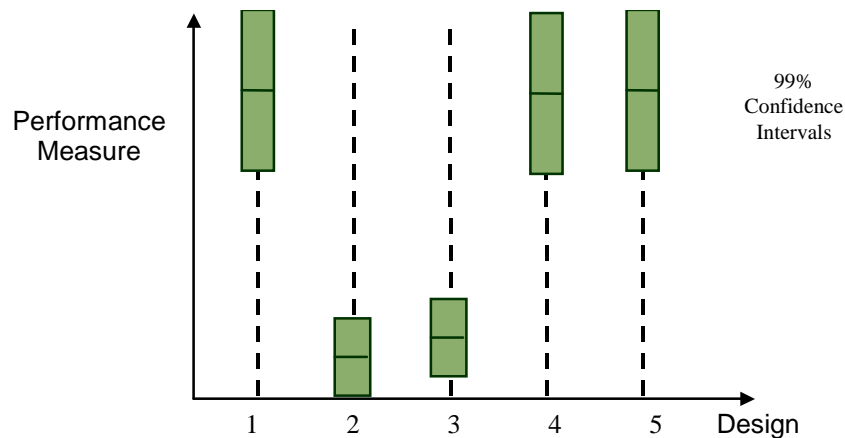


Figure 1. 99% confidence intervals for 5 design alternatives after some preliminary simulation. The intension is to determine a design with minimum mean performance measure.

As seen in Figure 1, while there is uncertainty in the estimation of the performance for each design, it is obvious that designs 2 and 3 are much better than the other designs, if we intend to find a design with minimum performance measure. And so only designs 2 and 3 need to be

further simulated to reduce estimation uncertainty in order to correctly identify the best design. By stopping simulations for designs 1, 4, and 5 earlier, we can save considerable computation cost.

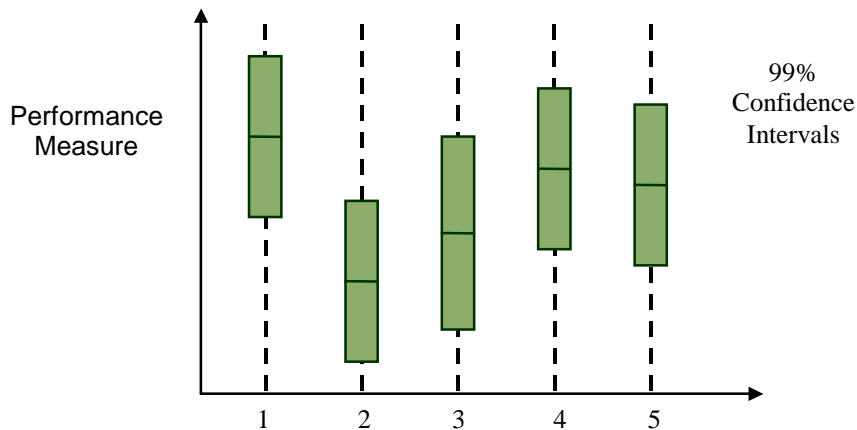


Figure 2. 99% confidence intervals after some preliminary simulation. The intension is to determine a design with minimum mean performance measure.

However, most cases are not as trivial as that in Figure 1. It is more common to see cases like the example shown in Figure 2, where some designs seem better, but not outstandingly better, than other designs. It is not straightforward to determine which designs can be eliminated from the simulation experiment in general, and when they should be eliminated. Ideally, one would like to allocate simulation trials to designs in a way that maximizes simulation efficiency. Chen et al. (2000) present a new computing budget allocation technique to asymptotically maximize the probability of correctly selecting the best design within a given computing budget. Numerical testing shows that the speedup factor can be several orders of magnitude compared to traditional approaches.

This paper addresses the issue of optimal computing budget allocation (OCBA) from a different perspective. In particular, we would like to minimize the total simulation cost for achieving a desired level of probability of correctly selecting the best design. We will formulate this goal as an optimization problem and obtain an asymptotic solution. While these two problems are quite different, the optimal budget allocation solution can to be identical to that in Chen et al. (2000).

