

## EFFICIENT SELECTION OF AN OPTIMAL SUBSET FOR OPTIMIZATION UNDER UNCERTAINTY

Chun-Hung Chen  
Donghai He

Michael Fu

Loo Hay Lee

Department of Systems Engineering &  
Operations Research  
George Mason University  
4400 University Drive, MS 4A6  
Fairfax, VA 22030

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

Department of Industrial & Systems  
Engineering  
The National University of Singapore  
Kent Ridge, Singapore 119260

### 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 present 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. As an ongoing research, we are integrating the efficient selection procedure with search algorithms, such as genetic algorithms or cross entropy method, for large-scale simulation-based optimization under uncertainty. We will present our development and new observations at the workshop.

### 1 INTRODUCTION

In this paper, we propose to consider simulation-based optimization under uncertainty. When applied to the simulation setting, several recent developments in global optimization 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).

To address this need, we first 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). 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 at least one design whose mean response is within a pre-specified distance from the optimal mean response. 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 Thümmel (2005), and Boesel, Nelson, and

Kim (2003), 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) develop a two-stage procedure for selecting all the  $m$  best designs. 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 Dudewicz 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.

The paper is organized as follows. In the next section, we formulate the optimal computing budget allocation problem for selecting the top- $m$  designs and present an asymptotically optimal allocation procedure. The performance of the technique is illustrated with a series of numerical examples in Section 3. Section 4 gives our thoughts for integrating the efficient selection procedure with evolutionary search methods for large-scale simulation-based optimization under uncertainty.

## 2 AN EFFICIENT SIMULATION BUDGET ALLOCATION FOR SELECTING AN OPTIMAL SUBSET

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  $i$ ,
- $\bar{J}_i$  = sample mean for design  $i$ ,
- $J_i$  = mean for design  $i$ ,
- $S_i^2$  = variance for design  $i$ .

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:

$$CS = \left\{ \bigcap_{i \in S_m} \bigcap_{j \notin S_m} (J_i \leq J_j) \right\} = \left\{ \max_{i \in S_m} J_i \leq \min_{i \notin S_m} J_i \right\}. \quad (1)$$

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

$$\begin{aligned} & \max_{N_1, \dots, N_k} P\{CS\} \\ & \text{s.t. } N_1 + N_2 + \dots + N_k = T. \end{aligned} \quad (2)$$

Here  $N_1 + N_2 + \dots + N_k$  denotes the total computational cost assuming the simulation execution times for different designs are roughly the same.

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  $\{\bar{J}_1, \bar{J}_2, \dots, \bar{J}_k\}$ , i.e.,  $\bar{J}_{i_1} = \bar{J}_{i_2} = \dots = \bar{J}_{i_k}$ . Then, the selected subset is given by

$$S_m = \{i_1, i_2, \dots, i_m\}.$$

We assume that the simulation output samples  $\{X_{ij}\}$  are normally distributed and independent from replication to replication as well as independent across designs. To solve the OCBA problem in (2), we estimate  $P\{CS\}$  using the Bayesian model presented in Chen et al. (2000) and He et al. (2006). For ease of computation, we adopt an approximation of  $P\{CS\}$  using a lower bound, which is re-

ferred as the *Approximate Probability of Correct Selection for m best (APCSm)*. Then the approximation *APCSm* can be asymptotically maximized (Chen et al. 2007), when

$$\frac{N_i}{N_j} = \left( \frac{s_i/d_i}{s_j/d_j} \right)^2, \quad i, j \in \{1, 2, \dots, k\}, \text{ and } i \neq j. \quad (3)$$

where  $d_i = \bar{J}_i - (\bar{J}_{i_m} + \bar{J}_{i_{m+1}}) / 2$ .

### 3 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  $J_{[1]} < J_{[2]} < \dots < J_{[k]}$ , so design [1] is the best and correct selection would be  $S_m = \{[1], [2], \dots, [m]\}$  (but this is unknown a priori).

In comparing the procedures, the measurement of effectiveness used is the  $P\{CS\}$  estimated by the fraction of times the procedure successfully finds *all* the true *m*-best designs out of 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).

