

Efficient Simulation Budget Allocation for Subset Selection Using Regression Metamodels

Fei Gao^a, Zhongshun Shi^b, Siyang Gao^c, Hui Xiao^d

^a*Network Planning Department, SF Technology, Shenzhen 518052, China*

^b*Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA*

^c*Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong*

^d*School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, China*

Abstract

This research considers the ranking and selection (R&S) problem of selecting the optimal subset from a finite set of alternative designs. Given the total simulation budget constraint, we aim to maximize the probability of correctly selecting the top- m designs. In order to improve the selection efficiency, we incorporate the information from across the domain into regression metamodels. In this research, we assume that the mean performance of each design is approximately quadratic. To achieve a better fit of this model, we divide the solution space into adjacent partitions such that the quadratic assumption can be satisfied within each partition. Using the large deviation theory, we propose an approximately optimal simulation budget allocation rule in the presence of partitioned domains. Numerical experiments demonstrate that our approach can enhance the simulation efficiency significantly.

Key words: Simulation optimization; ranking and selection; OCBA; subset selection; regression.

1 Introduction

Discrete-event systems (DES) simulation has played an important role in analyzing modern complex systems and evaluating decision problems, since these systems are usually too difficult to be described using analytical models. DES simulation has been a common analysis method of choice and widely used in many practical applications, such as the queueing systems, electric power grids, air and land traffic control systems, manufacturing plants and supply chains (Xu et al., 2015, 2016; Gao and Chen, 2016). However, running the simulation model is usually time consuming, and a large number of simulation replications are typically required to achieve an accurate estimate of a design decision (Lee et al., 2010). In addition, it could be computationally quite expensive to select the best design(s) when the number of alternatives is relatively large.

In this paper, we consider the problem of selecting the optimal subset of the top- m designs out of t alternatives, where the performance of each design is estimated based on simulation. In order to improve the selection efficiency, we aim to intelligently allocate the simulation replications to each design to maximize the probability of correctly selecting all the top- m designs. This problem setting falls in the well-established statistics branch known as ranking and selection (R&S) (Xu et al., 2015).

In the literature, several types of efficient R&S procedures have been developed. The indifference-zone (IZ) approach allocates the simulation budget to provide a guaranteed lower bound for the probability of correct selection (PCS)

Email addresses: fei.gao5@sffmail.sf-express.com (Fei Gao), zhongshun.shi@wisc.edu (Zhongshun Shi), siyangao@cityu.edu.hk (Siyang Gao), msxh@swufe.edu.cn (Hui Xiao).

(Kim and Nelson, 2001). Chen et al. (2000) proposed an optimal computing budget allocation (OCBA) approach for R&S problems. The OCBA approach allocates the simulation replications sequentially in order to maximize PCS under a simulation budget constraint. He et al. (2007), Gao and Shi (2015) and Gao et al. (2017a) further developed the OCBA method with the expected opportunity cost (EOC) measure, which focuses more on the consequence of a wrong selection compared to PCS . Brantley et al. (2013) proposed another approach called optimal simulation design (OSD) to select the best design with regression metamodels. It assumes that all designs fit a single quadratic line and the variance of each design is identically distributed. The OSD approach was further extended in Brantley et al. (2014), Xiao et al. (2015) and Gao et al. (2018) to consider more general problems by dividing the solution space into adjacent partitions. Although these studies are also based on partitioning or metamodeling, they do not aim to select the top m designs, and are therefore different in objectives from this research. Other variants of OCBA include selecting the best design considering resource sharing and allocation (Peng et al., 2013) and input uncertainty (Gao et al., 2017b).

Most of the existing R&S procedures focus on identifying the best design and return a single choice as the estimated optimum. However, decision makers may prefer to have several good alternatives instead of one and make the final selection by considering some qualitative criteria, such as political feasibility and environmental consideration, which might be neglected by computer models (Gao and Chen, 2016; Zhang et al., 2016). The selection procedure providing the top- m designs can help the decision makers make their final decision in a more flexible way.

The literature on the optimal subset selection is sparse. Chen et al. (2008) and Zhang et al. (2016) considered the optimal subset selection problem using the OCBA framework, which maximizes PCS under a simulation budget constraint. In Gao and Chen (2015), an optimal subset selection procedure was proposed to minimize the measure of EOC . The optimal subset selection problem was further extended in Gao and Chen (2016) for general underlying distributions using the large deviations theory.

The aforementioned R&S procedures could smartly allocate the computing budget given the simulation results. They estimate the performance of each design only by considering the sample information of the considered design itself. However, the designs nearby could also provide useful information since neighboring designs usually have similar performance. Based on this idea, we aim to improve the selection efficiency by incorporating the information from across the domain into some response surfaces. Unlike traditional R&S methods, the regression based approaches require simulation experiments on only a subset of all the designs under consideration. The performances of the other designs can be inferred based on the sample information of the simulated designs. This provides us an effective way to further improve the efficiency for solving the subset selection problem, which is the motivation of this paper.

In this research, we assume the underlying function is quadratic or approximately quadratic. This assumption could help utilize the structure information of the design space and led to significant improvement of the computational efficiency. It is commonly used in the literature, such as Brantley et al. (2013, 2014); Xiao et al. (2015); McConnell and Servaes (1990). Based on this assumption we built a quadratic regression metamodel to incorporate the information from across the domain. The first contribution of this work is that we propose an asymptotically optimal allocation rule that determines which designs need to be simulated and the number of simulation budget allocated to them, such that the PCS of the optimal subset could be maximized. We call this procedure the optimal computing budget allocation for selecting the top- m designs with regression (OCBA-mr). In order to further extend the OCBA-mr procedure to more general cases where the underlying function is partially quadratic or non-quadratic, we divide the solution space into adjacent partitions and build a quadratic regression metamodel within each partition. The underlying function in each partition could be well approximated by a quadratic function if the solution space is properly partitioned or each partition is small enough. According to the results in Brantley et al. (2014); Xiao et al. (2015); Gao et al. (2018), the use of partitioned domains along with regression metamodels could significantly improve the simulation efficiency. That means interpolating the solution space can be an effective way for us to have further improvement. For different problems, the solution space could be divided into discrete partitions using different criteria, such as the size of corporations, the type of industries and the temperature of chemical process (Xiao et al., 2015). Based on the idea mentioned above, we develop an asymptotically optimal computing budget allocation procedure for selecting the top- m designs with regression in partitioned domains (OCBA-mrp), which is an extension of the OCBA-mr procedure for more general cases. In order to maximize the PCS of the optimal subset, the OCBA-mrp procedure not only determines the optimal simulation budget allocation within each partition but also determines the optimal budget allocation between partitions.

The rest of the paper is organized as follows. In Section 2, we formulate the optimal subset selection problem with regression metamodel and derive an asymptotically optimal simulation budget allocation rule, called OCBA-mr. Section 3 extends the OCBA-mr method for more general cases with partitioned domains and derives another

asymptotically optimal simulation budget allocation rule, called OCBA-mrp. The performance of the proposed methods is illustrated with numerical examples in Section 4. Section 5 concludes the paper.

2 Optimal Subset Selection Strategy

In this section, we provide an optimal computing budget allocation rule for the subset selection problem based on the regression metamodel.

2.1 Problem formulation

Without loss of generality, the best design is defined as the design with the smallest mean performance. We introduce the following notations:

- t : total number of designs;
- x_i : location of design i ;
- m : design with the m -th smallest mean value;
- $f(x_i)$: mean performance value for design i ;
- $Y(x_i)$: simulation output for design i ;
- β : vector of $(\beta_0, \beta_1, \beta_2)$, representing the coefficients of the regression model;
- $\hat{f}(x_i)$: estimate of $f(x_i)$ based on the regression model;
- S_o : set of the true top- m designs;
- S_n : set of designs not in S_o (complement of S_o);
- n : total number of simulation replications;
- n_i : number of simulation replications allocated to design i ;
- α : vector of $(\alpha_1, \alpha_s, \alpha_t)$, where α_r is the proportion of the total simulation budget n allocated to design r , i.e., $\alpha_r = n_r/n$, $r = 1, s, t$.

In the last bulleted item, r takes values only at 1, s and t , where designs 1 and t are the first and last designs in the solution space, and design s is an intermediate design determined by (5) (the rationale of it will be explained in more detail in Theorem 2). The problem we considered in this paper is to select the top- m designs out of the t alternatives by allocating simulation replications to these three designs. The optimal set is

$$S_o = \left\{ i : \max_{i \in S_o} f(x_i) < \min_{j \in S_n} f(x_j), i, j = 1, \dots, t \text{ and } |S_o| = m \right\},$$

where $|S|$ is the size of subset S .

We study the problem where the expected performance value across the solution space is quadratic or approximately quadratic in nature. Then, the mean performance of design i can be written as

$$f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2, \quad i = 1, \dots, t.$$

The coefficients $\beta = (\beta_0, \beta_1, \beta_2)$ are unknown beforehand, but can be estimated via simulation samples. Assume that the noises (ε) of the simulation experiments of design i follow normal distribution $N(0, \sigma^2)$. For each design, the noise is independent from replication to replication. The simulation output is $Y(x_i) = f(x_i) + \varepsilon(x_i)$ with $\varepsilon(x_i) \sim N(0, \sigma^2)$.

Given a total of n samples, we define \mathbf{Y} as the vector of the n simulation samples $Y(x_i)$ and \mathbf{X} be an $n \times 3$ matrix with each row $(1, x_i, x_i^2)$ corresponding to each entry $Y(x_i)$ of \mathbf{Y} . We estimate the coefficients β using the ordinary least squares (OLS) method (Hayashi, 2000) and denote the estimates as $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$. Then, we have $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ and $Cov[\hat{\beta}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$, where T denotes transposition and $\mathbf{X}^T \mathbf{X}$ is known as the information matrix (Kiefer, 1959). The estimate of $f(x_i)$ can be written as

$$\hat{f}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2, \quad i = 1, \dots, t. \tag{1}$$

We can use the equation (1) which incorporates the sample information of the simulated designs to estimate the expected performance for each design across the solution space. As $\hat{f}(x_i)$ is a linear combination of $\hat{\beta}$, we have $Var[\hat{f}(x_i)] = \sigma^2 \mathbf{X}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_i$, where $\mathbf{X}_i^T = (1, x_i, x_i^2)$.

Due to the uncertainty of the estimate of the underlying function, a correct selection of the optimal subset S_o may not always occur. Therefore, we introduce a measure, the probability of correct selection (PCS), to formulate the R&S problem considered in this paper. The PCS is given by $PCS = P\left\{ \bigcap_{i \in S_o} \bigcap_{j \in S_n} \hat{f}(x_i) \leq \hat{f}(x_j) \right\}$.

Given a fixed simulation budget, the optimization problem can be written as follows:

$$\begin{aligned} \max_{n_i} PCS &= P\left\{ \bigcap_{i \in S_o} \bigcap_{j \in S_n} \hat{f}(x_i) \leq \hat{f}(x_j) \right\} \\ \text{s.t.} \quad \sum_{i=1}^t n_i &= n. \end{aligned} \tag{2}$$

In this section, we aim to solve problem (2) where the mean performance of each design is estimated using the regression metamodel. Due to the uncertainty in the simulation experiments, multiple simulation replications are needed to generate the estimates of the underlying function $\hat{f}(x_i)$ accurately. The variance of $\hat{f}(x_i)$ is a function of the information matrix ($\mathbf{X}^T \mathbf{X}$) and can be reduced if additional simulation replications are conducted (Xiao et al., 2015). We are interested in how to intelligently allocate the simulation budget to proper designs such that $\hat{f}(x_i)$ can be better estimated using regression metamodel, and the PCS in problem (2) can be maximized.

2.2 Optimization model under large deviations framework

A major difficulty in solving problem (2) is that the objective function PCS does not have a closed-form expression. We seek to solve this optimization problem under an asymptotic framework in which the PCS is maximized or the probability of false selection ($PFS = 1 - PCS$) is minimized as n goes to infinity.

