Adaptive-Region Sequential Design with Quantitative and Qualitative Factors in Application to HPC Configuration

(July 21, 2023)

Xia Cai[†], Li Xu[‡], C. Devon Lin[§], Yili Hong[‡] and Xinwei Deng[‡*]

[†]School of Science, Hebei University of Science and Technology, China

[§]Department of Mathematics and Statistics, Queen's University, Canada

[‡]Department of Statistics, Virginia Tech, USA

Abstract

Motivated by the need of finding optimal configuration in the high-performance computing (HPC) system, this work proposes an adaptive-region sequential design (ARSD) for optimization of computer experiments with qualitative and quantitative factors. Experiments with both qualitative and quantitative factors are also encountered in other applications. The proposed ARSD method considers a sequential design criterion under the additive Gaussian process to deal with both qualitative and quantitative factors. Moreover, the adaptiveness of the proposed sequential procedure allows the selection of next design point from the adaptive design region achieving a meaningful balance between exploitation and exploration for optimization. Theoretical justification of the adaptive design region is provided. The performance of the proposed method is evaluated by several numerical examples in simulations. The case study of HPC performance optimization further elaborates the merits of the proposed method.

Keywords: Adaptive design; Gaussian process; Design of experiment; Exploitation and exploration; Optimal Configuration.

*Address for correspondence: Xinwei Deng, Professor, Department of Statistics, Virginia Tech, Blacksburg, VA 24061 (E-mail: xdeng@vt.edu).

1 Introduction

In many areas of the fourth industrial revolution, high-performance computing (HPC) provides important infrastructures for enabling large-scale data analytics. Reliable computing performance is vital for cloud computing, data storage and management, and optimization (Sakellariou et al., 2018). Thus, the investigation of performance variability of HPC has drawn great attention in recent research (Cameron et al., 2019). The variability of HPC performance exists in several aspects, of which the input/output (IO) variability is of great interest. The IO performance is usually measured by the IO throughput (i.e., data transfer speed), which can vary from run to run. The variability of IO throughput can be affected by various system factors such as CPU frequency, the number of threads, IO operation mode, and IO scheduler, through a complicated relationship (Cameron et al., 2019).

To configure an HPC system with reliable IO performance, one important task is to find an optimal configuration (i.e., a certain level combination of system factors) that optimizes the IO performance measure. The search for the optimized configuration is a challenging task since the functional relationship between IO performance measure and system factors is unknown and complicated, especially for the HPC system containing both quantitative and qualitative inputs. To address this challenge, sequential designs in computer experiments (Sacks et al. 1989; Santner et al. 2003; Fang et al. 2005) can be used. It is a novel application of sequential designs of experiments for the HPC performance optimization.

The execution of computer experiments of HPC is time consuming. For example, it can take hours or days to collect the HPC IO performance in a single run under certain system configurations. Therefore, statistical surrogates are often adopted for statistical analysis and uncertainty quantification (Sacks et al. 1989; Bingham et al. 2014). One fundamental issue is the design of experiments, i.e., how to choose the settings of input variables to run computer experiments to obtain the output responses for the objectives of interest. The commonly used designs are space-filling designs (Joseph 2016; Wang et al. 2018). To entertain both qualitative and quantitative inputs, space-filling designs such as sliced Latin hypercube designs and marginally coupled designs have been introduced (Qian 2012; Deng et al. 2015; He et al. 2019). However, these designs are proposed

with the aim of building an accurate emulator and thus they are not designed for other objectives such as the optimization we consider here. An objective-oriented design approach is to use sequential designs which find the new input setting sequentially for the objective of interest (Picheny et al. 2016; Sauer et al. 2020). There are also works on adaptive design region by zooming the design region efficiently around the target regions (Picheny et al. 2010; Cortes et al., 2020). The sequential approach has appeared being efficient and advantageous as indicated in many applications (Gramacy 2020). For example, Bingham et al. (2014) adopted sequential designs for choosing input settings of a computer simulator for the maximization of the tidal power in the Bay of Fundy, Nova Scotia, Canada (Ranjan et al. 2011). One popular approach in the sequential design framework is to use an expected improvement (EI) criterion (Jones et al. 1998; Ponweiser et al. 2008). An EI criterion was initially introduced for the global optimization of black box functions (computer simulators) by Jones et al. (1998). Since then, various EI criteria have been proposed for other objectives such as contour estimation (Ranjan et al. 2008), quantile estimation (Roy 2008), estimating the probability of rare events and system failure (Bichon et al. 2009), and prediction (Yang et al. 2020). Other criteria in the sequential design framework include the upper confidence bound (Srinivas et al. 2012), the knowledge gradient method (Frazier et al. 2008; Scott et al. 2011), and hierarchical expected improvement (Chen et al. 2019). However, to the best of our knowledge, these sequential design approaches including those using EI criteria have exclusively focused on computer experiments with only quantitative inputs. These approaches may not be directly applicable to computer experiments, such as the HPC experiment, with both qualitative and quantitative factors.