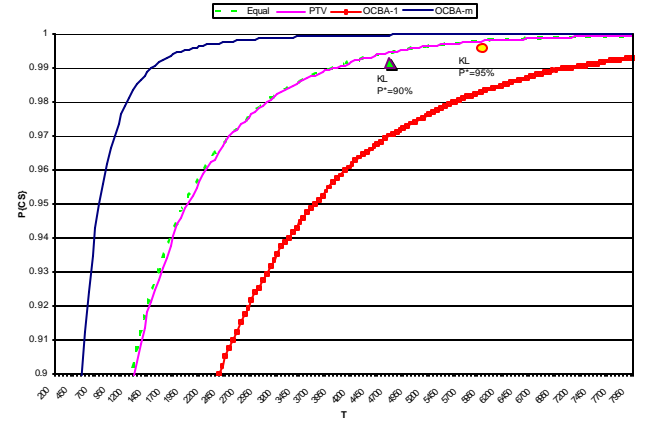
In the numerical experiment, there are 10 alternative designs, with distribution  $N(i, \sigma^2)$  for design  $i = 1, 2, \dots, 10$ . The goal is to identify the top-3 designs via simulation samples, i.e.,  $m=3$ .

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 1. For KL, we test two cases  $P^* = 0.9$  and  $P^* = 0.95$ , and the corresponding estimated  $P\{CS\}$  vs. the average total simulation replications are shown as two single points (the triangle and circle) in Figure 1.

We see that all procedures obtain a higher  $P\{CS\}$  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 1.**  $P\{CS\}$  vs.  $T$  using four sequential allocation procedures and KL ( triangle for  $P^*=90\%$  and circle for  $P^*=95\%$  ) for Exa mple 1.

### 4 INTEGRATION WITH EVOLUTIONARY ALGORITHMS

In developing a new approach for large-scale discrete ranking and selection or continuous stochastic optimization problems, as an ongoing research, we are integrating the OCBA procedure with evolutionary search algorithms, such as genetic algorithms or cross entropy method. The evolutionary population-based algorithms which are considered require the selection of an “elite” population in the evolutionary process. Simulation is applied to evaluate a set of candidate solutions in order to determine the elite subset, which is then used to reproduce a new set of candidate solutions for next iteration. In specific, such evolutionary methods can be summarized as follows.

**Step 0. Initialization**

Generate an initial set of  $k$  solutions

**Step 1. Simulation & Selection**

Simulate the set of  $k$  candidate solutions until a specified  $P\{CS\}$  level is met or a computing budget  $T$  is used. Then select the best top- $m$  solutions.

**Step 2. Stop or Not?**

If the stopping criterion is met, stop the algorithm. Otherwise, continue.

**Step 3. Reproduction**

The selected top- $m$  solutions from Step 1 are used to update the subsequent population or sampling distribution that drives the search for a new set of  $k$  candidate solutions for next iteration. Go to Step 1.

Since OCBA is much more efficient than other methods as shown in Section 3, we anticipate that the new OCBA-based evolutionary algorithm will be more efficient. The advantage of OCBA can be utilized in three ways in Step 1,

- For a same size of candidate solutions ( $k$ ) and same the  $P\{CS\}$  requirement, OCBA reduces the required computation cost.
- For a same size of candidate solutions ( $k$ ) and same amount of computing budget ( $T$ ), OCBA achieves a higher probability of correctly selecting top- $m$  solutions.
- For the same amount of computing budget ( $T$ ) and the same  $P\{CS\}$  requirement, OCBA can evaluate a bigger set of candidate solutions ( $k$ ) in one iteration, which leads to a higher overall convergence rate.

We are conducting research to evaluate the above three alternative options and will present our development and new observations at the workshop.

**ACKNOWLEDGMENTS**

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.