The PCS used in this paper is defined based on the quadratic regression model (1), which is constructed using the simulation information of only a fraction of the t designs. We call the designs receiving simulation replications the support designs. In order to construct the quadratic regression model, we need at least three support designs to obtain all of the information in $\mathbf{X}^T \mathbf{X}$ (Kiefer, 1959). For simplicity, in this research we let the number of support designs be three, two of which are at the extreme locations, i.e., x_1 and x_t (for this setting, see, e.g., Brantley et al. (2013, 2014); Xiao et al. (2015); Kiefer (1959)). The case with more than three designs can be similarly analyzed.

Lemma 1: Let $\hat{d}_{i,j} = \hat{f}(x_i) - \hat{f}(x_j)$, $i, j = 1, 2, \dots, t, i \neq j$. $\hat{d}_{i,j}$ follows normal distribution $N(f(x_i) - f(x_j), \varsigma_{ij})$, where

$$\begin{aligned} \varsigma_{ij} &= \sigma^2 (0, x_i - x_j, x_i^2 - x_j^2) (\mathbf{X}^T \mathbf{X})^{-1} \begin{pmatrix} 0 \\ x_i - x_j \\ x_i^2 - x_j^2 \end{pmatrix} \\ &= \frac{\sigma^2}{n} \left(\frac{\rho_{ij,1}^2}{\alpha_1} + \frac{\rho_{ij,s}^2}{\alpha_s} + \frac{\rho_{ij,t}^2}{\alpha_t} \right), \end{aligned}$$

and

$$\begin{cases} \rho_{ij,1} = \frac{(x_s - x_i)(x_t - x_i) - (x_s - x_j)(x_t - x_j)}{(x_1 - x_s)(x_1 - x_t)}, \\ \rho_{ij,s} = \frac{(x_1 - x_i)(x_t - x_i) - (x_1 - x_j)(x_t - x_j)}{(x_s - x_1)(x_s - x_t)}, \\ \rho_{ij,t} = \frac{(x_1 - x_i)(x_s - x_i) - (x_1 - x_j)(x_s - x_j)}{(x_t - x_1)(x_t - x_s)}. \end{cases}$$

The proof is similar to that is given in Eq.(22) of Brantley et al. (2013) (becomes the same when $m = 1$), and hence is omitted for brevity.

For any $i, j = 1, 2, \dots, t, i \neq j$, $\hat{d}_{i,j}$ is a normally distributed random variable. We can use the large deviations theory to derive the convergence rate function of the false selection probability.

Lemma 2: The convergence rate function of incorrect comparison probability for each design is:

$$\begin{aligned} & \left. \begin{aligned} & - \lim_{n \rightarrow \infty} \frac{1}{n} \log P\{\hat{f}(x_i) > \hat{f}(x_m)\}, i \in S_o, i \neq m \\ & - \lim_{n \rightarrow \infty} \frac{1}{n} \log P\{\hat{f}(x_i) < \hat{f}(x_m)\}, i \in S_n \end{aligned} \right\} \\ & = R_{m,i}(\boldsymbol{\alpha}) = \frac{(f(x_m) - f(x_i))^2/2}{\sigma^2 \left(\frac{\rho_{mi,1}^2}{\alpha_1} + \frac{\rho_{mi,s}^2}{\alpha_s} + \frac{\rho_{mi,t}^2}{\alpha_t} \right)}, \end{aligned}$$

Based on the results in Lemma 2, we can get an explicit expression of the convergence rate function of *PFS*.

Lemma 3: The convergence rate function of PFS is:

$$- \lim_{n \rightarrow \infty} \frac{1}{n} \log PFS = \min \left\{ \min_{\substack{i \in S_o \\ i \neq m}} R_{m,i}(\boldsymbol{\alpha}), \min_{j \in S_n} R_{m,j}(\boldsymbol{\alpha}) \right\}.$$

The main assertion of Lemma 3 is that the overall convergence rate of *PFS* is determined by the minimum convergence rate of the incorrect comparison for each design. Minimizing the *PFS* is asymptotically equivalent to maximizing the rate at which *PFS* goes to zero as a function of $\boldsymbol{\alpha}$, i.e., maximizing $-\lim_{n \rightarrow \infty} \frac{1}{n} \log PFS$. Based on Lemma 3, the asymptotical version of (2) becomes

$$\begin{aligned} & \max \min \left\{ \min_{i \in S_o, i \neq m} R_{m,i}(\boldsymbol{\alpha}), \min_{j \in S_n} R_{m,j}(\boldsymbol{\alpha}) \right\} \\ & \text{s.t. } \alpha_1 + \alpha_s + \alpha_t = 1 \\ & \alpha_1, \alpha_s, \alpha_t \geq 0. \end{aligned} \tag{3}$$

2.3 Asymptotically optimal solution

In this subsection, we seek to derive the optimality conditions for (3). Since the overall convergence rate of *PFS* is determined by the design with the minimum convergence rate. A false selection is most likely to happen at this key design. Therefore, it is enough for us to investigate the convergence rate of the key design across the solution space. We define i^* as the key design. That is

$$i^* = \arg \min_{i=1, \dots, t} \left\{ \min_{i \in S_o, i \neq m} R_{m,i}(\boldsymbol{\alpha}), \min_{i \in S_n} R_{m,i}(\boldsymbol{\alpha}) \right\}.$$

Theorem 1 *The optimization problem (2) can be asymptotically optimized with the following allocation rule:*

$$\alpha_r^* = \begin{cases} \frac{|\rho_{mi^*,r}|}{|\rho_{mi^*,1}| + |\rho_{mi^*,s}| + |\rho_{mi^*,t}|}, & r = 1, s, t, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The $\rho_{mi^*,r}$ is also known as the Lagrange interpolating polynomial coefficient (De la Garza et al., 1954; Burden and Faires, 2001). It represents the relative importance of each support design for estimating $\hat{d}_{mi} = \hat{f}(x_m) - \hat{f}(x_i)$. Theorem 1 indicates that the support design r will receive more simulation budget if it has larger $\rho_{mi^*,r}$.

Given the optimal allocation rule (Theorem 1), we next determine the optimal location for the support design s .

Theorem 2 *The rate function of PFS with allocation rule satisfying (4) can be maximized if the support design s satisfies the following equations.*

The support design

$$x_s = \begin{cases} x_{i^*} + x_m - x_1, & \frac{3x_1 + x_t}{4} \leq \frac{x_{i^*} + x_m}{2} < \frac{x_1 + x_t}{2}, \\ x_{i^*} + x_m - x_t, & \frac{x_1 + x_t}{2} < \frac{x_{i^*} + x_m}{2} \leq \frac{x_1 + 3x_t}{4}, \\ (x_1 + x_t)/2. & \text{otherwise.} \end{cases} \quad (5)$$

When the x_s derived from (5) does not correspond to any design available, we round it to the nearest one. The expression of Theorem 2 is similar to the results in Brantley et al. (2013). The differences are the selection of the key design x_i^* . We define x_i^* as the design with the minimum convergence rate of PFS, where a false selection of the optimal subset is most likely to happen.

3 Optimal Subset Selection Strategy for Partitioned Domains

The regression based method mentioned above can greatly improve the subset selection efficiency compared to the traditional methods. However, it is constrained with the typical assumptions such as the assumption of quadratic underlying function for the means. It is possible that the underlying function is neither quadratic nor approximately quadratic, so that we will fail to find the top- m designs. In order to extend our method to more general cases, we divide the solution space into adjacent partitions. The quadratic pattern can be expected when the solution space is properly partitioned or each partition is small enough.

3.1 Problem formulation

We first add the following notations for partitioned domains:

- l : number of partitions of the entire domain;
- k^h : number of designs in partition h , $h = 1, 2, \dots, l$;
- i_h : i th design in partition h , $i = 1, 2, \dots, k^h$ (when $i = k^h$, we denote design k^h as design k_h for notational simplicity);
- x_{i_h} : location of design i_h ;
- b : the partition containing the design with the m -th smallest mean value;
- m_b : design with the m -th smallest mean value;
- β_h : vector of $(\beta_{h0}, \beta_{h1}, \beta_{h2})$, representing the coefficients of the regression model in partition h ;
- $n_{h\bullet}$: number of simulation replications allocated to partition h ;
- n_{i_h} : number of simulation replications allocated to design i_h ;
- θ_h : the proportion of simulation budget allocated to partition h , i.e., $\theta_h = n_{h\bullet}/n$;
- α_h : vector of $(\alpha_{1_h}, \alpha_{s_h}, \alpha_{k_h})$, where α_{r_h} is the proportion of the simulation budget $n_{h\bullet}$ allocated to design r_h , i.e., $\alpha_{r_h} = n_{r_h}/n_{h\bullet}$, $r_h = 1_h, s_h, k_h$.

The notations $f(x_{i_h})$, $Y(x_{i_h})$ and $\hat{f}(x_{i_h})$ defined for partitioned domains are similar to those in Section 2, except the design i is replaced by i_h . The entire domain is divided into l adjacent partitions. Each partition contains k^h designs, i.e., there are $\sum_{h=1}^l k^h = t$ designs in total.

We assume there exists $\epsilon \in \mathbb{R}$ such that there are exactly m designs from the total t designs with mean performances less than ϵ and the rest $t - m$ designs with mean performances greater than ϵ . It ensures that the optimal subset is well defined and can be distinguished. We define the optimal subset S_o as

$$S_o = \left\{ i_h : \max_{i_h \in S_o} f(x_{i_h}) < \min_{j_g \in S_n} f(x_{j_g}), i_h = 1_h, \dots, k_h, \right. \\ \left. j_g = 1_g, \dots, k_g \quad h, g = 1, \dots, l \text{ and } |S_o| = m \right\}.$$

In this section, we assume that the expected performance value in each partition is quadratic or approximately quadratic when the solution space is properly partitioned and the mean performance within a partition is continuous and smooth. For this problem setting the *PCS* is defined as

$$PCS = P\left\{\bigcap_{i_h \in S_o} \bigcap_{j_g \in S_n} \hat{f}(x_{i_h}) \leq \hat{f}(x_{j_g})\right\}.$$

Given a fixed simulation budget, the optimization problem can be written as follows:

$$\begin{aligned} & \max PCS \\ & \text{s.t. } \sum_{h=1}^l \sum_{i_h=1}^{k_h} n_{i_h} = n. \end{aligned} \tag{6}$$

In this section, we aim to solve problem (6) in the presence of regression metamodels. We are interested in how to intelligently allocate the simulation budget to proper designs such that $\hat{f}(x_{i_h})$ can be better estimated using regression metamodels, and the *PCS* in problem (6) can be maximized. Note that this model does not require all the designs to be on the same axis, and therefore does not hinder it from being applied for multi-dimensional problems. For multi-dimensional problems, we can treat the range of the underlying function on each dimension as one or more partitions, and then apply this formulation.

3.2 Optimization model under large deviations framework

In order to solve the optimization problem (6), one challenge is how to derive an explicit expression of the *PCS*. We seek to solve (6) under an asymptotic framework in which the probability of false selection ($PFS = 1 - PCS$) is minimized as n goes to infinity.

Similar to the setting in Section 2, we let the number of support designs in each partition be three, two of which are at the extreme locations, i.e., x_{1_h} and x_{k_h} . We have $\hat{d}_{ij_h} = \hat{f}(x_{i_h}) - \hat{f}(x_{j_h})$ follow normal distribution $N(f(x_{i_h}) - f(x_{m_h}), \varsigma_{ij_h})$ and $\hat{d}_{i_h, j_g} = \hat{f}(x_{i_h}) - \hat{f}(x_{j_g})$ follow normal distribution $N(f(x_{i_h}) - f(x_{j_g}), \varsigma_{i_h, j_g})$ where $h \neq g$.