In this article, our scope is to develop a sequential design approach for efficient optimization of computer experiments with both qualitative and quantitative (QQ) factors. In the HPC application, the IO operation mode is a qualitative variable, while the CPU frequency is a quantitative variable. We propose an *adapative-region sequential design* (ARSD) method for the global optimization for computer experiments with QQ factors. The proposed ARSD method considers the additive Gaussian process (AGP) (Deng et al. 2017) as the surrogate model for searching follow-up design points. Similar to the EI and other criteria, the

proposed sequential design criterion aims to achieve the balance between exploitation and exploration when searching for the next input setting. What is fundamentally different and makes this criterion novel is that the search design region at each stage via the new criterion is adaptive in the sense that the design region changes with the data collected. Theoretical justifications are provided to support the choice of the adaptive design region. In addition, the proposed ARSD criterion has a simple expression with meaningful interpretation to choose the next design point sequentially based on the AGP as the surrogate. The sequential design procedure with the proposed criterion appears to be efficient in finding the optimal setting, i.e., the setting of optimizing the response output.

The remainder of this paper is organized as follows. Section 2 briefly reviews the additive Gaussian process model. Section 3 presents the details of the proposed ARSD method and its theoretical justification on the choice of adaptive design region. In Section 4, several numerical examples are conducted to illustrate the effectiveness of the proposed method. Section 5 presents the case study of HPC experiments, where the proposed method is demonstrated to efficiently find the optimal setting for HPC performance optimization. We conclude this work with some discussion in Section 6.

2 Brief Review of Additive Gaussian Process Model

Consider a computer experiment with p quantitative factors $\boldsymbol{x} = (x_1, \cdots, x_p)^T \in \boldsymbol{X} \subseteq R^p$ and q qualitative factors $\boldsymbol{z} = (z_1, \cdots, z_q)^T \in \boldsymbol{Z}$ with the jth qualitative factor having m_j levels, $j = 1, \cdots, q$, and the corresponding output is denoted by Y, where $\boldsymbol{Z}$ contains $M = \prod_{j=1}^{q} m_j$ elements. Suppose that the observed data are $(\boldsymbol{w}_t^T, y_t), t = 1, \cdots, n$, where $\boldsymbol{w}_t = (\boldsymbol{x}_t^T, \boldsymbol{z}_t^T)^T = (x_{t1}, \cdots, x_{tp}, z_{t1}, \cdots, z_{tq})^T$. To model the relationship between output Y and input $\boldsymbol{w}$, the AGP model assumes

$$Y(\boldsymbol{x}, z_1, \cdots, z_q) = \mu + G_1(\boldsymbol{x}, z_1) + \cdots + G_q(\boldsymbol{x}, z_q), \tag{1}$$

where μ is the overall mean, and the G_j's are independent Gaussian processes with mean zero and covariance function ϕ_j. For two inputs $\boldsymbol{w}_1 = (\boldsymbol{x}_1^T, \boldsymbol{z}_1^T)^T = (x_{11}, \cdots, x_{1p}, z_{11}, \cdots, z_{1q})^T$

and $\boldsymbol{w}_2 = (\boldsymbol{x}_2^T, \boldsymbol{z}_2^T)^T = (x_{21}, \cdots, x_{2p}, z_{21}, \cdots, z_{2q})^T$, the covariance function ϕ_j is given by

$$\phi_j(G_j(\boldsymbol{x}_1, z_{1j}), G_j(\boldsymbol{x}_2, z_{2j})) = \sigma_j^2 \tau^{(j)}_{z_{1j}, z_{2j}} R(\boldsymbol{x}_1, \boldsymbol{x}_2 | \boldsymbol{\theta}^{(j)}), \tag{2}$$

