Automatica 50 (2014) 1391–1400

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

An efficient simulation budget allocation method incorporating regression for partitioned domains ☆

Mark W. Brantley [a], Loo Hay Lee [b,c], Chun-Hung Chen [d,1], Jie Xu [d]

[a] Goose Point Analysis L.L.C., Covington, LA 70433, United States

[b] Department of Industrial and Systems Engineering, National University of Singapore, Singapore 119077, Singapore

[c] Logistics Engineering College, Shanghai Maritime University, China

[d] Department of Systems Engineering and Operations Research, George Mason University, Fairfax, VA 22030, United States

ARTICLE INFO

Article history:
Received 1 July 2012
Received in revised form
30 March 2013
Accepted 11 February 2014
Available online 13 April 2014

Keywords:
Simulation
Budget allocation
Regression

ABSTRACT

Simulation can be a very powerful tool to help decision making in many applications but exploring multiple courses of actions can be time consuming. Numerous ranking and selection (R&S) procedures have been developed to enhance the simulation efficiency of finding the best design. To further improve efficiency, one approach is to incorporate information from across the domain into a regression equation. However, the use of a regression metamodel also inherits some typical assumptions from most regression approaches, such as the assumption of an underlying quadratic function and the simulation noise is homogeneous across the domain of interest. To extend the limitation while retaining the efficiency benefit, we propose to partition the domain of interest such that in each partition the mean of the underlying function is approximately quadratic. Our new method provides approximately optimal rules for between and within partitions that determine the number of samples allocated to each design location. The goal is to maximize the probability of correctly selecting the best design. Numerical experiments demonstrate that our new approach can dramatically enhance efficiency over existing efficient R&S methods.

1. Introduction

Simulation optimization is a method to find a design consisting of a combination of input decision variable values of a simulated system that optimizes a particular output performance measure of the system. We propose to investigate stochastic problems on a discrete domain with a finite simulation budget consisting of runs conducted sequentially on a single computer. To assess the performance at a single design location on the domain, the uncertainty in the system performance measure requires multiple runs to obtain a good estimate of the performance measure.

When presented with a relatively small number of designs in the domain, the problem we consider is that of selecting the best design from among the finite number of choices. Ranking and Selection (R&S) procedures are statistical methods specifically developed to select the best design or a subset that contains the best design from a set of k competing design alternatives. Rinott (1978) developed two-stage procedures for selecting the best design or a design that is very close to the best system. Many researchers have extended this idea to more general R&S settings in conjunction with new developments (e.g., Bechhofer, Santner, & Goldsman, 1995).

To improve efficiency for R&S, several approaches have been explored for problems of selecting a single best design. Intuitively, to ensure a high probability of correct selection (PCS) of the best design, a larger portion of the computing budget should be allocated to those designs that are critical in the process of identifying the best design. A key consequence is the use of both the means and variances in the allocation procedures, rather than just the variances, as in Rinott (1978). Among examples of such approaches, the

☆ This work has been supported in part by National Science Foundation under Award CMMI-1233376, Department of Energy under Award DE-SC0002223, NIH under Grant 1R21DK088368-01, National Science Council of Taiwan under Award NSC-100-2218-E-002-027-MY3 and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Editor Berç Rüstem.

E-mail addresses: mark@goosepointanalysis.com (M.W. Brantley), iseleelh@nus.edu.sg (L.H. Lee), cchen9@gmu.edu (C.-H. Chen), jxu13@gmu.edu (J. Xu).

[1] Tel.: +1 703 9933572; fax: +1 703 9931521.

http://dx.doi.org/10.1016/j.automatica.2014.03.011
0005-1098/

Optimal Computing Budget Allocation (OCBA) approach by Chen, He, Fu, and Lee (2008), Chen, Lin, Yücesan, and Chick (2000), Lee et al. (2010) and Lee, Pujowidianto, Li, Chen, and Yap (2012) is the most relevant to this paper. OCBA maximizes a simple heuristic approximation of the *PCS*. The approach by Chick and Inoue (2001) estimates the *PCS* with Bayesian posterior distributions and allocates further samples using decision-theory tools to maximize the expected value of information in those samples. Branke, Chick, and Schmidt (2007) provide a nice overview and extensive comparison for some of relevant selection procedures.

Brantley, Lee, and Chen (2013) take an approach called optimal simulation design (OSD) that is different than most R&S methods by incorporating information from across the domain into a regression equation. Morrice, Brantley, and Chen (2008) extended the concepts from OSD to a method for selecting the best configuration based on a transient mean performance measure. Unlike traditional R&S methods, this regression based approach requires simulation of only a subset of the alternative design locations and so the simulation efficiency can be dramatically enhanced. While the use of a regression metamodel can dramatically enhance efficiency, the OSD method also inherits some typical assumptions from most DOE approaches. It is assumed that there is an underlying quadratic function for the means and the simulation noise is homogeneous across the domain of interest. Such assumptions are common in some of the DOE literature but become a limit for simulation optimization.

Motivated by iterative search methods (e.g., Newton's method in nonlinear programming) which rely upon a quadratic assumption only in a small local area of the search space during each iteration, we assume that we have several adjacent partitions and that in each partition the mean of the underlying function is approximately quadratic. Thus, we can utilize the efficiency benefit of a regression metamodel. From the perspective of simulation efficiency, we want to determine how to simulate each design point in the different partitions so that the overall simulation efficiency can be maximized.