**REFERENCES**

- Andradottir, S., D. Goldsman, B. W. Schmeiser, L. W. Schruben, and E. Yücesan. 2005. Analysis Methodology: Are We Done?. *Proceedings of the 2005 Winter Simulation Conference*, pp. 790-796.
- Bechhofer, R.E., T.J. Santner, and D.M. Goldsman. 1995. Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons, John Wiley & Sons.
- Boesel, J., B.L. Nelson, and S.H. Kim. 2003. Using Ranking and Selection to 'Clean up' After Simulation Optimization. *Operations Research*, 51, 814-825.
- Buchholz, P. and A. Thümmler. 2005, Enhancing Evolutionary Algorithms with Statistical Selection Procedures for Simulation Optimization. *Proceedings of the Winter Simulation Conference*, 842-852.
- Chambers, L. 1995. Practical Handbook of Genetic Algorithms, CRC Press.
- Chen, C. H., J. Lin, E. Yücesan, and S. E. Chick. 2000. 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.
- Chen, E. J. and W. D. Kelton. 2000. An Enhanced Two-Stage Selection Procedure. *Proceedings of the Winter Simulation Conference*, pp. 727-735.
- Chen, H. C., C. H. Chen, L. Dai, and E. Yücesan. 1997. New Development of Optimal Computing Budget Allocation For Discrete Event Simulation. *Proceedings of the 1997 Winter Simulation Conference*, pp. 334-341.
- Chick, S. and K. Inoue. 2001. New Two-Stage and Sequential Procedures for Selecting the Best Simulated System. *Operations Research*, Vol. 49, pp. 1609-1624.
- Chick, S. and K. Inoue. 2001. New Procedures to Select the Best Simulated System Using Common Random Numbers. *Management Science*, 47(8), pp. 1133-1149.
- Dudewicz, E. J. and S. R. Dalal. 1975. Allocation of Observations in Ranking and Selection with Unequal Variances. *Sankhya*, B37, pp. 28-78.
- Fu, M. C., J. Q. Hu, C. H. Chen, and X. Xiong. 2006. Simulation Allocation for Determining the Best Design in the Presence of Correlated Sampling. *INFORMS Journal on Computing*, accepted for publication.
- Fu, M. C., J. Hu, and S. I. Marcus. 2006. Model-Based Randomized Methods for Global Optimization. *Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems*, 355-363.
- Goldsman, D. and B. L. Nelson. 1998. Comparing Systems via Simulation. J. Banks, ed. *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, John Wiley & Sons, New York, pp. 273-306.
- Gupta, S. S.. 1965. On Some Multiple Decision (Selection and Ranking) Rules. *Technometrics*, 7: 225-245.
- He, D., S. E. Chick, C. H. Chen. 2006. The Opportunity Cost and OCBA Selection Procedures in Ordinal Optimization. to appear in *IEEE Transactions on Systems, Man, and Cybernetics--Part C*.
- Holland, J. H.. 1975. Adaptation in Natural and Artificial Systems, The University of Michigan Press.
- Hu, J., M. C. Fu, and S. I. Marcus. 2006a. A Model Reference Adaptive Search Algorithm for Global Optimization. *Operations Research*, accepted for publication.
- Hu, J., M. C. Fu, and S. I. Marcus. 2006b. A Model Reference Adaptive Search Algorithm for Stochastic Global Optimization. working paper.

- Hyden, P. and L. Schruben. 2000. Improved Decision Processes Through Simultaneous Simulation and Time Dilation. *Proceedings of the 2000 Winter Simulation Conference*, pp. 743-748.
- Inoue, K., and S. E. Chick. 1998. Comparison of Bayesian and Frequentist Assessments of Uncertainty for Selecting the Best System. *Proceedings of the 1998 Winter Simulation Conference*, pp. 727-734.
- Koenig, L. W. and A. M. Law. 1985. 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.
- Law, A. M. and W. D. Kelton. 2000. *Simulation Modeling & Analysis*. McGraw-Hill, Inc..
- Lee, L. H. and E. P. Chew. 2003. A Simulation Study on Sampling and Selecting under Fixed Computing Budget. *Proceedings of 2003 Winter Simulation Conference*, pp. 535-542.
- Rubinstein, R.Y. and D.P. Kroese. 2004. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning*. Springer.
- Sullivan, D. W. and J. R. Wilson. 1989. Restricted Subset Selection Procedures for Simulation. *Operations Research*, 37:52-71.
- Swisher, J.R., S.H. Jacobson, and E. Yücesan. 2003. Discrete-Event Simulation Optimization Using Ranking, Selection, and Multiple Comparison Procedures: A Survey. *ACM Transactions on Modeling and Computer Simulation* 13, 134-154.
- Trailovic, L. and L. Y. Pao. 2004. Computing Budget Allocation for Efficient Ranking and Selection of Variances with Application to Target Tracking Algorithms. to appear in *IEEE Transactions on Automatic Control*.

## AUTHOR BIOGRAPHIES

**CHUN HUNG CHEN** is a professor of Systems Engineering and Operations Research at George Mason University. His e-mail address is [cchen9@gmu.edu](mailto:cchen9@gmu.edu).

**DONGHAI HE** is a Ph.D. student of Systems Engineering and Operations Research at George Mason University. His e-mail address is [dhe1@gmu.edu](mailto:dhe1@gmu.edu).

**MICHAEL FU** is a professor of Robert H. Smith School of Business and Institute for Systems Research at University of Maryland at College Park. His e-mail address is [mfu@wam.umd.edu](mailto:mfu@wam.umd.edu).

**LOO HAY LEE** is a professor of Department of Industrial & Systems Engineering at The National University of Singapore. His e-mail address is [iseleelh@nus.edu.sg](mailto:iseleelh@nus.edu.sg).