The paper is organized as follows: In the next section, we formulate the optimal computing budget allocation problem. Section 3 presents an asymptotic allocation rule for OCBA. 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 consider a general simulation experiment for decision making defined as

$$\min_{\theta_i \in \Theta} J(\theta_i) \equiv E[L(\theta_i, \xi)], \quad (1)$$

where Θ , the design space, is an arbitrary, unstructured, very large but finite set; θ_i is the system design parameter vector for design i , $i = 1, 2, \dots, k$; J , the performance criterion which is the expectation of L , the sample performance, as a functional of θ , and ξ , a random vector that represents uncertain factors in the systems. $L(\theta, \xi)$ is available only in the form of a complex calculation via *simulation* for the systems considered in this paper. In the simulation experiment, it is intended to simulate all designs in order to find the best design. The system constraints are implicitly involved in the simulation process, and so are not shown in (1). The standard approach is to estimate $E[L(\theta_i, \xi)]$ by the sample mean performance measure

$$\bar{J}_i \equiv \frac{1}{N_i} \sum_{j=1}^{N_i} L(\theta_i, \xi_{ij}),$$

where ξ_{ij} represents the j -th sample of ξ and N_i represents the number of simulation samples for design i . Denote by

σ_i^2 : the variance for design i , i.e., $\sigma_i^2 = \text{Var}(L(\theta_i, \xi))$. Assume σ_i^2 is finite. In practice, σ_i^2 is unknown beforehand and so is approximated by sample variance.

b : the design having the smallest sample mean performance measure, i.e., $\bar{J}_b \leq \min_i \bar{J}_i$,

$$\delta_{b,i} \equiv \bar{J}_b - \bar{J}_i,$$

$$\sigma_{b,i}^2 \equiv \frac{\sigma_b^2}{N_b} + \frac{\sigma_i^2}{N_i}.$$

While the design with the smallest sample mean (design b) is usually picked, design b is not necessarily the one with the smallest unknown mean performance. *Correct selection* (CS) is therefore defined as the event that design b is actually the best design (i.e., with the smallest population mean). As N_i increases, \bar{J}_i becomes a better approximation to $J(\theta_i)$ in the sense that its corresponding confidence interval becomes narrower, and the probability of correct selection, $P\{\text{CS}\}$, becomes larger as well. Note that each sample of $L(\theta_i, \xi_{ij})$ requires one simulation run. To ensure that $P\{\text{CS}\}$ is sufficiently large, the required number of samples of $L(\theta_i, \xi_{ij})$ for all designs may become large, making the simulation experiment very time consuming.

Chen et al. (2000) formulate the problem of simulation experiment efficiency as an optimization problem. The purpose is to optimally allocate a computing budget to the designs under evaluation in a way that maximizes simulation efficiency. Stating this more precisely, it is

intended to find N_1, N_2, \dots, N_k such that $P\{\text{CS}\}$ is maximized, subject to a limited computing budget T ; i.e.,

$$\begin{aligned} & \max_{N_1, \dots, N_k} P\{\text{CS}\} \\ & \text{s.t. } N_1 + N_2 + \dots + N_k = T. \\ & N_i \in \Gamma, i = 1, \dots, k. \end{aligned} \tag{2}$$

Here Γ is the set of non-negative integers and $N_1 + N_2 + \dots + N_k$ denotes the total computational cost assuming the simulation execution times for different designs are roughly the same. Chen et al. (2000) propose an asymptotic solution to problem (2).

In this paper, we want to consider the simulation efficiency problem from a different perspective. Instead of asking "what is the highest $P\{\text{CS}\}$ we can achieve with a fixed computing budget," one may ask the question "what is the most efficient way (with minimum computation cost) to conduct simulation experiments in order to achieve a desired $P\{\text{CS}\}$, say 95%?" The question can be formulated as that of determining the minimum number of total simulation runs to achieve a desired $P\{\text{CS}\}$, say P^* , in the simulation experiments. That is,

$$\begin{aligned} & \min_{N_1, \dots, N_k} [N_1 + N_2 + \dots + N_k] \\ & \text{s.t. } P\{\text{CS}\} = P^*. \\ & N_i \in \Gamma, i = 1, \dots, k. \end{aligned} \tag{3}$$

In this paper, we will introduce an asymptotic solution to problem (3). Furthermore, we will show that the asymptotic solutions to problems (2) and (3) are identical under some conditions, while these two problem formulations are different. In the next section, an asymptotic solution to problem (3) will be presented.