Similar to the proof of Lemma 1, we have

$$\begin{aligned} \varsigma_{ij_h} &= \sigma_h^2(0, x_{i_h} - x_{j_h}, x_{i_h}^2 - x_{j_h}^2)(\mathbf{X}_h^T \mathbf{X}_h)^{-1} \begin{pmatrix} 0 \\ x_{i_h} - x_{j_h} \\ x_{i_h}^2 - x_{j_h}^2 \end{pmatrix} \\ &= \frac{\sigma_b^2}{\theta_h n} \left(\frac{\rho_{ij_h,1}^2}{\alpha_{1_h}} + \frac{\rho_{ij_h,s}^2}{\alpha_{s_h}} + \frac{\rho_{ij_h,k}^2}{\alpha_{k_h}} \right), \\ \begin{cases} \rho_{ij_h,1} &= \frac{(x_{s_h} - x_{i_h})(x_{k_h} - x_{i_h}) - (x_{s_h} - x_{j_h})(x_{k_h} - x_{j_h})}{(x_{1_h} - x_{s_h})(x_{1_h} - x_{k_h})}, \\ \rho_{ij_h,s} &= \frac{(x_{1_h} - x_{i_h})(x_{k_h} - x_{i_h}) - (x_{1_h} - x_{j_h})(x_{k_h} - x_{j_h})}{(x_{s_h} - x_{1_h})(x_{s_h} - x_{k_h})}, \\ \rho_{ij_h,k} &= \frac{(x_{1_h} - x_{i_h})(x_{s_h} - x_{i_h}) - (x_{1_h} - x_{j_h})(x_{s_h} - x_{j_h})}{(x_{k_h} - x_{1_h})(x_{k_h} - x_{s_h})}, \end{cases} \end{aligned}$$

and

$$\begin{aligned} \varsigma_{i_h, j_g} &= \sigma_h^2(0, x_{i_h} - x_{j_g}, x_{i_h}^2 - x_{j_g}^2)(\mathbf{X}_h^T \mathbf{X}_h)^{-1} \begin{pmatrix} 0 \\ x_{i_h} - x_{j_g} \\ x_{i_h}^2 - x_{j_g}^2 \end{pmatrix} \\ &= \frac{\sigma_h^2}{\theta_h n} \left(\frac{\eta_{i_h,1}^2}{\alpha_{1_b}} + \frac{\eta_{i_h,s}^2}{\alpha_{s_b}} + \frac{\eta_{i_h,k}^2}{\alpha_{k_b}} \right) + \frac{\sigma_g^2}{\theta_g n} \left(\frac{\eta_{j_g,1}^2}{\alpha_{1_h}} + \frac{\eta_{j_g,s}^2}{\alpha_{s_h}} + \frac{\eta_{j_g,k}^2}{\alpha_{k_h}} \right), \end{aligned}$$

$$\begin{cases} \eta_{i_h,1} = \frac{(x_{s_h} - x_{i_h})(x_{k_h} - x_{i_h})}{(x_{1_h} - x_{s_h})(x_{1_h} - x_{k_h})}, \\ \eta_{i_h,s} = \frac{(x_{1_h} - x_{i_h})(x_{k_h} - x_{i_h})}{(x_{s_h} - x_{1_h})(x_{s_h} - x_{k_h})}, \\ \eta_{i_h,k} = \frac{(x_{1_h} - x_{i_h})(x_{s_h} - x_{i_h})}{(x_{k_h} - x_{1_h})(x_{k_h} - x_{s_h})}. \end{cases}$$

For any $h = 1, 2, \dots, l$ and $i_h = 1_h, 2_h, \dots, k_h$, $\hat{f}(x_{i_h})$ is a normally distributed random variable. We can use the large deviations theory to derive the convergence rate function of the false selection probability.

Lemma 4: The convergence rate functions of incorrect comparison probability for each design are provided as follows:

$$\begin{aligned} & \left. \begin{aligned} & - \lim_{n \rightarrow \infty} \frac{1}{n} \log P\{\hat{f}(x_{i_h}) > \hat{f}(x_{m_b})\}, i_h \in S_o \\ & - \lim_{n \rightarrow \infty} \frac{1}{n} \log P\{\hat{f}(x_{i_h}) < \hat{f}(x_{m_b})\}, i_h \in S_n \end{aligned} \right\} = R_{m_b, i_h}(\theta_b, \theta_h, \alpha_b, \alpha_h) \\ & = \frac{(f(x_{m_b}) - f(x_{i_h}))^2 / 2}{\frac{\sigma_b^2}{\theta_b} \left(\frac{\eta_{m_b,1}^2}{\alpha_{1_b}} + \frac{\eta_{m_b,s}^2}{\alpha_{s_b}} + \frac{\eta_{m_b,k}^2}{\alpha_{k_b}} \right) + \frac{\sigma_h^2}{\theta_h} \left(\frac{\eta_{i_h,1}^2}{\alpha_{1_h}} + \frac{\eta_{i_h,s}^2}{\alpha_{s_h}} + \frac{\eta_{i_h,k}^2}{\alpha_{k_h}} \right)}, h \neq b, \end{aligned} \quad (7)$$

and

$$\begin{aligned} & \left. \begin{aligned} & - \lim_{n \rightarrow \infty} \frac{1}{n} \log P\{\hat{f}(x_{i_b}) > \hat{f}(x_{m_b})\}, i_b \in S_o, i_b \neq m_b \\ & - \lim_{n \rightarrow \infty} \frac{1}{n} \log P\{\hat{f}(x_{i_b}) < \hat{f}(x_{m_b})\}, i_b \in S_n \end{aligned} \right\} = R_{m_b, i_b}(\theta_b, \alpha_b) \\ & = \frac{(f(x_{m_b}) - f(x_{i_b}))^2 / 2}{\frac{\sigma_b^2}{\theta_b} \left(\frac{\rho_{m_b,1}^2}{\alpha_{1_b}} + \frac{\rho_{m_b,s}^2}{\alpha_{s_b}} + \frac{\rho_{m_b,k}^2}{\alpha_{k_b}} \right)}. \end{aligned} \quad (8)$$

According to the Bonferroni inequality, we have $PFS \leq \sum_{\substack{i_h \in S_o \\ i_h \neq m_b}} P\{\hat{f}(x_{i_h}) \geq \hat{f}(x_{m_b})\} + \sum_{j_g \in S_n} P\{\hat{f}(x_{j_g}) \leq \hat{f}(x_{m_b})\}$. The PFS is bounded below by

$$\begin{aligned} & \max \left\{ \max_{\substack{1 \leq h \leq l \\ h \neq b}} \left\{ \max_{i_h \in S_o} P\{\hat{f}(x_{i_h}) \geq \hat{f}(x_{m_b})\}, \right. \right. \\ & \quad \left. \max_{j_h \in S_n} P\{\hat{f}(x_{j_h}) \leq \hat{f}(x_{m_b})\} \right\}, \\ & \max \left\{ \max_{\substack{i_b \in S_o \\ i_b \neq m_b}} P\{\hat{f}(x_{i_b}) \geq \hat{f}(x_{m_b})\}, \right. \\ & \quad \left. \max_{j_b \in S_n} P\{\hat{f}(x_{j_b}) \leq \hat{f}(x_{m_b})\} \right\} \end{aligned}$$

and bounded above by

$$\begin{aligned} & |S_o| \times |S_n| \times \max \left\{ \max_{\substack{1 \leq h \leq l \\ h \neq b}} \left\{ \max_{i_h \in S_o} P\{\hat{f}(x_{i_h}) \geq \hat{f}(x_{m_b})\}, \right. \right. \\ & \quad \left. \max_{j_h \in S_n} P\{\hat{f}(x_{j_h}) \leq \hat{f}(x_{m_b})\} \right\}, \\ & \max \left\{ \max_{\substack{i_b \in S_o \\ i_b \neq m_b}} P\{\hat{f}(x_{i_b}) \geq \hat{f}(x_{m_b})\}, \right. \\ & \quad \left. \max_{j_b \in S_n} P\{\hat{f}(x_{j_b}) \leq \hat{f}(x_{m_b})\} \right\} \end{aligned}$$

Therefore, according to the results in Lemma 4, the convergence rate functions of PFS is given by

$$\begin{aligned}
 - \lim_{n \rightarrow \infty} \frac{1}{n} \log PFS &= \min \left\{ \min_{\substack{1 \leq h \leq l \\ h \neq b}} \left\{ \min_{i_h \in S_o} R_{m_b, i_h}(\theta_b, \theta_h, \boldsymbol{\alpha}_b, \boldsymbol{\alpha}_h), \right. \right. \\
 &\quad \left. \left. \min_{j_h \in S_n} R_{m_b, j_h}(\theta_b, \theta_h, \boldsymbol{\alpha}_b, \boldsymbol{\alpha}_h) \right\}, \right. \\
 &\quad \left. \min \left\{ \min_{\substack{i_b \in S_o \\ i_b \neq m_b}} R_{m_b, i_b}(\theta_b, \boldsymbol{\alpha}_b), \min_{j_b \in S_n} R_{m_b, j_b}(\theta_b, \boldsymbol{\alpha}_b) \right\} \right\},
 \end{aligned}$$

Minimizing the PFS is asymptotically equivalent to maximizing the rate at which PFS goes to zero as a function of θ_h and $\boldsymbol{\alpha}_h$, $h = 1, 2, \dots, l$. Similarly, the asymptotical version of (6) becomes

$$\begin{aligned}
 \max \min &\left\{ \min_{\substack{1 \leq h \leq l \\ h \neq b}} \left\{ \min_{i_h \in S_o} R_{m_b, i_h}(\theta_b, \theta_h, \boldsymbol{\alpha}_b, \boldsymbol{\alpha}_h), \right. \right. \\
 &\quad \left. \left. \min_{j_h \in S_n} R_{m_b, j_h}(\theta_b, \theta_h, \boldsymbol{\alpha}_b, \boldsymbol{\alpha}_h) \right\}, \right. \\
 &\quad \left. \min \left\{ \min_{\substack{i_b \in S_o \\ i_b \neq m_b}} R_{m_b, i_b}(\theta_b, \boldsymbol{\alpha}_b), \min_{j_b \in S_n} R_{m_b, j_b}(\theta_b, \boldsymbol{\alpha}_b) \right\} \right\} \tag{9} \\
 \text{s.t. } &\alpha_{1_h} + \alpha_{s_h} + \alpha_{k_h} = 1, \quad h = 1, \dots, l, \\
 &\sum_{h=1}^l \theta_h = 1, \\
 &\theta_h, \alpha_{1_h}, \alpha_{s_h}, \alpha_{k_h} \geq 0, \quad h = 1, \dots, l.
 \end{aligned}$$

3.3 Asymptotically optimal solution

In this section, we seek to derive the optimality conditions for (9). We want to determine (i) the number of simulation replication allocated to each partition, (ii) the locations of the designs should be simulated in each partition and (iii) the number of simulation replications allocated to those selected designs.

