Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

BAYESIAN SUBSET SELECTION FOR NEAR-OPTIMAL SYSTEMS

Javier Gatica

Instituto de Ingeniería Matemática y Computacional
Pontificia Universidad Católica de Chile
Avda. Vicuña Mackenna 4860
Santiago, 7820436, CHILE

Jinbo Zhao
David J. Eckman

Industrial and Systems Engineering
Texas A&M University
3131 TAMU
College Station, TX 77843, USA

ABSTRACT

We study the ranking-and-selection problem of selecting a subset of simulated systems that with high probability contains a system with near-optimal performance. The posterior probability that at least one system in a given subset is near optimal – referred to as the posterior probability of good inclusion (pPGI) – can be expressed in terms of a sum of one-dimensional integrals and computed via numerical integration. Still, enumerating all possible subsets and computing their associated pPGI is impractical for large problem instances, thus we explore approximate solution methods. In particular, we investigate a greedy algorithm that builds a subset by iteratively adding the system that increases the pPGI the most.

1 INTRODUCTION

Research on how to incorporate the Bayesian statistical philosophy into ranking and selection (R&S) has predominantly focused on the problem of how best to allocate scarce simulation effort to maximize some criterion based on selecting a single system (Chen et al. 2015). Far less attention has been given to subset selection, which refers to the task of selecting a subset of simulated systems having some desired property. The Bayesian framework makes delivering statistical guarantees for R&S procedures convenient, but is often accompanied by computational challenges, especially as the number of systems grows. This is because Bayesian R&S procedures involve computing posterior quantities of interest, such as the posterior probability of correct selection. Quantities such as this can be expressed as integrals of the posterior distribution on the mean performances of all systems over some (possibly non-convex) region.

Suppose there are k systems under consideration and the mean performance of each system is unknown, but can be estimated via simulation. Systems are assumed to be simulated independently and endowed with independent prior distributions. Adopting a Bayesian perspective, we let W_i denote the mean performance of System i, for $i = 1, 2, \ldots, k$. Here, W_i is viewed as a random variable whose posterior distribution represents the remaining uncertainty after having obtained experimental evidence $\mathscr{E}$.

We explore the subset-selection problem of finding a subset of systems that contains one having near-optimal mean performance with high (posterior) probability. More specifically, given a subset of systems $S \subseteq \{1, 2, ..., k\}$ and user-specified tolerance $\delta \geq 0$, we define the posterior probability of good inclusion (pPGI) of the subset as $\text{pPGI}(S) := \Pr(\cup_{i \in S}\{W_i \geq W_{\max} - \delta\}|\mathscr{E})$, where larger performance is assumed to be better, $W_{\max} := \max_{i=1,2,\ldots,k} W_i$, and $\Pr(\cdot \mid \mathscr{E})$ denotes the posterior probability. Our goal is to find the smallest subset with a posterior probability of good inclusion exceeding $1-\alpha$, where $1-\alpha \in (1/k, 1)$ is a user-specified confidence level. In other words, we consider the combinatorial optimization problem

$$\min_{S \subseteq \{1,\ldots,k\}} |S| \quad \text{such that} \quad \text{pPGI}(S) \geq 1 - \alpha. \tag{1}$$

2 APPROACH

By conditioning on the identity and mean performance of the best system, the posterior probability of good inclusion of a given subset can be expressed in terms of a sum of one-dimensional integrals:

$$\text{pPGI}(S) = 1 - \sum_{i \in S^c} \int_{-\infty}^{\infty} \left[\prod_{j \in S} F_{W_j|\mathscr{E}}(w-\delta) \prod_{\ell \in S^c \setminus \{i\}} F_{W_\ell|\mathscr{E}}(w) \right] f_{W_i|\mathscr{E}}(w)\, dw. \tag{2}$$

A similar technique was applied in Eckman and Henderson (2022) to simplify the calculation of the posterior probability of good selection. In (2), $F_{W_i|\mathscr{E}}(\cdot)$ and $f_{W_i|\mathscr{E}}(\cdot)$ represent the cdf and pdf, respectively, of the marginal posterior distribution for the mean performance of System i; in the standard R&S setup, these are normal or Student's t distributions, depending on whether the true variances are known or unknown. While (2) is computationally tractable for many problem instances, as the number of systems increases the complexity and number of the integrals increases, as does the number of possible subsets. Thus, exhaustively enumerating all subsets and computing their pPGI is generally impractical.

We investigate a computationally cheap heuristic approach for solving Problem (1). For the related problem of finding the smallest subset that contains the best system with high probability, Eckman et al. (2020) showed that an optimal subset can be found via a simple greedy algorithm. Motivated by this, we inspect a greedy algorithm that iteratively builds a subset S by each time adding the system (not yet in S) that leads to the largest increase in the pPGI until the pPGI exceeds $1-\alpha$.

Conjecture 1 The subset returned by the greedy algorithm is an optimal solution to Problem (1).

The aforementioned increase in the pPGI from adding System i to a subset S can be expressed as a sum of one-dimensional integrals and has a similar computational cost to that of (2):

$$\begin{aligned}\Delta(i \mid S) = \text{pPGI}(S \cup \{i\}) - \text{pPGI}(S) &= \int_{-\infty}^{\infty} \left[\prod_{j \in S} F_{W_j|\mathscr{E}}(w-\delta) \prod_{\ell \in S^c \setminus \{i\}} F_{W_\ell|\mathscr{E}}(w) \right] f_{W_i|\mathscr{E}}(w)\, dw \\ &+ \sum_{\ell \in S^c \setminus \{i\}} \int_{-\infty}^{\infty} \left[\prod_{j \in S} F_{W_j|\mathscr{E}}(w-\delta) \prod_{m \in S^c \setminus \{i,\ell\}} F_{W_m|\mathscr{E}}(w) \left[F_{W_i|\mathscr{E}}(w) - F_{W_i|\mathscr{E}}(w-\delta) \right] \right] f_{W_\ell|\mathscr{E}}(w)\, dw. \end{aligned} \tag{3}$$

We conjecture that there exists a partial ordering of the pPGI increments of different systems in terms of their posterior means and variances, denoted by μ_i and ρ_i^2, respectively. Eckman and Henderson (2022) established a similar relationship for the posterior probability of good selection.

Conjecture 2 Suppose that the true variances were known. For any pair of Systems i and j satisfying $\mu_i \le \mu_j - \delta/2$ and $\rho_i^2 \le \rho_j^2$, $\Delta(i \mid S) \le \Delta(j \mid S)$ for all $S \subseteq \{1,2,\ldots,k\} \setminus \{i,j\}$.

If Conjecture 2 were to be proven true, it could be used to accelerate the greedy algorithm by decreasing the number of systems for which it needs to evaluate $\Delta(i \mid S)$ at each iteration.

3 NUMERICAL EXPERIMENTS

Experiments on random problem instances have so far failed to find counter-examples disproving Conjectures 1 and 2. Additional experiments and efforts to rigorously prove these two conjectures are ongoing.

REFERENCES

Chen, C.-H., S. E. Chick, L. H. Lee, and N. A. Pujowidianto. 2015. "Ranking and Selection: Efficient Simulation Budget Allocation". In *Handbook of Simulation Optimization*, edited by M. C. Fu, 45–80. New York: Springer.

Eckman, D. J., and S. G. Henderson. 2022. "Posterior-Based Stopping Rules for Bayesian Ranking-and-Selection Procedures". *INFORMS Journal on Computing* 34(3):1711–1728.

Eckman, D. J., M. Plumlee, and B. L. Nelson. 2020. "Revisiting Subset Selection". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. G. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 2972–2983. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.