3. An Asymptotic Allocation Rule

To solve problem (3), we must be able to first estimate $P\{\text{CS}\}$. We follow the Bayesian model introduced in Chen et al. (2000) and assume that the simulation output is normally distributed. Let \tilde{J}_i denote the random variable whose probability distribution is the posterior distribution for design i , which is constructed based on two pieces of information: (i) prior knowledge of the system's performance, and (ii) observed simulation output. Chen et al. (2000) offer a relatively fast and inexpensive way of estimating $P\{\text{CS}\}$ given in the following Lemma.

Lemma 1. A good approximation to $P\{\text{CS}\}$ is given by

$$APCS \equiv 1 - \sum_{i=1, i \neq b}^k P\{\tilde{J}_b > \tilde{J}_i\}$$

We refer to this approximation as the *Approximate Probability of Correct Selection (APCS)*. In fact, $APCS \leq P\{\text{CS}\}$ and $APCS$ is asymptotically close to $P\{\text{CS}\}$.

In this paper, we consider non-informative prior distributions. This implies that no prior knowledge is available about the performance of any design alternative before conducting the simulation.

Lemma 2. If the simulation output is normally distributed and if no prior knowledge is available about the performance before conducting the simulation, then (DeGroot 1970)

$$\tilde{J}_i \sim N(\bar{J}_i, \frac{\sigma_i^2}{N_i}), \text{ for } i = 1, \dots, k.$$

With Lemma 1 and Lemma 2, the approximation of $P\{\text{CS}\}$ can be expressed as

$$\begin{aligned} APCS &= 1 - \sum_{i=1, i \neq b}^k P\{\tilde{J}_b > \tilde{J}_i\} \\ &= 1 - \sum_{\substack{i=1 \\ i \neq b}}^k \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma_{b,i}} e^{-\frac{(x-\delta_{b,i})^2}{2\sigma_{b,i}^2}} dx \\ &= 1 - \sum_{\substack{i=1 \\ i \neq b}}^k \int_{-\frac{\delta_{b,i}}{\sigma_{b,i}}}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt. \end{aligned} \tag{4}$$

Note that $\delta_{b,i}$ and $\sigma_{b,i}^2$ have been defined in section 2.

Lemma 3. Let the total number of simulation samples be $n = N_1 + N_2 + \dots + N_k$. If $APCS \rightarrow 1.0$, then $n \rightarrow \infty$.

This lemma can be easily established. If $APCS \rightarrow 1.0$, then all the terms within the summation in (4) must approach 0. This implies that $\sigma_{b,i} \rightarrow 1.0$ and so $N_i \rightarrow \infty$ for $i = 1, 2, \dots, k$. Since $APCS$ is a lower bound of $P\{\text{CS}\}$, $P\{\text{CS}\} \rightarrow 1.0$ as $APCS \rightarrow 1.0$.

We consider the following problem:

$$\begin{aligned} &\min_{N_1, \dots, N_k} [N_1 + N_2 + \dots + N_k] \\ &s.t. \ 1 - \sum_{\substack{i=1 \\ i \neq b}}^k \int_{-\frac{\delta_{b,i}}{\sigma_{b,i}}}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = P^*. \\ &N_i \in \Gamma, i = 1, \dots, k. \end{aligned} \tag{5}$$

First, we assume the variables, N_i 's, are continuous. Second, our strategy is to tentatively ignore all non-negativity constraints; all N_i 's can therefore assume any real value. Let $N_i = \alpha_i n$. Thus, $\sum_{i=1}^k \alpha_i = 1$. Before the end of this section, we will show how all α_i 's become positive and hence all N_i 's are positive. Based on this non-integral assumption, we first consider the following:

$$\begin{aligned} \min_{N_1, \dots, N_k} & [N_1 + N_2 + \dots + N_k] \\ \text{s.t. } & 1 - \sum_{\substack{i=1 \\ i \neq b}}^k \int_{-\frac{\delta_{b,i}}{\sigma_{b,i}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = P^*. \end{aligned} \quad (6)$$

Let F be the Lagrangian of (6):

$$F = \sum_{i=1}^k N_i - \lambda \left(1 - \sum_{\substack{i=1 \\ i \neq b}}^k \int_{-\frac{\delta_{b,i}}{\sigma_{b,i}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt - P^* \right).$$