According to Xiao et al. (2015), we have $\theta_h/\theta_b \rightarrow 0$ as the number of partitions l goes to infinity. That means the fraction of the simulation budget allocated to the partition b containing the m -th smallest mean value far exceeds the fraction given to any other partition when l goes to infinity. Given that, the rate function $R_{m_b, i_h}(\theta_b, \theta_h, \boldsymbol{\alpha}_b, \boldsymbol{\alpha}_h)$ converges to $\tilde{R}_{m_b, i_h}(\theta_h, \boldsymbol{\alpha}_h)$ where

$$\begin{aligned}
 \tilde{R}_{m_b, i_h}(\theta_h, \boldsymbol{\alpha}_h) &= \lim_{l \rightarrow \infty} R_{m_b, i_h}(\theta_b, \theta_h, \boldsymbol{\alpha}_b, \boldsymbol{\alpha}_h) \\
 &= \frac{(f(x_{m_b}) - f(x_{i_h}))^2/2}{\frac{\sigma_h^2}{\theta_h} \left(\frac{\eta_{i_h, 1}^2}{\alpha_{1_h}} + \frac{\eta_{i_h, s}^2}{\alpha_{s_h}} + \frac{\eta_{i_h, k}^2}{\alpha_{k_h}} \right)}.
 \end{aligned}$$

The problem (9) can be asymptotically rewritten as

$$\begin{aligned}
& \max \min \left\{ \min_{\substack{1 \leq h \leq l \\ h \neq b}} \left\{ \min_{i_h \in S_o} \tilde{R}_{m_b, i_h}(\theta_h, \boldsymbol{\alpha}_h), \min_{j_h \in S_n} \tilde{R}_{m_b, j_h}(\theta_h, \boldsymbol{\alpha}_h) \right\}, \right. \\
& \quad \left. \min \left\{ \min_{\substack{i_b \in S_o \\ i_b \neq m_b}} R_{m_b, i_b}(\theta_b, \boldsymbol{\alpha}_b), \min_{j_b \in S_n} R_{m_b, j_b}(\theta_b, \boldsymbol{\alpha}_b) \right\} \right\} \\
& \text{s.t. } \alpha_{1_h} + \alpha_{s_h} + \alpha_{k_h} = 1, h = 1, \dots, l \\
& \quad \sum_{h=1}^l \theta_h = 1 \\
& \quad \theta_h, \alpha_{1_h}, \alpha_{s_h}, \alpha_{k_h} \geq 0, h = 1, \dots, l.
\end{aligned} \tag{10}$$

In order to better analyze problem (10) above, we decompose it as follows:

$$\begin{aligned}
& \max \min_{\substack{1 \leq h \leq l \\ h \neq b}} \left\{ \min_{i_h \in S_o} \tilde{R}_{m_b, i_h}(\theta_h, \boldsymbol{\alpha}_h), \min_{j_h \in S_n} \tilde{R}_{m_b, j_h}(\theta_h, \boldsymbol{\alpha}_h) \right\} \\
& \text{s.t. } \alpha_{1_h} + \alpha_{s_h} + \alpha_{k_h} = 1, \alpha_{1_h}, \alpha_{s_h}, \alpha_{k_h} \geq 0,
\end{aligned} \tag{11}$$

for $h = 1, 2, \dots, l, h \neq b$, and

$$\begin{aligned}
& \max \min \left\{ \min_{\substack{i_b \in S_o \\ i_b \neq m_b}} R_{m_b, i_b}(\theta_b, \boldsymbol{\alpha}_b), \min_{j_b \in S_n} R_{m_b, j_b}(\theta_b, \boldsymbol{\alpha}_b) \right\} \\
& \text{s.t. } \alpha_{1_b} + \alpha_{s_b} + \alpha_{k_b} = 1, \alpha_{1_b}, \alpha_{s_b}, \alpha_{k_b} \geq 0,
\end{aligned} \tag{12}$$

for $h = b$.

Lemma 5: Let θ_h^ and $\boldsymbol{\alpha}_h^*$ be the optimal solution to (10). The θ_h^* and $\boldsymbol{\alpha}_h^*$ are independent and can be solved separately. In addition, the l $\boldsymbol{\alpha}_h^*$'s ($h = 1, 2, \dots, l$) corresponding to the l partitions are also mutually independent and can be solved separately using (11) and (12).*

Similar to Section 2, we define a key design of each partition h , denoted as i_h^* . A false selection is most likely to happen at these key designs. That is

$$i_h^* = \begin{cases} \arg \min_{1_h \leq i_h \leq k_h} \left\{ \min_{i_h \in S_o} R_{m_b, i_h}(\theta_b, \theta_h, \boldsymbol{\alpha}_b, \boldsymbol{\alpha}_h), \right. \\ \quad \left. \min_{i_h \in S_n} R_{m_b, i_h}(\theta_b, \theta_h, \boldsymbol{\alpha}_b, \boldsymbol{\alpha}_h) \right\}, h \neq b, \\ \arg \min_{1_b \leq i_b \leq k_b} \left\{ \min_{\substack{i_b \in S_o \\ i_b \neq m_b}} R_{m_b, i_b}(\theta_b, \boldsymbol{\alpha}_b), \right. \\ \quad \left. \min_{i_b \in S_n} R_{m_b, i_b}(\theta_b, \boldsymbol{\alpha}_b) \right\}, h = b. \end{cases}$$

3.3.1 Determine $\boldsymbol{\alpha}_h^*$ and locations of support designs

Given Lemma 5, we can determine $\boldsymbol{\alpha}_h^*$ separately for each partition by solving the optimization problems (11) and (12).

Theorem 3 *The optimization problem (6) can be asymptotically optimized with the following allocation rule:*

$$\alpha_{r_h}^* = \begin{cases} \frac{|z_{r_h}|}{|z_{1_h}| + |z_{s_h}| + |z_{k_h}|}, & r_h = 1_h, s_h, k_h, \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

where

$$z_{r_h}^2 = \begin{cases} \rho_{m_i^*, r}^2, & h=b, \\ \eta_{i_h^*, r}^2, & \text{otherwise,} \end{cases}$$

for $r_h = 1_h, s_h, k_h$.

Given the optimal allocation rule (Theorem 3), we next determine the optimal location for the support design s_h within each partition.

Theorem 4 *The rate function of PFS with allocation rule satisfying (13) can be asymptotically maximized if the support design s_h satisfies the following equations.*

(a) For all $h \neq b$, support design

$$x_{s_h} = \begin{cases} x_{i_h^*}, & i_h^* \neq 1_h, k_h, \\ (x_{1_h} + x_{k_h})/2, & i_h^* = 1_h, k_h, \end{cases} \quad (14)$$

and the optimal allocation rule (13) becomes

$$\alpha_{r_h}^* = \begin{cases} 1, & r_h = i_h^*, \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

(b) For $h = b$, support design

$$x_{s_b} = \begin{cases} x_{i_b^*} + x_{m_b} - x_{1_b}, & \frac{3x_{1_b} + x_{k_b}}{4} \leq \frac{x_{i_b^*} + x_{m_b}}{2} < \frac{x_{1_b} + x_{k_b}}{2}, \\ x_{i_b^*} + x_{m_b} - x_{k_b}, & \frac{x_{1_b} + x_{k_b}}{2} < \frac{x_{i_b^*} + x_{m_b}}{2} \leq \frac{x_{1_b} + 3x_{k_b}}{4}, \\ (x_{1_b} + x_{k_b})/2, & \text{otherwise,} \end{cases} \quad (16)$$

and the approximate optimal allocation rule (13) is

$$\alpha_{r_b}^* = \begin{cases} \frac{|\rho_{m_i^*, r}|}{|\rho_{m_i^*, 1}| + |\rho_{m_i^*, s}| + |\rho_{m_i^*, k}|}, & r_b = 1_b, s_b, k_b, \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

When the x_{s_h} derived from (14) and (16) does not correspond to any design available, we round it to the nearest one.

3.3.2 Determine θ_h^*

In this subsection, we aim to determine the optimal budget allocation rules between the l partitions. Since α_h^* and θ_h^* , can be solved separately for each partition, $R_{m_b, i_h}(\theta_b, \theta_h, \alpha_b^*, \alpha_h^*)$ is a function of θ_b and θ_h only. Similarly, $R_{m_b, i_b}(\theta_b, \alpha_b^*)$ is a function of θ_b only. Optimization problem (9) can be rewritten as

$$\begin{aligned} \max \min & \left\{ \min_{\substack{1 \leq h \leq l \\ h \neq b}} \left\{ \min_{i_h \in S_o} R_{m_b, i_h}(\theta_b, \theta_h, \alpha_b^*, \alpha_h^*), \right. \right. \\ & \left. \left. \min_{j_h \in S_n} R_{m_b, j_h}(\theta_b, \theta_h, \alpha_b^*, \alpha_h^*) \right\}, \right. \\ & \left. \min \left\{ \min_{\substack{i_b \in S_o \\ i_b \neq m_b}} R_{m_b, i_b}(\theta_b, \alpha_b^*), \min_{j_b \in S_n} R_{m_b, j_b}(\theta_b, \alpha_b^*) \right\} \right\} \\ \text{s.t.} & \sum_{h=1}^l \theta_h = 1, \theta_h \geq 0, h = 1, \dots, l. \end{aligned} \quad (18)$$

By solving the optimization model (18), we can get the optimal θ_h^* for all $h = 1, \dots, l$.

Theorem 5 *The optimal allocation that asymptotically minimizes the PFS for the problem (6) is that $\theta_h^* = \frac{\gamma_h^*}{\sum_{i=1}^l \gamma_i^*}$,*

where $\gamma_b^* = \sigma_b \sqrt{\left(\frac{\eta_{m_b,1}^2}{\alpha_{1_b}^*} + \frac{\eta_{m_b,s}^2}{\alpha_{s_b}^*} + \frac{\eta_{m_b,k}^2}{\alpha_{k_b}^*}\right) \sum_{h \neq b} \frac{\gamma_h^{*2}}{\sigma_h^2}}$ and $\gamma_h^* = \sigma_h^2 / (f(x_{m_b}) - f(x_{i_h^*}))^2$, $h \neq b$.

The results of Theorem 5 indicate the optimal simulation budget allocation between the l partitions. The results of Theorem 4 determine which designs should be selected for simulation and the number of simulation replications allocated to these selected designs. Since the solution space is divided into adjacent partitions, this optimal budget allocation procedure can be used for more general cases where the underlying function is non-quadratic. By conducting simulation replications on fewer designs, the simulation efficiency is dramatically enhanced compared to the existing methods.

4 Sequential Budget Allocation Algorithm

Based on the discussion above, we propose two sequential budget allocation algorithm to implement the optimality conditions. They are Optimal Optimal Computing Budget Allocation for selecting the top- m designs with regression (OCBA-mr) based on Theorems 1 and 2 and Optimal Optimal Computing Budget Allocation for selecting the top- m designs with regression in partitioned domains (OCBA-mrp) based on Theorems 4 and 5.

(1) OCBA-mr Algorithm

INITIALIZE: $\kappa \leftarrow 0$; Perform n_0 simulation replications for three design locations in each partition; by convention we use the D-optimal support designs, i.e., $n_1^\kappa = n_{((1+t)/2)}^\kappa = n_t^\kappa = n_0$ (round as needed). The incremental simulation budget is Δ .

LOOP: WHILE $\sum_{i=1}^t n_i^\kappa < n$ **DO**

UPDATE: Estimate a quadratic regression equation $\hat{f}(x_i)$ using the least squares method based on the information from all prior simulation runs.

Calculate the mean and variance of each design using $\hat{f}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$.

Determine the observed subset \hat{S}_o and \hat{S}_n ; the design with the m -th smallest sample mean value; the observed key design \hat{i}^* .

Calculate $\hat{\alpha}^*$ and determine the support designs using Theorems 1 and 2.

ALLOCATE: Increase the simulation budget by Δ and calculate the new budget allocation rule using $\hat{\alpha}^*$.

SIMULATE: Perform $\max(0, n_i^\kappa - n_i^{\kappa-1})$ simulation replications for design i and $i = 1, s, t$. $\kappa \leftarrow \kappa + 1$.

END OF LOOP.

(2) OCBA-mrp Algorithm

INITIALIZE: $\kappa \leftarrow 0$; Perform n_0 simulation replications for three design locations in each partition; by convention we use the D-optimal support designs, i.e., $n_{1_h}^\kappa = n_{((1+k(h))/2)_h}^\kappa = n_{k_h}^\kappa = n_0$ (round as needed). The incremental simulation budget is Δ .

LOOP: WHILE $\sum_{h=1}^l \sum_{i_h=1}^{k_h} n_{i_h}^\kappa < n$ **DO**

UPDATE: Estimate the quadratic regression equations $\hat{f}(x_{i_h})$ for each partition using the least squares method based on the information from all prior simulation runs.

Calculate the means and variances of each design using $\hat{f}(x_{i_h}) = \hat{\beta}_{h0} + \hat{\beta}_{h1}x_{i_h} + \hat{\beta}_{h2}x_{i_h}^2$, $i_h = 1_h, 2_h, \dots, k_h, h = 1, 2, \dots, l$.