where σ_j^2 is the variance component associated with G_j, $\tau^{(j)}_{r,s}$ is the correlation of the rth level and the sth level of the qualitative factor $z_j, j = 1, \cdots, q$. That is, $\tau^{(j)}_{r,s}$ is the (r,s)th element in correlation matrix $\boldsymbol{T}^{(j)} = (\tau^{(j)}_{r,s})_{m_j \times m_j}$, $j = 1, \ldots, q$. Note that matrix $\boldsymbol{T}^{(j)}$ needs to be a valid correlation matrix, i.e., $\boldsymbol{T}^{(j)} = (\tau^{(j)}_{r,s})_{m_j \times m_j}$ needs to be a positive definite matrix with unit diagonal elements. To satisfy this requirement, the hypersphere parameterization approach in Zhou et al. (2011) is adopted here to parameterize $\boldsymbol{T}^{(j)}$ for $j = 1, \ldots, q$. The details of the hypersphere parameterization are given in the appendix. A common choice of the correlation function is the Gaussian correlation function $R(\boldsymbol{x}_1, \boldsymbol{x}_2 | \boldsymbol{\theta}^{(j)}) = \exp\left\{-\sum_{i=1}^p \theta_i^{(j)} (x_{1i} - x_{2i})^2\right\}$ for any two quantitative inputs $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, where $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, \cdots, \theta_p^{(j)})^T$ (Deng et al. 2017). Then, the response Y follows a Gaussian process with mean zero and the covariance function ϕ specified by

$$\begin{aligned}
\phi(Y(\boldsymbol{w}_1), Y(\boldsymbol{w}_2)) &= cov(Y(\boldsymbol{x}_1, \boldsymbol{z}_1), Y(\boldsymbol{x}_2, \boldsymbol{z}_2)) \\
&= \sum_{j=1}^q \sigma_j^2 \tau^{(j)}_{z_{1j}, z_{2j}} R(\boldsymbol{x}_1, \boldsymbol{x}_2 | \boldsymbol{\theta}^{(j)}) \\
&= \sum_{j=1}^q \sigma_j^2 \tau^{(j)}_{z_{1j}, z_{2j}} \exp\left\{-\sum_{i=1}^p \theta_i^{(j)} (x_{1i} - x_{2i})^2\right\}.
\end{aligned} \tag{3}$$

We denote $Y_0 = Y(\boldsymbol{w}_0)$ as the prediction of Y at a new setting $\boldsymbol{w}_0 = (\boldsymbol{x}_0^T, \boldsymbol{z}_0^T)^T$. Let $\boldsymbol{y}_n = (y_1, \cdots, y_n)^T$ be n outputs from the input $(\boldsymbol{w}_1^T, \cdots, \boldsymbol{w}_n^T)^T$. Based on the AGP, it is easy to obtain that $Y_0 | \boldsymbol{y}_n$ follows a normal distribution with

$$E(Y_0 | \boldsymbol{y}_n) = \mu_{0|n} = \mu + \boldsymbol{r}_0^T \boldsymbol{\Phi}^{-1} (\boldsymbol{y}_n - \mu \boldsymbol{1}_n), \tag{4}$$

$$Var(Y_0 | \boldsymbol{y}_n) = \sigma^2_{0|n} = \sum_{j=1}^q \sigma_j^2 - \boldsymbol{r}_0^T \boldsymbol{\Phi}^{-1} \boldsymbol{r}_0, \tag{5}$$

where $\boldsymbol{\Phi}$ is the covariance matrix of $\boldsymbol{y}_n$, and $\boldsymbol{r}_0 = (\phi_{01}, \cdots, \phi_{0n})^T$ with ϕ_{0t} given by $\phi_{0t} = \phi(Y(\boldsymbol{w}_0), Y(\boldsymbol{w}_t)) = \sum_{j=1}^q \sigma_j^2 \tau^{(j)}_{z_{0j}, z_{tj}} \exp\left\{-\sum_{i=1}^p \theta_i^{(j)} (x_{0i} - x_{ti})^2\right\}$, $t = 1, 2, \ldots, n$.