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

$$\frac{\partial F}{\partial N_i} = 1 + \frac{\lambda}{2\sqrt{2\pi}} \exp\left[-\frac{\delta_{b,i}^2}{2\sigma_{b,i}^2}\right] \frac{\delta_{b,i}\sigma_i^2}{N_i^2(\sigma_{b,i}^2)^{3/2}} = 0, \text{ for } i = 1, 2, \dots, k, \text{ and } i \neq b. \quad (7)$$

$$\frac{\partial F}{\partial N_b} = 1 + \sum_{\substack{i=1 \\ i \neq b}}^k \frac{\lambda}{2\sqrt{2\pi}} \exp\left[-\frac{\delta_{b,i}^2}{2\sigma_{b,i}^2}\right] \frac{\delta_{b,i}\sigma_b^2}{N_b^2(\sigma_{b,i}^2)^{3/2}} = 0, \quad (8)$$

$$\lambda \left(1 - \sum_{\substack{i=1 \\ i \neq b}}^k \int_{-\frac{\delta_{b,i}}{\sigma_{b,i}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt - P^* \right) = 0, \text{ and } \lambda \geq 0.$$

We now examine the relationship between N_b and N_i for $i = 1, 2, \dots, k$, and $i \neq b$. From Eq. (7),

$$\frac{\lambda}{2\sqrt{2\pi}} \exp\left[-\frac{\delta_{b,i}^2}{2\sigma_{b,i}^2}\right] \frac{\delta_{b,i}}{(\sigma_{b,i}^2)^{3/2}} = -\frac{N_i^2}{\sigma_i^2}, \text{ for } i = 1, 2, \dots, k, \text{ and } i \neq b. \quad (9)$$

Plugging (9) into (8), we have

$$1 - \sum_{\substack{i=1 \\ i \neq b}}^k \frac{N_i^2 \sigma_b^2}{N_b^2 \sigma_i^2} = 0.$$

Then

$$N_b = \sigma_b \sqrt{\sum_{i=1, i \neq b}^k \frac{N_i^2}{\sigma_i^2}} \quad \text{or} \quad \alpha_b = \sigma_b \sqrt{\sum_{i=1, i \neq b}^k \frac{\alpha_i^2}{\sigma_i^2}}. \quad (10)$$

We further investigate the relationship between N_i and N_j , for any $i, j \in \{1, 2, \dots, k\}$, and $i \neq j \neq b$. From Eq. (7),

$$\exp\left(\frac{-\delta_{b,i}^2}{2\left(\frac{\sigma_b^2}{N_b} + \frac{\sigma_i^2}{N_i}\right)}\right) \cdot \frac{\delta_{b,i} \sigma_i^2 / N_i^2}{\left(\frac{\sigma_b^2}{N_b} + \frac{\sigma_i^2}{N_i}\right)^{3/2}} = \exp\left(\frac{-\delta_{b,j}^2}{2\left(\frac{\sigma_b^2}{N_b} + \frac{\sigma_j^2}{N_j}\right)}\right) \cdot \frac{\delta_{b,j} \sigma_j^2 / N_j^2}{\left(\frac{\sigma_b^2}{N_b} + \frac{\sigma_j^2}{N_j}\right)^{3/2}}. \quad (11)$$

Taking the log on Eq. (11) and assuming that $N_b \gg N_i$ (note that this assumption is not too bad if we examine the equation (10)), we have

$$\frac{\delta_{b,j}^2}{\sigma_j^2} N_j + \log(N_j) = \frac{\delta_{b,i}^2}{\sigma_i^2} N_i + \log(N_i) + 2 \log\left(\frac{\delta_{b,j} \sigma_i}{\delta_{b,i} \sigma_j}\right),$$

or

$$\frac{\delta_{b,j}^2}{\sigma_j^2} \alpha_j n + \log(\alpha_j n) = \frac{\delta_{b,i}^2}{\sigma_i^2} \alpha_i n + \log(\alpha_i n) + 2 \log\left(\frac{\delta_{b,j} \sigma_i}{\delta_{b,i} \sigma_j}\right),$$

which yields

$$\frac{\delta_{b,j}^2}{\sigma_j^2} \alpha_j n + \log(\alpha_j) = \frac{\delta_{b,i}^2}{\sigma_i^2} \alpha_i n + \log(\alpha_i) + 2 \log\left(\frac{\delta_{b,j} \sigma_i}{\delta_{b,i} \sigma_j}\right). \quad (12)$$