Determine the observed subset \hat{S}_o and \hat{S}_n ; the design with the m -th smallest sample mean value; the observed key design \hat{i}_h^* of each partition.

Calculate $\hat{\alpha}_h^*$ and determine the support designs in each partition using Theorem 4. Calculate $\hat{\theta}_h^*$ using Theorem 5.

ALLOCATE: Increase the simulation budget by Δ and calculate the new budget allocation rule using $\hat{\alpha}_h^*$ and $\hat{\theta}_h^*$.

SIMULATE: Perform $\max(0, n_{i_h}^\kappa - n_{i_h}^{\kappa-1})$ simulation replications for design i_h in each partition $h = 1, 2, \dots, l$ and $i_h = 1_h, s_h, k_h$. $\kappa \leftarrow \kappa + 1$.

END OF LOOP.

5 Numerical Experiments

In this section, we test our proposed simulation budget allocation rules, OCBA-mr and OCBA-mrp, with some existing methods on several typical optimal subset selection problems. We use the following three budget allocation approaches for comparison. The top- m designs are selected based on their mean values.

- *Equal Allocation (EA):* The EA is the most commonly used and simplest method, which allocates the simulation budget equally to each of the design.
- *OCBAm+ Allocation:* The OCBAm+ is an efficient simulation budget allocation procedure for selecting the top m designs. It was developed based on the OCBA framework. For detail see Zhang et al. (2016).
- *OCBA-ss Allocation:* The OCBA-ss procedure was proposed in Gao and Chen (2016) and seek to solve the optimal subset R&S problems for general underlying distributions using the large deviations theory.
- *OSD Allocation:* The OSD procedure was proposed in Brantley et al. (2013) and seek to solve the single best R&S problems using regression metamodel.

The OCBA-mrp, OCBA-mr and OSD estimate the performance of each design based on some regression metamodels. The EA, OCBAm+ and OCBA-ss use the mean value of each design for comparison, and the mean performance value is computed directly from the simulation output. Unlike our proposed methods and OSD, they do not rely on any response surface to estimate the performance value for any design. In order to compare the performance of these allocation approaches, we test them empirically on the following experiments.

- Experiment 1: Consider the optimization problem

$$\min f(x) = (x - 5)^2.$$

The experiment assumes that the noise for simulation has normal distribution $N(0, 2^2)$ for the objective value. We discretize the domain of the function into 100 evenly spaced points from 0 to 10. Since the underlying function is quadratic, we can use the OCBA-mr method directly for this problem. In order to utilize the OCBA-mrp method, we divide the 100 points into 5 adjacent partitions and there are 20 points in each partition. We want to select the top $m = 5$ designs.

- Experiment 2: Consider the Griewank function

$$\min f(x) = 10 \times (1 + (1/4000) \times x^2 - \cos(x)).$$

We discretize the domain of the function into 100 evenly spaced points from 0 to 20. The nature of this underlying function does not allow us to apply the OCBA-mr method directly. To test the performance of the OCBA-mr method for the non-quadratic problem, we make a slight modification by dividing the solution space into 5 adjacent partitions. Then, we allocate the simulation budget equally to each partition and use the OCBA-mr method within each partition. In this experiment, the OCBA-mrp is also tested under this partition pattern. The experiment assumes that the noise for simulation has normal distribution $N(0, 0.2^2)$. We want to select the top $m = 3$ designs.

- Experiment 3: Consider the optimization problem

$$\min f(x) = \sin(x) + \sin(10x/3) + \log(x) - 0.84x + 3.$$

The entire domain consists of 200 discrete points for $x \in [0, 8]$. The 200 points are further divided into 10 adjacent partitions so that there are 20 points in each partition. This experiment assumes that the simulation noise is a standard normal random variable. Similar to the setting in experiment 2, the OCBA-rm method is modified to handle this non-quadratic problem. We want to select the top $m = 5$ designs.

- Experiment 4: Consider the optimization problem

$$\min f(x) = 2(x - 0.75)^2 + \sin(8\pi x - \pi/2).$$

The experiment assumes that the noise for simulation has standard normal distribution for the objective measure. We discretize the domain of the function into 200 evenly spaced points from 0 to 2. The 200 points are further divided into 20 adjacent partitions so that there are 10 points in each partition. Similar to the setting in experiment 2, the OCBA-rm method is modified to handle this non-quadratic problem. We want to select the top $m = 3$ designs.

- Experiment 5: Consider the optimization problem

$$\min f(x_1, x_2) = \frac{1}{40}(x_1^2 + x_2^2) - \cos(x_1) \cos\left(\frac{x_2}{\sqrt{2}}\right) + 1.$$

The experiment assumes that the noise for simulation has normal distribution $N(0, 2^2)$ for the objective measure. x_1 and x_2 are continuous variables with $-5 \leq x_1 \leq 5$ and $-5 \leq x_2 \leq 5$. We discretize the solution space into 11×11 discrete points $x_1 = \{-5, -4, \dots, 5\}$ and $x_2 = \{-5, -4, \dots, 5\}$. We divide the 11×11 designs into 11 adjacent partitions with 11 designs in each partition. For each $i = 1, 2, \dots, 11$, partition i consists of design points $(-5, i - 6)$, $(-4, i - 6)$, ..., $(5, i - 6)$. We want to select the top $m = 3$ designs, which are $(0, -1)$, $(0, 0)$ and $(0, 1)$.

- Experiment 6: Consider the (s, S) inventory problem in the Simulation Optimization Library (http://simopt.org/wiki/index.php?title=SS_Inventory). The decision variable (s, S) is the inventory strategy. When the inventory of design (s, S) on hand below s , an order is incurred to supplement the inventory to S . The optimization problem is to minimize the the $E[\text{Total cost per period}]$. The domain is discretized into 20×20 discrete points with $s = \{810, 820, \dots, 1000\}$ and $S = \{1510, 1520, \dots, 1700\}$. The discrete points are divided into 20 partitions with 20 points in each partition. For each $i = 1, 2, \dots, 20$, partition i consists of points $(800 + 10i, 1510)$, $(800 + 10i, 1520)$, ..., $(800 + 10i, 1700)$. In each partition, we assume the performances of the $E[\text{Total cost per period}]$ of each point can be modeled as quadratic functions, where the independent variable is S and s is kept as a constant. We want to select the top $m = 3$ designs, which are $(810, 1510)$, $(810, 1520)$ and $(820, 1510)$.

For the budget allocation procedures mentioned above, we let the initial number of replication n_0 be 10 and the incremental budget Δ be 100 for all the five experiments. Figures 1 reports the comparison results in terms of PCS . The estimate of PCS is based on the average of 2000 independent replications of each procedure for experiments 1-6.

It is observed that all the procedures obtain higher PCS as the available simulation budget increases. Our proposed procedures OCBA-mr and OCBA-mrp perform the best on the tested examples. It shows that our proposed procedures have dramatically improved the selection quality of the top- m designs compared to the existing methods. When the considered problem is quadratic or approximately quadratic, such as example 1, the OCBA-mr method has the best performance. That is because the OCBA-mr method is developed for the quadratic or approximately quadratic problem setting. It can make full use of its advantages by simulating fewer designs once the quadratic assumption is satisfied. When the considered problem is partially quadratic or non-quadratic, such as examples 2-6, the OCBA-mrp method performs better than the modified OCBA-mr method. Because the OCBA-mrp can intelligently allocate the simulation budget both between and within each partition.

Compared to the existing approaches, our proposed approaches incorporate the information from the simulated designs into regression metamodel(s). We only need to simulate a subset of the alternative designs to build up the regression metamodel(s). The performances of the designs that have not been simulated can be inferred based on the metamodel(s). It dramatically improves the selection efficiency.

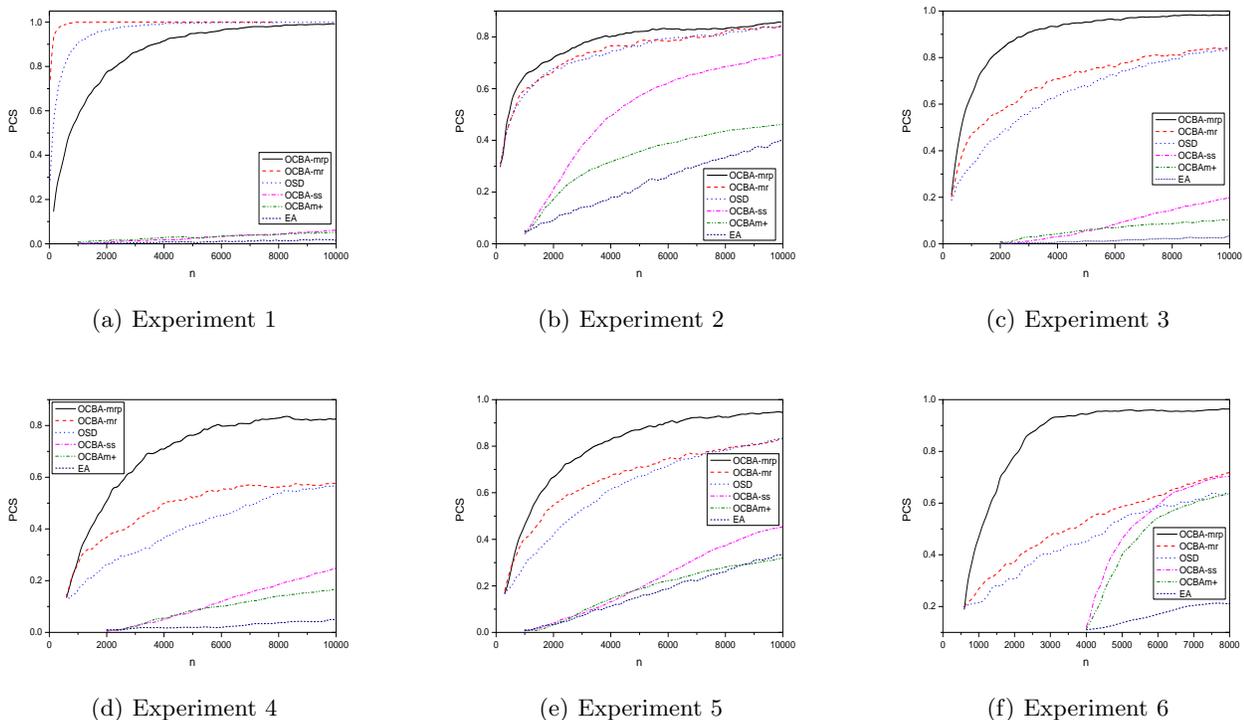


Fig. 1. Comparison results for different test examples.

6 Conclusions

In this study, we further enhanced the simulation efficiency of finding the top- m designs by incorporating quadratic regression equation(s). Using the large deviations theory, we formulate the problem of selecting the top- m designs as that of maximizing the minimum convergence rate of the false selection probability. We derived two asymptotically optimal budget allocation rules, OCBA-mr and OCBA-mrp, for one partition and multi-partition problem setting, respectively. Numerical results suggest that the proposed approaches can dramatically improve the selection efficiency compared to the existing R&S approaches. When the underlying function can be approximated by a quadratic function across the solution space, the OCBA-mr is more efficient. For the non-quadratic problem setting, the OCBA-mrp demonstrates the best performance.