Clearly the mean and variance of $Y_0 | \boldsymbol{y}_n$, i.e., $\mu_{0|n}$ and $\sigma^2_{0|n}$, involve the parameters μ, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \cdots, \sigma_q^2)$, $\boldsymbol{T} = (\boldsymbol{T}^{(1)}, \cdots, \boldsymbol{T}^{(q)})$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(q)})$. There are $1 + q +$

$\sum_{j=1}^{q} m_j(m_j-1)/2 + pq$ parameters. To estimate these parameters, Deng et al. (2017) considered the maximum likelihood estimation as

$$\{\hat{\mu}, \hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{T}}, \hat{\boldsymbol{\theta}}\} = \underset{\mu, \boldsymbol{\sigma}^2, \boldsymbol{T}, \boldsymbol{\theta}}{\operatorname{argmax}} \left[-\frac{1}{2}\log|\boldsymbol{\Phi}| - \frac{1}{2}(\boldsymbol{y}_n - \mu\mathbf{1}_n)^T \boldsymbol{\Phi}^{-1}(\boldsymbol{y}_n - \mu\mathbf{1}_n) \right]. \tag{6}$$

With the estimates obtained from (6), one can calculate $\hat{\mu}_{0|n}, \hat{\sigma}^2_{0|n}$ and subsequently compute the predictive distribution of $Y_0|\boldsymbol{y}_n$. The details of the maximum likelihood estimation can be found in the appendix.

3 The Proposed Adaptive-Region Sequential Design

In this section, we describe the proposed ARSD based on the additive Gaussian process model for computer experiments with quantitative and qualitative factors. The proposed ARSD focuses on the efficient global optimization, i.e., efficiently finding the optimum through the sequential design procedure. Without loss of generality, we consider the minimization problem. That is, given n collected data points $(\boldsymbol{w}_t^T, y_t), t = 1, \cdots, n$, the key interest is to find the next design point $\boldsymbol{w}_{n+1} \in \mathcal{A}$ for the computer experiment such that we can promptly find the optimal setting of $\boldsymbol{w}^*$ to reach the smallest value of output $y(\boldsymbol{w})$. Here $\mathcal{A}$ is the whole design region of $\boldsymbol{w}$, i.e., $\mathcal{A} = \{(\boldsymbol{x}, \boldsymbol{z}) | \boldsymbol{x} \in \boldsymbol{X}, \boldsymbol{z} \in \boldsymbol{Z}\}$.

Specifically, Section 3.1 presents the proposed ARSD method. Section 3.2 provides some theoretical justification on the adaptiveness regarding the design region for the ARSD method.

3.1 The Adaptive-Region Sequential Design Criterion

For a computer experiment with quantitative and qualitative factors, suppose that the collected data are $(\boldsymbol{w}_t^T, y_t), t = 1, \cdots, n$. We can use the AGP to fit the data and obtain the predictive normal distribution of $Y_0|\boldsymbol{y}_n$ for any input $\boldsymbol{w}_0$ with mean $\mu_{0|n}(\boldsymbol{w}_0)$ and variance $\sigma^2_{0|n}(\boldsymbol{w}_0)$ as described in (4) and (5). To find the design point for minimizing the response output, one would encourage local exploitation as well as the flexibility of exploration to other regions. Following Auer (2002), the exploitation is to makes decisions (i.e., design points)