To further facilitate the computations, we intend to find an asymptotic allocation rule. Namely, we consider the case where $P^* \rightarrow 1$ or $APCS \rightarrow 1.0$, and so $n \rightarrow \infty$. While it is impossible to have an infinite computing budget in practice, our allocation rule provides a simple means for allocating simulation budget in a way that efficiency can be significantly improved. As $n \rightarrow \infty$, all the log terms become much smaller than the other terms and are therefore negligible. This implies

$$\frac{\delta_{b,j}^2}{\sigma_j^2} \alpha_j = \frac{\delta_{b,i}^2}{\sigma_i^2} \alpha_i.$$

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

$$\frac{N_i}{N_j} = \frac{\alpha_i}{\alpha_j} = \left(\frac{\sigma_i / \delta_{b,i}}{\sigma_j / \delta_{b,j}}\right)^2 \quad \text{for } i = 1, 2, \dots, k, \text{ and } i \neq j \neq b. \quad (13)$$

Now we return to the nonnegativity constraints for N_i , which we have temporarily ignored. Note that from Eq. (10) and Eq. (13), all α_i 's have the same sign. Since $\sum_{i=1}^k \alpha_i = 1$ and $N_i = \alpha_i n$, it implies that all α_i 's ≥ 0 , and hence N_i 's ≥ 0 , for $i = 1, 2, \dots, k$.

In conclusion, if a solution satisfies Eq. (10) and Eq. (13), then KKT conditions must hold asymptotically. According to the KKT Sufficient Condition, this solution is a locally optimal solution to Eq. (5). We therefore have the following result:

Theorem 1. Suppose there are k competing designs whose performance is depicted by random variables with means $J(\theta_1), J(\theta_2), \dots, J(\theta_k)$, and finite variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2$, respectively. Given a desired level of *Approximate Probability of Correct Selection (APCS)*, P^* , the total number of simulation samples, $N_1 + N_2 + \dots + N_k$, can be asymptotically minimized when

$$(1) \frac{N_i}{N_j} = \left(\frac{\sigma_i / \delta_{b,i}}{\sigma_j / \delta_{b,j}} \right)^2, \quad i, j \in \{1, 2, \dots, k\}, \text{ and } i \neq j \neq b,$$

$$(2) N_b = \sigma_b \sqrt{\sum_{i=1, i \neq b}^k \frac{N_i^2}{\sigma_i^2}},$$

where N_i is the number of samples allocated to design i , $\delta_{b,i} = \bar{J}_b - \bar{J}_i$, and $\bar{J}_b \leq \min_i \bar{J}_i$.

The asymptotic solution given in Theorem 1 for problem (3) is actually identical to the asymptotic solution given by Chen et al. (2000) for problem (2), although these problems look different. In fact, they have the same objective to maximize simulation efficiency. In actual simulation experiments, the situation can be quite different from asymptotic condition; furthermore, the means and variance are usually unknown. It is therefore a good idea to apply Theorem 1 in a multi-stage sequential sampling setting. Initially, n_0 simulation replications for each of k design are conducted to get some information about the performance of each design during the first stage. As simulation proceeds, the sample means and sample variances of each design are computed from the data already collected up to that stage. Then *APCS* can be calculated using Lemma 1. If *APCS* is not sufficiently high, an incremental computing budget, Δ , is allocated based on Theorem 1 at each stage. Ideally, each new replication should bring us closer to the optimal solution. This procedure is continued until the desired *APCS* level is achieved or the entire computing budget T is exhausted.

4. Numerical Testing and Comparison with Other Allocation Procedures

In this section, we test our OCBA algorithm and compare it with different allocation procedures by performing several numerical experiments. In particular, we will emphasize how OCBA performs as we increase the number of alternative designs. In all the numerical illustrations, we

estimate $P\{\text{CS}\}$ by counting the number of times we successfully find the true best design (design 1 in this example) out of 10,000 independent applications of each selection procedure. $P\{\text{CS}\}$ is then obtained by dividing this number by 10,000, representing the correct selection frequency.