References

- Brantley, M. W., Lee, L. H., Chen, C.-H., and Chen, A. (2013). Efficient simulation budget allocation with regression. *IIE Transactions*, 45(3):291–308.
- Brantley, M. W., Lee, L. H., Chen, C.-H., and Xu, J. (2014). An efficient simulation budget allocation method incorporating regression for partitioned domains. *Automatica*, 50(5):1391–1400.
- Burden, R. L. and Faires, J. D. (2001). Numerical analysis. 2001. *Brooks/Cole, USA*.
- Chen, C.-H., He, D., Fu, M., and Lee, L. H. (2008). Efficient simulation budget allocation for selecting an optimal subset. *INFORMS Journal on Computing*, 20(4):579–595.
- Chen, C. H., Lin, J., Yücesan, E., and Chick, S. E. (2000). Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270.
- De la Garza, A. et al. (1954). Spacing of information in polynomial regression. *The Annals of Mathematical Statistics*, 25(1):123–130.
- Gao, F., Gao, S., Xiao, H., and Shi, Z. (2018). Advancing constrained ranking and selection with regression in partitioned domains. *IEEE Transactions on Automation Science and Engineering*.
- Gao, S. and Chen, W. (2015). Efficient subset selection for the expected opportunity cost. *Automatica*, 59:19–26.
- Gao, S. and Chen, W. (2016). A new budget allocation framework for selecting top simulated designs. *IIE Transactions*, 48(9):855–863.

- Gao, S., Chen, W., and Shi, L. (2017a). A new budget allocation framework for the expected opportunity cost. *Operations Research*, 65(3):787–803.
- Gao, S. and Shi, L. (2015). Selecting the best simulated design with the expected opportunity cost bound. *IEEE Transactions on Automatic Control*, 60(10):2785–2790.
- Gao, S., Xiao, H., Zhou, E., and Chen, W. (2017b). Robust ranking and selection with optimal computing budget allocation. *Automatica*, 81:30–36.
- Glynn, P. and Juneja, S. (2004). A large deviations perspective on ordinal optimization. In Ingalls, R. G., Rossetti, M. D., Smith, J. S., and Peters, B. A., editors, *Proceedings of the 2004 Winter Simulation Conference*, pages 577–585. IEEE.
- Hayashi, F. (2000). *Econometrics*. princeton. *New Jersey, USA: Princeton University*.
- He, D., Chick, S. E., and Chen, C.-H. (2007). The opportunity cost and OCBA selection procedures in ordinal optimization. *IEEE Transactions on Systems, Man, and Cybernetics–Part C*, 37(5):951–961.
- Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 272–319.
- Kim, S. H. and Nelson, B. L. (2001). A fully sequential procedure for indifference-zone selection in simulation. *ACM Transactions on Modeling and Computer Simulation*, 11(3):251–273.
- Lee, L. H., Chen, C., Chew, E. P., Li, J., Pujowidianto, N. A., and Zhang, S. (2010). A review of optimal computing budget allocation algorithms for simulation optimization problem. *International Journal of Operations Research*, 7(2):19–31.
- McConnell, J. J. and Servaes, H. (1990). Additional evidence on equity ownership and corporate value. *Journal of Financial economics*, 27(2):595–612.
- Peng, Y., Chen, C.-H., Fu, M. C., and Hu, J.-Q. (2013). Efficient simulation resource sharing and allocation for selecting the best. *IEEE Transactions on Automatic Control*, 58(4):1017–1023.
- Xiao, H., Lee, L. H., and Chen, C.-H. (2015). Optimal budget allocation rule for simulation optimization using quadratic regression in partitioned domains. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(7):1047–1062.
- Xu, J., Huang, E., Chen, C.-H., and Lee, L. H. (2015). Simulation optimization: A review and exploration in the new era of cloud computing and big data. *Asia-Pacific Journal of Operational Research*, 32(03):1–34.
- Xu, J., Huang, E., Hsieh, L., Lee, L. H., Jia, Q. S., and Chen, C.-H. (2016). Simulation optimization in the era of industrial 4.0 and the industrial internet. *Journal of Simulation*, 10(4):310–320.
- Zhang, S., Lee, L. H., Chew, E. P., Xu, J., and Chen, C.-H. (2016). A simulation budget allocation procedure for enhancing the efficiency of optimal subset selection. *IEEE Transactions on Automatic Control*, 61(1):62–75.

Proof of Lemma 2

Let $\Lambda_i^{(d)}(\theta) = \ln E(e^{\theta \hat{d}_{i,m}})$ denote the log-moment generating function of $\hat{d}_{i,m}$ and $I_i^{(d)}(\cdot)$ denote the Fenchel-Legendre transform of $\Lambda_i^{(d)}$, that is

$$I_i^{(d)}(x) = \sup_{\theta \in \mathbb{R}} (\theta x - \Lambda_i^{(d)}(\theta)).$$

For any set A , let A° denotes its interior and for any function $f(\cdot)$, let $f'(x)$ denote the derivative of f at x . The effective domain of Λ_i is $D_{\Lambda_i} = \{\theta \in \mathbb{R} : \Lambda_i(\theta) < \infty\}$ and $F_i = \{\Lambda_i'(\theta) : \theta \in D_{\Lambda_i}^\circ\}$. Let $f_{min} = \min_{i \in \{1,2,\dots,t\}} f(x_i)$ and $f_{max} = \max_{i \in \{1,2,\dots,t\}} f(x_i)$. We assume that the interval $[f_{min}, f_{max}] \subset \bigcap_{i=1}^t F_i^\circ$. This assumption can be satisfied by some common families of distributions such Normal, Bernoulli, Poisson and Gamma family (Glynn and Juneja, 2004). It ensures that the range of $\hat{f}(x_i)$ include $[f_{min}, f_{max}]$ and $P(\hat{f}(x_i) > \hat{f}(x_m))$ and $P(\hat{f}(x_j) < \hat{f}(x_m))$ are positive for $i \in S_o$ and $j \in S_n$.

Then, $\hat{d}_{i,m} = \hat{f}(x_i) - \hat{f}(x_m)$ satisfies the large deviation principle with good rate functions

$$\left. \begin{aligned} & - \lim_{n \rightarrow \infty} \frac{1}{n} \log P\{\hat{d}_{i,m} > 0\}, i \in S_o, i \neq m \\ & - \lim_{n \rightarrow \infty} \frac{1}{n} \log P\{\hat{d}_{i,m} < 0\}, i \in S_n \end{aligned} \right\} = R_{m,i}(\alpha) = I_i^{(d)}(0),$$

Since $\hat{d}_{i,m} = \hat{f}(x_i) - \hat{f}(x_m)$ are all normally distributed random variables, the rate function can be expressed as follow according to the results in Glynn and Juneja (2004).

$$R_{m,i}(\alpha) = \frac{(f(x_m) - f(x_i))^2/2}{\sigma^2 \left(\frac{\rho_{mi,1}^2}{\alpha_1} + \frac{\rho_{mi,s}^2}{\alpha_s} + \frac{\rho_{mi,t}^2}{\alpha_t} \right)}.$$

Proof of Lemma 3

According to the definition of *PCS*, we have

$$\begin{aligned} PCS &= P \left\{ \bigcap_{i \in S_o} \bigcap_{j \in S_n} \hat{f}(x_i) \leq \hat{f}(x_j) \right\} \\ &\geq P \left\{ \left(\bigcap_{\substack{i \in S_o \\ i \neq m}} \hat{f}(x_i) \leq \hat{f}(x_m) \right) \cap \left(\bigcap_{j \in S_n} \hat{f}(x_j) \geq \hat{f}(x_m) \right) \right\} \\ &\geq 1 - \sum_{\substack{i \in S_o \\ i \neq m}} P\{\hat{f}(x_i) \geq \hat{f}(x_m)\} - \sum_{j \in S_n} P\{\hat{f}(x_j) \leq \hat{f}(x_m)\}. \end{aligned}$$

The last inequality follows from the Bonferroni inequality. Since $PFS = 1 - PCS$, we can get that

$$PFS \leq \sum_{\substack{i \in S_o \\ i \neq m}} P\{\hat{f}(x_i) \geq \hat{f}(x_m)\} + \sum_{j \in S_n} P\{\hat{f}(x_j) \leq \hat{f}(x_m)\}.$$

Therefore, a lower bound on *PFS* can be derived as

$$\max \left\{ \max_{\substack{i \in S_o \\ i \neq m}} P\{\hat{f}(x_i) \geq \hat{f}(x_m)\}, \max_{j \in S_n} P\{\hat{f}(x_j) \leq \hat{f}(x_m)\} \right\}$$

and an upper bound on PFS can be derived as

$$|S_o| \times |S_n| \times \max \left\{ \max_{\substack{i \in S_o \\ i \neq m}} P\{\hat{f}(x_i) \geq \hat{f}(x_m)\}, \max_{j \in S_n} P\{\hat{f}(x_j) \leq \hat{f}(x_m)\} \right\}.$$

According to Lemma 2, as n goes to infinity we can get the convergence rate function shown in Lemma 3.

Proof of Theorem 1

Given the definition of the key design i^* , we rewrite (3) as

$$\begin{aligned} & \max R_{m,i^*}(\boldsymbol{\alpha}) \\ & \text{s.t. } \alpha_1 + \alpha_s + \alpha_t = 1, \alpha_1, \alpha_s, \alpha_t > 0. \end{aligned} \quad (.1)$$

Since $R_{m,i^*}(\boldsymbol{\alpha})$ is concave and strictly increasing functions of $\boldsymbol{\alpha}$ (Xiao et al., 2015; Glynn and Juneja, 2004), the optimization problem is a convex programming problem. We define a Lagrangian function $L = -R_{m,i^*}(\boldsymbol{\alpha}) + \lambda(\alpha_1 + \alpha_s + \alpha_t - 1)$ and determine the optimal allocation based on its Karush-Kuhn-Tucker (KKT) conditions.

$$\begin{aligned} \frac{\partial L}{\partial \alpha_r} &= -\frac{(f(x_m) - f(x_{i^*}))^2/2}{\sigma^2 \left(\frac{\rho_{mi^*,1}^2}{\alpha_1} + \frac{\rho_{mi^*,s}^2}{\alpha_s} + \frac{\rho_{mi^*,t}^2}{\alpha_t} \right)^2} \frac{\rho_{mi^*,r}^2}{\alpha_r^*{}^2} + \lambda = 0, \\ & r = 1, s, t. \end{aligned} \quad (.2)$$

Plug (.2) into $\alpha_1^* + \alpha_s^* + \alpha_t^* = 1$, we can establish

$$\alpha_r^* = \frac{|\rho_{mi^*,r}|}{|\rho_{mi^*,1}| + |\rho_{mi^*,s}| + |\rho_{mi^*,t}|}, r = 1, s, t. \quad (.3)$$

Then, we can get the conclusions in Theorem 1.

Proof of Theorem 2

Plugging (4) into $R_{m,i^*}(\boldsymbol{\alpha})$, we have

$$R_{m,i^*}(\boldsymbol{\alpha}) = \frac{(f(x_m) - f(x_{i^*}))^2/2}{\sigma^2 (|\rho_{mi^*,1}| + |\rho_{mi^*,s}| + |\rho_{mi^*,t}|)^2}.$$

$$\begin{aligned} \frac{\partial R_{m,i^*}(\boldsymbol{\alpha})}{\partial x_s} &= \frac{-(f(x_m) - f(x_{i^*}))^2}{\sigma^2 (|\rho_{mi^*,1}| + |\rho_{mi^*,s}| + |\rho_{mi^*,t}|)^3} \\ & \times \left(\frac{\partial |\rho_{mi^*,1}|}{\partial x_s} + \frac{\partial |\rho_{mi^*,s}|}{\partial x_s} + \frac{\partial |\rho_{mi^*,t}|}{\partial x_s} \right). \end{aligned}$$

where

$$\begin{aligned} \frac{\partial \rho_{mi^*,1}}{\partial x_s} &= \frac{(x_{i^*} - x_m)(x_{i^*} + x_m - x_1 - x_t)}{(x_1 - x_s)^2(x_1 - x_t)}, \\ \frac{\partial \rho_{mi^*,s}}{\partial x_s} &= -\frac{(x_{i^*} - x_m)(x_{i^*} + x_m - x_1 - x_t)(2x_s - x_1 - x_t)}{(x_s - x_1)^2(x_s - x_t)^2}, \\ \frac{\partial \rho_{mi^*,t}}{\partial x_s} &= \frac{(x_{i^*} - x_m)(x_{i^*} + x_m - x_1 - x_t)}{(x_t - x_1)(x_t - x_s)^2}. \end{aligned}$$

In order to analyze the optimal location for x_s , we consider the following five cases.