to maximize its current estimated rewards (i.e., responses) based on limited knowledge, while the exploration is to improve the knowledge about the reward generating process, but not necessarily maximize the current rewards. In our content, the design point with a small value of $\mu_{0|n}(\boldsymbol{w}_0)$ will support local exploitation. But there is only limited knowledge about the minimization. One might decide to do exploration in a wider area. The design point with a large value of $\sigma_{0|n}(\boldsymbol{w}_0)$ will encourage the exploration. Thus the idea of considering both $\mu_{0|n}(\boldsymbol{w}_0)$ and $\sigma^2_{0|n}(\boldsymbol{w}_0)$ is natural thinking for choosing the next design point. Intuitively, we would like to sequentially choose the next point $\boldsymbol{w}_0$ when the mean $\mu_{0|n}(\boldsymbol{w}_0)$ is small and the standard deviation $\sigma_{0|n}(\boldsymbol{w}_0)$ is large, which is to encourage balance between exploitation and exploration. Under this consideration, it would be reasonable to consider a criterion of choosing the next point $\boldsymbol{w}_{n+1}$ as $\min_{\boldsymbol{w}_0}[\hat{\mu}_{0|n}(\boldsymbol{w}_0) - \rho\hat{\sigma}_{0|n}(\boldsymbol{w}_0)]$, where $\rho \geq 0$ is a tuning parameter. Note that if ρ is chosen to be $z_{\alpha/2}$, the $\alpha/2$ upper quantile of the standard normal distribution, then $\hat{\mu}_{0|n}(\boldsymbol{w}_0) - \rho\hat{\sigma}_{0|n}(\boldsymbol{w}_0) = \hat{\mu}_{0|n}(\boldsymbol{w}_0) - z_{\alpha/2}\hat{\sigma}_{0|n}(\boldsymbol{w}_0)$ is the lower confidence limit of $[Y_0|\boldsymbol{y}_n]$ with the confidence level $1-\alpha$. It implies that, instead of minimizing the mean of $[Y_0|\boldsymbol{y}_n]$, it is to minimize the lower confidence limit of $[Y_0|\boldsymbol{y}_n]$ when searching for the input of achieving the minimum of the response surface. Note that such a criterion can be easily modified for the maximization problem as $\max_{\boldsymbol{w}_0}[\mu_{0|n}(\boldsymbol{w}_0) + \rho\sigma_{0|n}(\boldsymbol{w}_0)]$, and is closely related to upper confidence bound in the literature (Wang et al., 2021).

However, the optimization requires the search over the whole design space $\mathcal{A} = \{(\boldsymbol{x}, \boldsymbol{z})|\boldsymbol{x} \in \boldsymbol{X}, \boldsymbol{z} \in \boldsymbol{Z}\}$ in each iteration of the sequential design procedure. When the inputs of experiments contain both quantitative factors and qualitative factors, the design space $\mathcal{A}$ is discontinuous in nature due to the qualitative factors. The optimization will be complicated to obtain the global optimum especially when there are a large number of qualitative factors with many levels. Moreover, as the sequential design procedure is conducted with more collected data points, there should be more information on the design region where the minimum is located. Therefore, we propose an adaptive-region sequential design (ARSD) criterion for finding the next design point where the design region is adaptive in each iteration of the sequential procedure. The proposed ARSD criterion borrows the intuition from LCB (Wang et al., 2021) and theorizes the intuition by defining the adaptive region. We provide

a theoretical justification that our strategy is to minimize an LCB criterion but restrict to a region where the optimal solution should lie in with a high probability. Specifically, the proposed ARSD criterion is to choose the next point $\boldsymbol{w}_{n+1}$ as

$$\boldsymbol{w}_{n+1} = \underset{\boldsymbol{w}_0 \in \mathcal{A}_n}{\operatorname{argmin}} \left\{ \hat{\mu}_{0|n}(\boldsymbol{w}_0) - \rho \hat{\sigma}_{0|n}(\boldsymbol{w}_0) \right\}, \tag{7}$$

where $\mathcal{A}_n \subset \mathcal{A}$ is the adaptive design region as

$$\mathcal{A}_n = \left\{ \boldsymbol{w}_0 \in \mathcal{A} : \hat{\mu}_{0|n}(\boldsymbol{w}_0) - \sqrt{\beta_{0|n}} \hat{\sigma}_{0|n}(\boldsymbol{w}_0) \leq \min_{\boldsymbol{w}_0} [\hat{\mu}_{0|n}(\boldsymbol{w}_0) + \sqrt{\beta_{0|n}} \hat{\sigma}_{0|n}(\boldsymbol{w}_0)] \right\}, \tag{8}$$

where $\beta_{0|n} = 2\log(\pi^2 n^2 M/6\alpha)$ with $M = |\boldsymbol{Z}| = \prod_{j=1}^{q} m_j$ being the size of $\boldsymbol{Z}$, and $\mu_{0|n}(\boldsymbol{w}_0)$ and $\sigma_{0|n}(\boldsymbol{w}_0)$ are given in (4) and (5). Note that $\beta_{0|n}$ is a very complicated function of n, M and α. When the sample size n or the size M is large, the value of $\beta_{0|n}$ can be large, which may over-emphasize the role of the predictive standard deviation. Thus a relatively simple number ρ is used in (7). Regarding to the stopping criterion, we follows Jones et al. (1998) to stop searching the next design point when the objective value in (7) is less than 1% of the current objective value.