We compare OCBA with two other well-known simulation procedures: Equal Allocation, which simulates all design alternatives equally, and Rinott's method that assigns more simulation time to alternatives with higher estimated variances (this method is highly popular in the simulation literature). We briefly summarize these allocation procedures as follows.

Equal Allocation

This is the simplest way to conduct simulation experiments and has been widely applied. The simulation budget is equally allocated to all designs. Namely, we simulate all design alternatives through an equal number of replications, that is, $N_i = T/k$ for each i . The performance of equal allocation will serve as a benchmark for comparison.

Modified Rinott Procedure

The two-stage procedure of Rinott (1978) has been widely applied in the simulation literature. Unlike the OCBA approach, the two-stage procedures are developed based on the classical (frequentist) statistical model. See Bechhofer et al. (1995) for a systematic discussion of two-stage procedures. 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, the number of additional simulation samples for each design in the second stage is determined by:

$$N_i = \max(0, \lceil S_i^2 h^2 / d^2 \rceil - n_0), \text{ for } i = 1, 2, \dots, k,$$

where $\lceil \bullet \rceil$ is the integer "round-up" function, d is the size of the indifference zone, h is a constant, which solves Rinott's integral. In short, the computing budget is allocated proportionally to the estimated sample variances. To put Rinott procedure in a sequential setting for a fair comparison, we modify the procedure in a way that we sequentially determine N_i based on the newly updated sample variances and that N_i is proportional to the sample variance of design i .

4.1 Comparison of Different Procedures

We consider ten design alternatives. Suppose $L(\theta_i, \xi) \sim N(i, \sigma^2)$, $i = 1, 2, \dots, 10$. We want to find a design with the smallest mean. It is obvious that design 1 is the actual best design. Note that the information about design 1 being the true best design is used only in the calculation of the resulting $P\{\text{CS}\}$, but not used in any of the simulation procedures. In the numerical experiment, we compare the convergence of $P\{\text{CS}\}$ for different allocation procedures. We have $n_0 = 10$ and $\Delta = 20$.

Different computing budgets are tested. Figure 3 shows the test results using OCBA and the other two allocation procedures. We see that all procedures obtain a higher $P\{\text{CS}\}$ as the

available computing budget increases. However, OCBA achieves the same $P\{CS\}$ with a lower amount of computing budget than other procedures. In particular, Figure 3 indicates the computation costs in order to attain $P\{CS\} = 99\%$ for different procedures. In this setting, OCBA reduces the simulation time by 75%.

It is worth noting that Rinott’s procedure does not perform much better than the simple equal allocation scheme. This is because Rinott’s procedure determines the number of simulation samples for all designs using only the information on sample variances. This handicap becomes even more striking when Rinott’s procedure is compared with OCBA: when determining budget allocation, OCBA exploits the information on both sample means and variances, while Rinott’s procedure does not utilize the (readily available) information on sample means. The sample means can provide valuable information on the relative differences across the design space (as was illustrated in Figure 1).