Case I: $\frac{x_{i^*}+x_m}{2} < \frac{3x_1+x_t}{4}$

We have $\partial R_{m,i^*}(\boldsymbol{\alpha})/\partial x_s > 0$ when $x_s < (x_1 + x_t)/2$; $\partial R_{m,i^*}(\boldsymbol{\alpha})/\partial x_s < 0$ when $x_s > (x_1 + x_t)/2$. Therefore, when $(x_{i^*} + x_m)/2 < (3x_1 + x_t)/4$, we choose $x_s = (x_1 + x_t)/2$.

Case II: $\frac{3x_1+x_t}{4} \leq \frac{x_{i^*}+x_m}{2} < \frac{x_1+x_t}{2}$

We have $\partial R_{m,i^*}(\boldsymbol{\alpha})/\partial x_s > 0$ when $x_s < x_{i^*} + x_m - x_1$; $\partial R_{m,i^*}(\boldsymbol{\alpha})/\partial x_s < 0$ when $x_s > x_{i^*} + x_m - x_1$. Therefore, when $(3x_1 + x_t)/4 \leq (x_{i^*} + x_m)/2 < (x_1 + x_t)/2$, we choose $x_s = x_{i^*} + x_m - x_1$.

Case III: $\frac{x_{i^*}+x_m}{2} > \frac{x_1+3x_t}{4}$

We have $\partial R_{m,i^*}(\boldsymbol{\alpha})/\partial x_s > 0$ when $x_s < (x_1 + x_t)/2$; $\partial R_{m,i^*}(\boldsymbol{\alpha})/\partial x_s < 0$ when $x_s > (x_1 + x_t)/2$. Therefore, when $(x_{i^*} + x_m)/2 > (x_1 + 3x_t)/4$, we choose $x_s = (x_1 + x_t)/2$.

Case IV: $\frac{x_1+x_t}{2} < \frac{x_{i^*}+x_m}{2} \leq \frac{x_1+3x_t}{4}$

We have $\partial R_{m,i^*}(\boldsymbol{\alpha})/\partial x_s > 0$ when $x_s < x_{i^*} + x_m - x_t$; $\partial R_{m,i^*}(\boldsymbol{\alpha})/\partial x_s < 0$ when $x_s > x_{i^*} + x_m - x_t$. Therefore, when $(3x_1 + x_t)/4 \leq (x_{i^*} + x_m)/2 < (x_1 + x_t)/2$, we choose $x_s = x_{i^*} + x_m - x_t$.

Case V: $\frac{x_{i^*}+x_m}{2} = \frac{x_1+x_t}{2}$

The location of x_s has no influence on the results. Therefore, we choose $x_s = (x_1 + x_t)/2$ for consistency with Case II and Case III.

Analyzing the above the results, we can get the conclusions in Theorem 2.

Proof of Lemma 4

Let $\Lambda_{i_h}(\theta) = \ln E(e^{\theta \hat{f}(x_{i_h})})$ denote the log-moment generating function of $\hat{f}(x_{i_h})$ and $I_{i_h}(\cdot)$ denote the Fenchel-Legendre transform of Λ_{i_h} , that is

$$I_{i_h}(x) = \sup_{\theta \in \mathbb{R}} (\theta x - \Lambda_{i_h}(\theta)).$$

Let the scaled cumulant generating function of $(\hat{f}(x_{m_b}), \hat{f}(x_{i_h}))$ be denoted as follows:

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{n} \Lambda(\hat{f}(x_{m_b}), \hat{f}(x_{i_h}))(n\theta_b, n\theta_h) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \ln E(e^{n\theta_b \hat{f}(x_{m_b}) + n\theta_h \hat{f}(x_{i_h})}), \end{aligned}$$

By the Gärtner-Ellis theorem, $(\hat{f}(x_{m_b}), \hat{f}(x_{i_h}))$ satisfy the large deviation principle with good rate functions

$$\begin{aligned} & \left. \begin{aligned} & - \lim_{n \rightarrow \infty} \frac{1}{n} \log P\{\hat{f}(x_{i_h}) > \hat{f}(x_{m_b})\}, i_h \in S_o \\ & - \lim_{n \rightarrow \infty} \frac{1}{n} \log P\{\hat{f}(x_{i_h}) < \hat{f}(x_{m_b})\}, i_h \in S_n \end{aligned} \right\} \\ &= R_{m_b, i_h}(\theta_b, \theta_h, \boldsymbol{\alpha}_b, \boldsymbol{\alpha}_h) = \inf_v (\theta_b I_{m_b}(v) + \theta_h I_{i_h}(v)), h \neq b. \end{aligned}$$

Since $(\hat{f}(x_{m_b}), \hat{f}(x_{i_h}))$ are all normally distributed random variables, the rate function can be expressed as follows according to the results in Glynn and Juneja (2004). Then, we can get

$$R_{m_b, i_h}(\theta_b, \theta_h, \boldsymbol{\alpha}_b, \boldsymbol{\alpha}_h) = \frac{(f(x_{m_b}) - f(x_{i_h}))^2 / 2}{\frac{\sigma_b^2}{\theta_b} \left(\frac{\eta_{m_b, 1}^2}{\alpha_{1_b}} + \frac{\eta_{m_b, s}^2}{\alpha_{s_b}} + \frac{\eta_{m_b, k}^2}{\alpha_{k_b}} \right) + \frac{\sigma_h^2}{\theta_h} \left(\frac{\eta_{i_h, 1}^2}{\alpha_{1_h}} + \frac{\eta_{i_h, s}^2}{\alpha_{s_h}} + \frac{\eta_{i_h, k}^2}{\alpha_{k_h}} \right)}, h \neq b,$$

The proof of (8) is similar to that is given for Lemma 2, and hence is omitted for brevity.

Proof of Lemma 5

We first rewrite optimization model (10) as

$$\begin{aligned} & \max z \\ & \text{s.t. } \tilde{R}_{i_h}(\theta_h, \boldsymbol{\alpha}_h) - z \geq 0, h = 1, \dots, m, h \neq b, i_h = 1_h, \dots, k_h \\ & \quad R_{i_b}(\theta_b, \boldsymbol{\alpha}_b) - z \geq 0, i_b = 1_b, \dots, k_b \\ & \quad \alpha_{1_h} + \alpha_{s_h} + \alpha_{k_h} = 1, h = 1, \dots, m \\ & \quad \sum_{h=1}^l \theta_h = 1, \theta_h, \alpha_{1_h}, \alpha_{s_h}, \alpha_{k_h} \geq 0, h = 1, \dots, m. \end{aligned} \tag{.4}$$

Similarly, the optimization models (11) and (12) can be rewritten as follows:

$$\begin{aligned} & \max z \\ & \text{s.t. } \tilde{R}_{i_h}(\theta_h, \boldsymbol{\alpha}_h) - z \geq 0, i_h = 1_h, \dots, k_h \\ & \quad \alpha_{1_h} + \alpha_{s_h} + \alpha_{k_h} = 1, \alpha_{1_h}, \alpha_{s_h}, \alpha_{k_h} \geq 0, \end{aligned} \tag{.5}$$

for $h = 1, 2, \dots, m, h \neq b$, and

$$\begin{aligned} & \max z \\ & \text{s.t. } R_{i_b}(\theta_b, \boldsymbol{\alpha}_b) - z \geq 0, i_b = 1_b, \dots, k_b \\ & \quad \alpha_{1_b} + \alpha_{s_b} + \alpha_{k_b} = 1, \alpha_{1_b}, \alpha_{s_b}, \alpha_{k_b} \geq 0. \end{aligned} \tag{.6}$$

Let $\tilde{\boldsymbol{\alpha}}_h^*$ and $\tilde{\boldsymbol{\alpha}}_b^*$ be the optimal solutions to (.5) and (.6). The optimization model (.4), (.5) and (.6) have the same objective function, and the domain of (.4) is a subset of (.5) and (.6). Therefore, if $\tilde{\boldsymbol{\alpha}}_h^*$ and $\tilde{\boldsymbol{\alpha}}_b^*$ are feasible to (.4), they are the optimal solutions of (.4). According to the formula of $R_{i_b}(\theta_b, \boldsymbol{\alpha}_b)$ and $\tilde{R}_{i_h}(\theta_h, \boldsymbol{\alpha}_h)$, we can conclude that $\tilde{\boldsymbol{\alpha}}_h^*$ and $\tilde{\boldsymbol{\alpha}}_b^*$ do not depend on the value of θ_h . Therefore, we can conclude that $\tilde{\boldsymbol{\alpha}}_h^*$ and $\tilde{\boldsymbol{\alpha}}_b^*$ are feasible to (.4), and they are optimal to (10). Then we get the conclusions in Lemma 5.

Proof of Theorem 3

In order to determine the optimal $\boldsymbol{\alpha}_h^*$ for each partition, we consider two cases ($h \neq b$ and $h = b$).

For $h \neq b$, we solve the optimization model (11). Given the definition of the key design i_h^* , we rewrite (11) as

$$\begin{aligned} & \max \tilde{R}_{m_b, i_h^*}(\theta_h, \boldsymbol{\alpha}_h) \\ & \text{s.t. } \alpha_{1_h} + \alpha_{s_h} + \alpha_{k_h} = 1, \alpha_{1_h}, \alpha_{s_h}, \alpha_{k_h} > 0, \end{aligned} \tag{.7}$$

for $h = 1, 2, \dots, m$ and $h \neq b$.

Since $\tilde{R}_{i_h^*}(\theta_h, \boldsymbol{\alpha}_h)$ is concave and strictly increasing functions of $\boldsymbol{\alpha}_h$ (Xiao et al., 2015; Glynn and Juneja, 2004), the optimization problem is a convex programming problem. We define a Lagrangian function $L = -\tilde{R}_{m_b, i_h^*}(\theta_h, \boldsymbol{\alpha}_h) + \lambda_h(\alpha_{1_h} + \alpha_{s_h} + \alpha_{k_h} - 1)$ and determine the optimal allocation based on its Karush-Kuhn-Tucker (KKT) conditions.

$$\tilde{R}_{m_b, i_h^*}(\theta_h, \boldsymbol{\alpha}_h) = \frac{(f(x_{m_b}) - f(x_{i_h^*}))^2/2}{\frac{\sigma_h^2}{\theta_h} \left(\frac{\eta_{i_h^*,1}^2}{\alpha_{1_h}} + \frac{\eta_{i_h^*,s}^2}{\alpha_{s_h}} + \frac{\eta_{i_h^*,k}^2}{\alpha_{k_h}} \right)}.$$

$$\frac{\partial L}{\partial \alpha_{r_h}} = -\frac{(f(x_{m_b}) - f(x_{i_h^*}))^2/2}{\frac{\sigma_h^2}{\theta_h} \left(\frac{\eta_{i_h^*,1}^2}{\alpha_{1_h}} + \frac{\eta_{i_h^*,s}^2}{\alpha_{s_h}} + \frac{\eta_{i_h^*,k}^2}{\alpha_{k_h}} \right)^2} \frac{\eta_{i_h^*,r}^2}{\alpha_{r_h}^2} + \lambda_h = 0,$$

$$r_h = 1_h, s_h, k_h.$$

Using the fact that $\alpha_{1_h}^* + \alpha_{s_h}^* + \alpha_{k_h}^* = 1$, we can establish

$$\alpha_{r_h}^* = \frac{|\eta_{i_h^*,r}|}{|\eta_{i_h^*,1}| + |\eta_{i_h^*,s}| + |\eta_{i_h^*,k}|}, r_h = 1_h, s_h, k_h. \quad (.8)$$

For $h = b$, we solve the optimization model (12), which could be rewritten as

$$\begin{aligned} \max \quad & R_{m_b, i_b^*}(\theta_b, \boldsymbol{\alpha}_b) \\ \text{s.t.} \quad & \alpha_{1_b} + \alpha_{s_b} + \alpha_{k_b} = 1, \alpha_{1_b}, \alpha_{s_b}, \alpha_{k_b} > 0. \end{aligned} \quad (.9)$$

The process of solving the (.9) is similar to that is given for Theorem 1, and hence is omitted for brevity. Then, we can get the conclusions in Theorem 3.