For the adaptive region $\mathcal{A}_n$ in (8), it is easy to see that the lower bound $\hat{\mu}_{0|n}(\boldsymbol{w}_0) - \sqrt{\beta_{0|n}} \hat{\sigma}_{0|n}(\boldsymbol{w}_0)$ is always smaller than the upper bound $\hat{\mu}_{0|n}(\boldsymbol{w}_0) + \sqrt{\beta_{0|n}} \hat{\sigma}_{0|n}(\boldsymbol{w}_0)$ for every $\boldsymbol{w}_0 \in \mathcal{A}$. In (8), the design region $\mathcal{A}_n$ consists of the points $\hat{\mu}_{0|n}(\boldsymbol{w}_0) - \sqrt{\beta_{0|n}} \hat{\sigma}_{0|n}(\boldsymbol{w}_0)$ is less than the minimum of $\hat{\mu}_{0|n}(\boldsymbol{w}_0) + \sqrt{\beta_{0|n}} \hat{\sigma}_{0|n}(\boldsymbol{w}_0)$. It implies that the inequality in Eq. (8) would eliminate regions where the function value is suboptimal with a high probability. One can see that design region $\mathcal{A}_n$, as a subset of the whole region $\mathcal{A}$, varies with the data points collected sequentially. In the numerical example in Section 4, we illustrate how $\mathcal{A}_n$ changes as the data arrives. Solving the optimization with the ARSD criterion will be more efficient because the search for next input setting in each iteration is confined in $\mathcal{A}_n$ rather than the whole design region $\mathcal{A}$.

3.2 Theoretical Justification for the Adaptive Design Region

For notation convenience, we use $\mu_{0|n}$ and $\sigma_{0|n}$ rather than their estimates in the presentation of theoretical investigation and technical proofs. These theoretical results still hold when

the estimates are used. The technical proofs can be found in the appendix. Now we focus on finding the adaptive region $\mathcal{A}_n$ based on the properties of the predictive mean $\mu_{0|n}(\boldsymbol{w}_0)$. First, we present Lemma 1 below.

Lemma 1. *For a given quantitative factors $\boldsymbol{x}_0 \in \boldsymbol{X}$ from a design point $\boldsymbol{w}_0 = (\boldsymbol{x}_0^T, \boldsymbol{z}_0^T)^T \in \mathcal{A}$, let $y(\boldsymbol{w}_0)$ be a sample from the Gaussian process in (1). For all $\alpha \in (0,1)$, we have*

$$P\left(|y(\boldsymbol{w}_0) - \mu_{0|n}(\boldsymbol{w}_0)| \leq \sqrt{\beta_{0|n}}\sigma_{0|n}(\boldsymbol{w}_0), \forall \boldsymbol{z}_0 \in \boldsymbol{Z}, \forall n \geq 1\right) \geq 1 - \alpha, \tag{9}$$

where $\beta_{0|n} = 2\log(\pi^2 n^2 M/6\alpha)$ with $M = |\boldsymbol{Z}| = \prod_{j=1}^{q} m_j$ being the size of $\boldsymbol{Z}$, and $\mu_{0|n}(\boldsymbol{w}_0)$ and $\sigma_{0|n}(\boldsymbol{w}_0)$ are given in (4) and (5).

Lemma 1 is based on Lemma 1 in Jala et al. (2016), which established similar results for a finite space. Because $\boldsymbol{Z}$ is a finite discrete space and $\boldsymbol{x}_0 \in \boldsymbol{X}$ is fixed, the design space in Lemma 1 is finite. We can easily extend their proof to ensure that Lemma 1 holds, and thus we skip the proof of Lemma 1.