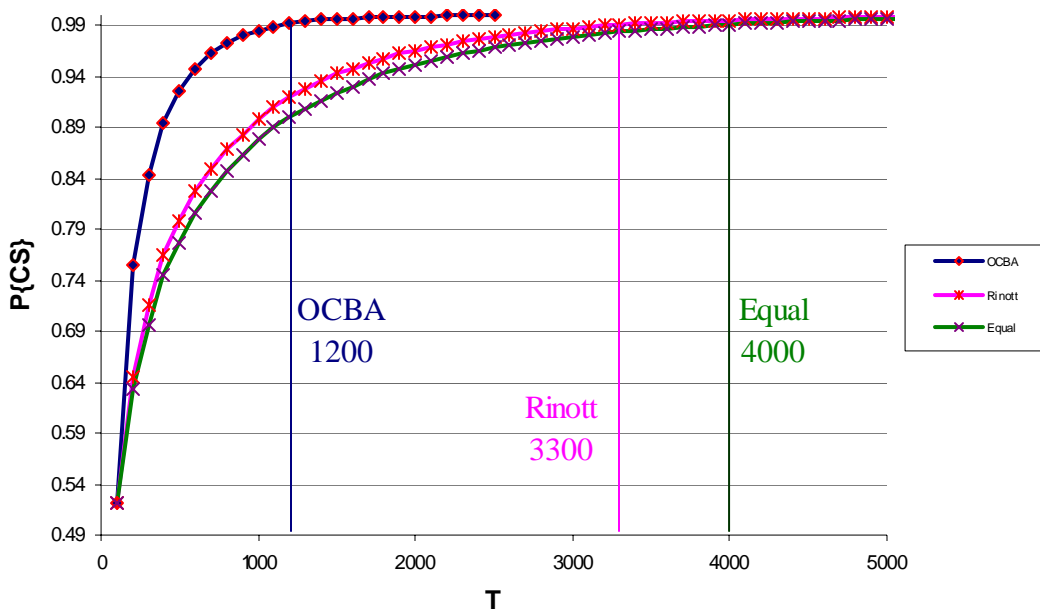


Figure 3. $P\{CS\}$ vs. T using three different allocation procedures in experiment 1.

4.2 Speed-up Factors Using OCBA

In the second experiment, we study how well OCBA performs for different numbers of design alternatives – here, up to 100 alternatives. Suppose $L(\theta_i, \xi) \sim N(10i/k, 1^2)$, $i = 1, 2, \dots, k$, where k is the number of designs. Under this setting, regardless of the value of k , the range of the means for these k designs is the same as those in the earlier 10-design experiment, namely, from 1 to 10. In this test, we compare OCBA and equal allocation, and focus on the speedup factors under OCBA. For both procedures, we record the minimum computation cost to reach $P\{CS\} = 99\%$: T_{OCBA} and T_{EA} . The speedup factor using OCBA is given by the ratio T_{EA} / T_{OCBA} . It is also

useful to measure the so-called *Equivalent Number of Alternatives with a Fixed Computing Budget*, $ENAFCB(k)$, which is defined as the number of alternatives that can be simulated under the equal allocation procedure using the computing budget that is needed for OCBA to simulate k alternatives for reaching $P\{CS\}=99\%$. For example, in the case of 10 alternatives, OCBA is 3.42 times faster than equal allocation. Thus, OCBA can simulate 10 alternatives in the same time that equal allocation can simulate only $10/3.42 = 2.93$ alternatives. In this case, $ENAFCB(10)$ for OCBA is 2.93. An alternative interpretation of this statement is that, under OCBA, we can simulate 10 alternatives with only the equivalent effort of 2.93 alternatives. The speedup factors and ENAFCB's for different number of alternatives are given in Table 1.

Table 1. The speedup factor of using OCBA as compared with the use of equal allocation.

Number of designs, k	4	10	20	50	75	100
Speedup factor using OCBA	1.75	3.42	6.45	12.8	16.3	19.8
$ENAFCB(k)$	2.29	2.93	3.10	3.90	4.59	5.05

We see that OCBA is even more efficient as the number of designs increases. The higher efficiency is obtained because a larger design space gives the OCBA algorithm more flexibility in allocating the computing budget. In particular, $ENAFCB(10) = 2.93$ and $ENAFCB(50) = 3.90$. This means that with an equivalent effort of less than 4 alternatives, OCBA can simulate 50 alternatives!

To thoroughly examine the speedup property of OCBA, we further increase the number of designs in the testing. We extend the test by increasing the number of design alternatives from 100 to 1150. Figure 4 shows that the speedup factor initially increases almost linearly when the number of alternatives is small. However, when the number of alternatives is large, the speedup converges to some maximum level. This is due to the initial sampling cost for OCBA. Initial sampling is needed to gather the first estimates for means and variances so that the optimal computing budget allocation can be determined. Because we do not assume any prior knowledge about the topology of the design space, this initial sampling cost is proportional to the number of alternatives and will become dominating when the number of alternatives is large.