Proof of Theorem 4

In order to determine the location of the support design s_h for each partition, we consider two cases ($h \neq b$ and $h = b$).

For $h \neq b$, we plug (13) into $\tilde{R}_{m_b, i_h^*}(\theta_h, \boldsymbol{\alpha}_h)$, we have

$$\tilde{R}_{m_b, i_h^*}(\theta_h, \boldsymbol{\alpha}_h) = \frac{(f(x_{m_b}) - f(x_{i_h^*}))^2/2}{\frac{\sigma_h^2}{\theta_h} \left(|\eta_{i_h^*,1}| + |\eta_{i_h^*,s}| + |\eta_{i_h^*,k}| \right)^2}.$$

$$\frac{\partial \tilde{R}_{m_b, i_h^*}(\theta_h, \boldsymbol{\alpha}_h)}{\partial x_{s_h}} = \frac{-(f(x_{m_b}) - f(x_{i_h^*}))^2}{\frac{\sigma_h^2}{\theta_h} \left(|\eta_{i_h^*,1}| + |\eta_{i_h^*,s}| + |\eta_{i_h^*,k}| \right)^3} \times \left(\frac{\partial |\eta_{i_h^*,1}|}{\partial x_{s_h}} + \frac{\partial |\eta_{i_h^*,s}|}{\partial x_{s_h}} + \frac{\partial |\eta_{i_h^*,k}|}{\partial x_{s_h}} \right).$$

where

$$\frac{\partial \eta_{i_h^*,1}}{\partial x_{s_h}} = \frac{(x_{k_h} - x_{i_h^*})(x_{1_h} - x_{i_h^*})}{(x_{1_h} - x_{s_h})^2(x_{1_h} - x_{k_h})},$$

$$\frac{\partial \eta_{i_h^*, s}}{\partial x_{s_h}} = -\frac{(x_{1_h} - x_{i_h^*})(x_{k_h} - x_{i_h^*})(2x_{s_h} - x_{1_h} - x_{k_h})}{(x_{s_h} - x_{1_h})^2(x_{s_h} - x_{k_h})^2},$$

$$\frac{\partial \eta_{i_h^*, k}}{\partial x_{s_h}} = \frac{(x_{1_h} - x_{i_h^*})(x_{k_h} - x_{i_h^*})}{(x_{k_h} - x_{1_h})(x_{k_h} - x_{s_h})^2}.$$

When $x_{s_h} > x_{i_h^*}$,

$$\frac{\partial \tilde{R}_{m_b, i_h^*}(\theta_h, \boldsymbol{\alpha}_h)}{\partial x_{s_h}} < 0.$$

When $x_{s_h} < x_{i_h^*}$,

$$\frac{\partial \tilde{R}_{m_b, i_h^*}(\theta_h, \boldsymbol{\alpha}_h)}{\partial x_{s_h}} > 0.$$

Therefore, $\tilde{R}_{m_b, i_h^*}(\theta_h, \boldsymbol{\alpha}_h)$ is maximized when

$$x_{s_h} = x_{i_h^*}.$$

According to equation (13), we can get

$$\alpha_{i_h^*}^* = \begin{cases} 1, & i_h = i_h^*, \\ 0, & \text{otherwise.} \end{cases}$$

When i_h^* is at the extreme locations (1_h and k_h),

$$\tilde{R}_{m_b, i_h^*}(\theta_h, \boldsymbol{\alpha}_h) = \frac{(f(x_{m_b}) - f(x_{i_h^*}))^2 / 2}{\sigma_h^2 \eta_{i_h^*, r}^2 / \theta_h} \alpha_{r_h}, r_h = i_h^*.$$

In this setting, the location of the support design s_h has no effect on $\tilde{R}_{m_b, i_h^*}(\theta_h, \boldsymbol{\alpha}_h)$. For simplicity, we use the D-optimal support design and let $x_{s_h} = (x_{1_h} + x_{k_h})/2$ (round to the nearest design as needed) (Kiefer, 1959).

For $h = b$, the proof is similar to that is given for Theorem 2, and hence is omitted for brevity. Then, we can get the conclusions in Theorem 4.

Proof of Theorem 5

Given the definition of key design i_h^* , we rewrite the optimization model (18) as

$$\begin{aligned} & \max \min \left\{ R_{m_b, i_b^*}(\theta_b, \boldsymbol{\alpha}_b^*), \min_{\substack{1 \leq h \leq m \\ h \neq b}} \left(R_{m_b, i_h^*}(\theta_b, \theta_h, \boldsymbol{\alpha}_b^*, \boldsymbol{\alpha}_h^*) \right) \right\}, \\ & \text{s.t. } \sum_{h=1}^l \theta_h = 1, \theta_h \geq 0, h = 1, \dots, l. \end{aligned}$$

The $R_{m_b, i_b^*}(\theta_b, \boldsymbol{\alpha}_b^*)$ and $R_{m_b, i_h^*}(\theta_b, \theta_h, \boldsymbol{\alpha}_b^*, \boldsymbol{\alpha}_h^*)$ are concave and strictly increasing functions of θ_h and θ_b (Glynn and Juneja, 2004). Therefore, the optimization problem is a convex programming problem, and the first order condition is also the optimality condition.

Rewrite the optimization model above as

$$\begin{aligned} & \max z \\ & \text{s.t. } R_{m_b, i_h^*}(\theta_b, \theta_h, \boldsymbol{\alpha}_b^*, \boldsymbol{\alpha}_h^*) - z \geq 0, h = 1, \dots, l, h \neq b \\ & \quad R_{m_b, i_b^*}(\theta_b, \boldsymbol{\alpha}_b^*) - z \geq 0 \\ & \quad \sum_{h=1}^l \theta_h = 1, \theta_h \geq 0, h = 1, \dots, l. \end{aligned} \tag{.10}$$

From the KKT conditions, there exist $\lambda_h \geq 0$, $h = 1, 2, \dots, l$ and $\mu > 0$ such that

$$1 - \sum_{h=1}^m \lambda_h = 0 \quad (.11)$$

$$\mu - \lambda_h \frac{\partial R_{m_b, i_h^*}(\theta_b^*, \theta_h^*, \boldsymbol{\alpha}_b^*, \boldsymbol{\alpha}_h^*)}{\partial \theta_h} = 0, h = 1, 2, \dots, l, h \neq b \quad (.12)$$

$$\mu - \sum_{\substack{h=1 \\ h \neq b}}^l \lambda_h \frac{\partial R_{m_b, i_h^*}(\theta_b^*, \theta_h^*, \boldsymbol{\alpha}_b^*, \boldsymbol{\alpha}_h^*)}{\partial \theta_b} - \lambda_b \frac{\partial R_{m_b, i_b^*}(\theta_b^*, \boldsymbol{\alpha}_b^*)}{\partial \theta_b} = 0 \quad (.13)$$

$$\lambda_h (z - R_{m_b, i_h^*}(\theta_b^*, \theta_h^*, \boldsymbol{\alpha}_b^*, \boldsymbol{\alpha}_h^*)) = 0, h = 1, 2, \dots, l, h \neq b \quad (.14)$$

$$\lambda_b (z - R_{m_b, i_b^*}(\theta_b^*, \boldsymbol{\alpha}_b^*)) = 0 \quad (.15)$$

From (.11), it can be concluded that there must exist some $\lambda_h > 0$, $h = 1, 2, \dots, l$. If there is one $\lambda_h = 0$, $h = 1, 2, \dots, l, h \neq b$, it results $\mu = 0$ based on (.12). That means all the $\lambda_h = 0$, as $\partial R_{m_b, i_h^*}(\theta_b, \theta_h, \boldsymbol{\alpha}_b^*, \boldsymbol{\alpha}_h^*) / \partial \theta_h$ is strictly positive. Therefore, we can conclude that $\lambda_h \neq 0$, $h = 1, 2, \dots, l, h \neq b$. According to (.14),

$$z^* = R_{m_b, i_h^*}(\theta_b^*, \theta_h^*, \boldsymbol{\alpha}_b^*, \boldsymbol{\alpha}_h^*), h = 1, 2, \dots, l, h \neq b. \quad (.16)$$

In addition, since $\theta_h / \theta_b \rightarrow 0$ as l goes to infinity and $R_{m_b, i_b^*}(\theta_b^*, \boldsymbol{\alpha}_b^*) > R_{m_b, i_h^*}(\theta_b^*, \theta_h^*, \boldsymbol{\alpha}_b^*, \boldsymbol{\alpha}_h^*)$, $h = 1, 2, \dots, l, h \neq b$, it can be concluded that $\lambda_b = 0$ according to (.15). Therefore, we have

$$\sum_{\substack{h=1 \\ h \neq b}}^l \frac{\partial R_{m_b, i_h^*}(\theta_b^*, \theta_h^*, \boldsymbol{\alpha}_b^*, \boldsymbol{\alpha}_h^*) / \partial \theta_b}{\partial R_{m_b, i_h^*}(\theta_b^*, \theta_h^*, \boldsymbol{\alpha}_b^*, \boldsymbol{\alpha}_h^*) / \partial \theta_h} = 1. \quad (.17)$$

According to Theorem 4, we have

$$R_{m_b, i_h^*}(\theta_b^*, \theta_h^*, \boldsymbol{\alpha}_b^*, \boldsymbol{\alpha}_h^*) = \frac{(f(x_{m_b}) - f(x_{i_h^*}))^2 / 2}{\frac{\sigma_b^2}{\theta_b^*} \left(\frac{\eta_{m_b, 1}^2}{\alpha_{1_b}^*} + \frac{\eta_{m_b, s}^2}{\alpha_{s_b}^*} + \frac{\eta_{m_b, k}^2}{\alpha_{k_b}^*} \right) + \frac{\sigma_h^2}{\theta_h^*}}. \quad (.18)$$

Then, we can get

$$\begin{aligned} & \frac{\partial R_{m_b, i_h^*}(\theta_b^*, \theta_h^*, \boldsymbol{\alpha}_b^*, \boldsymbol{\alpha}_h^*)}{\partial \theta_b^*} \\ &= \frac{\sigma_b^2 (f(x_{m_b}) - f(x_{i_h^*}))^2 \left(\frac{\eta_{m_b, 1}^2}{\alpha_{1_b}^*} + \frac{\eta_{m_b, s}^2}{\alpha_{s_b}^*} + \frac{\eta_{m_b, k}^2}{\alpha_{k_b}^*} \right) / \theta_b^{*2}}{2 \left(\frac{\sigma_b^2}{\theta_b^*} \left(\frac{\eta_{m_b, 1}^2}{\alpha_{1_b}^*} + \frac{\eta_{m_b, s}^2}{\alpha_{s_b}^*} + \frac{\eta_{m_b, k}^2}{\alpha_{k_b}^*} \right) + \frac{\sigma_h^2}{\theta_h^*} \right)^2}, \end{aligned} \quad (.19)$$

and

$$\frac{\partial R_{m_b, i_h^*}(\theta_b^*, \theta_h^*, \boldsymbol{\alpha}_b^*, \boldsymbol{\alpha}_h^*)}{\partial \theta_h^*} = \frac{\sigma_h^2 (f(x_{m_b}) - f(x_{i_h^*}))^2 / \theta_h^{*2}}{2 \left(\frac{\sigma_b^2}{\theta_b^*} \left(\frac{\eta_{m_b, 1}^2}{\alpha_{1_b}^*} + \frac{\eta_{m_b, s}^2}{\alpha_{s_b}^*} + \frac{\eta_{m_b, k}^2}{\alpha_{k_b}^*} \right) + \frac{\sigma_h^2}{\theta_h^*} \right)^2}, \quad (.20)$$

As l goes to infinity, plug the results above into (.16) and (.17), we can get the equations in Theorem 5.