Specifically, we want to determine (i) how much simulation budget to allocate to each partition; (ii) which design points in each partition must be simulated from the predetermined set of design points; (iii) how many replications should we simulate for those design points? This paper develops a Partitioning Optimal Simulation Design (POSD) method to address these issues. Numerical testing demonstrates that partitioning the domain and then efficiently allocating within the partitions can enhance simulation efficiency, even compared with some existing efficient R&S methods such as OCBA. By incorporating efficient allocations between the partitions in addition to efficient allocation within the partitions, the POSD method offers dramatic further improvements. As compared with only efficiently allocating within each partition, the POSD method offers an improvement over not only the well-known D-optimality approach in DOE literature (by 70%–74% reduction) but also the OSD method developed in Brantley et al. (2013) (by 55%–65% reduction). The rest of the paper is organized as follows. In Section 2, we introduce the simulation optimization problem setting and Bayesian framework. Section 3 develops an approximate *PCS* while Section 4 provides heuristic approximations of the optimal simulation allocations to maximize the approximate *PCS*. Numerical experiments comparing the results using the new partitioned OSD (POSD) method and other methods are provided in Section 5. Finally, Section 6 provides the conclusions and suggestions for future work using the concepts introduced here.

2. Problem setting and Bayesian framework

This paper explores a problem with the principal goal of selecting the best of multiple alternative design locations. Without loss of generality, we assume that we have m adjacent partitions and that each partition has k design locations. We aim to find the minimization problem shown below in (1) where the "best" design location is the one with smallest expected performance measure

$$\min_{x_{hi}} y(x_{hi}) = E\left[f(x_{hi})\right]; \\ x_{hi} \in \left[x_{11}, \ldots, x_{1k}, x_{21}, \ldots, x_{2k}, x_{m1}, \ldots, x_{mk}\right]. \tag{1}$$

Addressing how the domain is partitioned is not within the scope of this paper and we assume this partitioning scheme is derived from knowledge of the domain, through iterative refinement, or through an optimal selection procedure such as multi-variate adaptive regression splines (MARS) (Friedman, 1991).

In this paper, we consider that the expectation of the unknown underlying function for each partition is quadratic or approximately quadratic in nature on the prescribed domain, i.e., for each partition h,

$$y(x_{hi}) = \beta_{h0} + \beta_{h1}x_{hi} + \beta_{h2}x_{hi}^2. \tag{2}$$

For ease of notation, we define $\beta_h = [\beta_{h0}, \beta_{h1}, \beta_{h2}]$. In (2), the parameters β_h are unknown and we consider a common case where $y(x_{hi})$ must be estimated via simulation with noise. The simulation output $f(x_{hi})$ is independent from replication to replication such that

$$f(x_{hi}) = y(x_{hi}) + \varepsilon_h; \quad i = 1, \ldots, k, \ \varepsilon_h \sim N(0, \sigma_h^2). \tag{3}$$

The parameters β_h are unknown so $y(x_{hi})$ are also unknown. However, we can estimate expected performance measure at x_{hi}, that we define as $\hat{y}(x_{hi})$, by using a least squares estimate of the form shown in (4) below where $\hat{\beta}_{h0}$, $\hat{\beta}_{h1}$, and $\hat{\beta}_{h2}$ are the least squares parameter estimates for the corresponding parameters associated with the constant, linear, and quadratic terms in (2).

$$\hat{y}(x_{hi}) = \hat{\beta}_{h0} + \hat{\beta}_{h1}x_{hi} + \hat{\beta}_{h2}x_{hi}^2. \tag{4}$$

In a similar manner, we define $\hat{\beta}_h = [\hat{\beta}_{h0}, \hat{\beta}_{h1}, \hat{\beta}_{h2}]$. In order to obtain the least squares parameter estimates for each partition, we take n_h samples on any choice of x_{hi} (on at least three design locations for each partition to avoid singular solutions). We assume that these x_{hi} are given beforehand and we can only take samples from these points. Given the n_h samples, we define F_h as the n_h dimensional vector containing the replication output measures $f(x_{hi})$ and X_h as the $n_h \times 3$ matrix composed of rows consisting of $[1, x_{hi}, x_{hi}^2]$ with each row corresponding to its respective entry of $f(x_{hi})$ in F_h. Using the matrix notation and a superscript t to indicate the transpose of a matrix, for each partition we determine the least squares estimate for the parameters β_h which minimize the sum of the squares of the error terms $(F_h - X_h\beta_h)^t(F_h - X_h\beta_h)$. As shown in many regression texts, we obtain the least squares estimate for the parameters as $\hat{\beta}_h = (X_h^tX_h)^{-1}X_h^tF_h$.

Our problem is to select the design location associated with the smallest mean performance measure from among the mk design locations within the constraint of a computing budget with only T simulation replications. Given the least squares estimates for the parameters, we can use (4) to estimate the expected performance measure at each design location. We designate the design location with the smallest estimated mean performance measure in each partition as x_{hb} so that $\hat{y}(x_{hb}) = \min_i \hat{y}(x_{hi})$ and designate x_{Bb} as the design location with the smallest estimated mean performance measure across the entire domain so that $\hat{y}(x_{Bb}) = \min_h \hat{y}(x_{hb})$. Given the uncertainty of the estimate of the underlying function, x_{Bb} is a random variable and we define Correct Selection as the event where x_{Bb} is indeed the best location. We define N_{hi} as the number of simulation replications conducted at design location x_{hi}. Since the simulation is expensive and the computing budget is restricted, we seek to develop an allocation rule for each N_{hi}