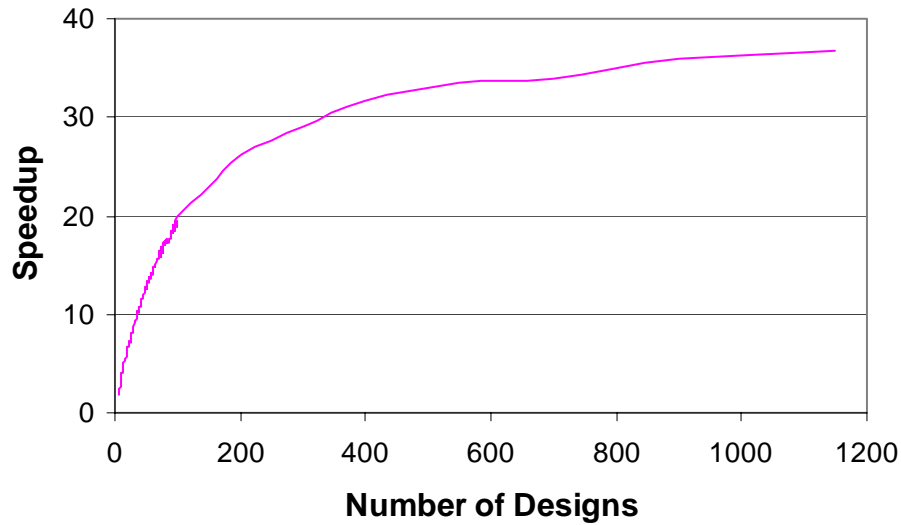


Figure 4. Speedup factor vs. T using OCBA.

5. Conclusions

We present a new procedure to enhance the efficiency of simulation experiments for decision making under uncertainty. The objective is to maximize the simulation efficiency, expressed as the probability of correct selection within a given computing budget. We consider two ways of maximizing simulation efficiency: i) the first is to minimize the total computation cost, with a constraint that the simulation quality must be higher than a desired level; ii) an alternative is to maximize the simulation quality with the constraint that the total computation cost can not exceed a given budget. While these two problems appear to be different, we show that their asymptotic solutions are identical. Comparisons with other procedures show that our approach can achieve a speedup factor of 3~4 for a 10-design example. The speedup factor is even higher with the problems having a larger number of designs. For example, the speedup factor is 19.8 when there are 100 designs in the simulation experiment. Indeed, this represents significant savings in computation cost. We believe that OCBA can change the way of thinking about many stochastic optimization methodologies.

References

1. 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, Inc., 1995.
2. Bratley, P., B. L. Fox, and L. E. Schrage, *A Guide to Simulation*. 2nd ed. Springer-Verlag, 1987.

3. 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.
4. Chen, C. H., S. D. Wu, and L. Dai, "Ordinal Comparison of Heuristic Algorithms Using Stochastic Optimization," *IEEE Transactions on Robotics and Automation*, Vol. 15, No. 1, pp. 44-56, February 1999.
5. Chen, C. H., E. Yücesan, Y. Yuan, H. C. Chen and L. Dai, "Computing Budget Allocation for Simulation Experiments with Different System Structures," *Proceedings of the 1998 Winter Simulation Conference*, pp. 735-741, December 1998.
6. Cheng, R. C. H. "The Use of Antithetic Variates in Computer Simulations," *Journal of Operational Research Society*, 15, pp. 227-236, 1984.
7. Cheng, R. C. H. "Antithetic Variate Methods for Simulation of Processes with Peaks and Troughs," *European Journal of Operations Research*, 33, pp. 229-237, 1982.
8. 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.
9. Fishman, G. *Monte Carlo: Concepts, Algorithms, and Applications*, Springer-Verlag, 1996.
10. Glasserman, P. and D. D. Yao, "Some Guidelines and Guarantees for Common Random Numbers," *Management Science*, Vol. 38, No. 6, pp. 884-908, 1992.
11. Heidelberger, P. "Fast Simulation of Rare Events in Queueing and Reliability Models," In *Performance Evaluation of Computer and Communication Systems*, ed. L. Donatiello and R. Nelson, pp. 165-202, Springer Verlag, 1993.
12. 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.
13. 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.
14. Rinott, Y., "On Two-stage Selection Procedures and Related Probability Inequalities," *Communications in Statistics A7*, 799-811, 1978.
15. Walker, R. C. *Introduction to Mathematical Programming*, Prentice Hall, Upper Saddle River, NJ, 1999
16. Trailovic, L. and L. Y. Pao, "Computing Budget Allocation for Optimization of Sensor Processing Order in Sequential Multi-sensor Fusion Algorithms," *Proceedings of American Control Conference*, June 2001.
17. 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.