Lemma 1 gives the lower bound and upper bound for the prediction of $y(\boldsymbol{w}_0)$. Let denote the lower bound $\mu_{0|n}^L(\boldsymbol{w}_0)$ and the upper bound $\mu_{0|n}^U(\boldsymbol{w}_0)$ as follows:

$$\mu_{0|n}^L(\boldsymbol{w}_0) = \mu_{0|n}(\boldsymbol{w}_0) - \sqrt{\beta_{0|n}}\sigma_{0|n}(\boldsymbol{w}_0),\ \mu_{0|n}^U(\boldsymbol{w}_0) = \mu_{0|n}(\boldsymbol{w}_0) + \sqrt{\beta_{0|n}}\sigma_{0|n}(\boldsymbol{w}_0), \tag{10}$$

Then Lemma 1 implies that given $\boldsymbol{x}_0 \in \boldsymbol{X}$, $y(\boldsymbol{w}_0)$ belongs to the interval $[\mu_{0|n}^L(\boldsymbol{w}_0), \mu_{0|n}^U(\boldsymbol{w}_0)]$ with the probability greater than $1-\alpha$. Moreover, Lemma 2 below shows that $\min_{\boldsymbol{w}_0 \in \mathcal{A}} y(\boldsymbol{w}_0)$ belongs to the interval $[\min_{\boldsymbol{w}_0 \in \mathcal{A}} \mu_{0|n}^L(\boldsymbol{w}_0), \min_{\boldsymbol{w}_0 \in \mathcal{A}} \mu_{0|n}^U(\boldsymbol{w}_0)]$ with the probability greater than $1 - \alpha$. Let us define $y_{\min}, \tilde{\mu}_{\min,n}, \tilde{\mu}_{\min,n}^L, \tilde{\mu}_{\min,n}^U$ as follows:

$$\begin{aligned} y_{\min} = \min_{\boldsymbol{w}_0 \in \mathcal{A}} y(\boldsymbol{w}_0),\ \tilde{\mu}_{\min,n} = \min_{\boldsymbol{w}_0 \in \mathcal{A}} \mu_{0|n}(\boldsymbol{w}_0), \\ \tilde{\mu}_{\min,n}^L = \min_{\boldsymbol{w}_0 \in \mathcal{A}} \mu_{0|n}^L(\boldsymbol{w}_0),\ \tilde{\mu}_{\min,n}^U = \min_{\boldsymbol{w}_0 \in \mathcal{A}} \mu_{0|n}^U(\boldsymbol{w}_0). \end{aligned} \tag{11}$$

Lemma 2. *Let $y_{\min}, \tilde{\mu}_{\min,n}^L, \tilde{\mu}_{\min,n}^U$ be the quantifies as defined in (11). Then for all $\alpha \in (0,1)$,*

$$P\left(y_{\min} \in [\tilde{\mu}_{\min,n}^L, \tilde{\mu}_{\min,n}^U], \forall n \geq 1\right) \geq 1 - 2\alpha. \tag{12}$$

The proof of Lemma 2 can be found in the appendix. Now we can obtain the bound for the discrepancy between of the minimum of the response $y_{\min}$ and its estimate $\tilde{\mu}_{\min,n}$ in a probabilistic manner.

Theorem 1. *Let* $\mu^L_{0|n}(\boldsymbol{w}_0), y_{\min}, \tilde{\mu}_{\min,n}, \tilde{\mu}^U_{\min,n}$ *be the quantitie as defined in* (10) *and* (11). *Then for all* $\alpha \in (0,1)$, *we have*

$$P\left(|\tilde{\mu}_{\min,n} - y_{\min}| \leq \sqrt{\beta_{0|n}} \sup_{\boldsymbol{w}_0 \in \mathcal{A}_n} \sigma_{0|n}(\boldsymbol{w}_0), \forall n \geq 1\right) \geq 1 - 4\alpha, \tag{13}$$

where $\mathcal{A}_n = \{\boldsymbol{w}_0 \in \mathcal{A} : \mu^L_{0|n}(\boldsymbol{w}_0) \leq \tilde{\mu}^U_{\min,n}\}$.

Clearly, the definition of $\mathcal{A}_n$ here is the same as in (8). It is easy to see that the lower bound $\mu^L_{0|n}(\boldsymbol{w}_0)$ is smaller than the upper bound $\mu^U_{0|n}(\boldsymbol{w}_0)$ for every $\boldsymbol{w}_0 \in \mathcal{A}$. In Theorem 1, the design region $\mathcal{A}_n$ consists of the points whose the lower bound $\mu^L_{0|n}(\boldsymbol{w}_0)$ is less than the minimum of its upper bound $\tilde{\mu}^U_{\min,n}$. Thus $\mathcal{A}_n$ can have a smaller size than the whole design region $\mathcal{A}$. Note that $\mathcal{A}_n$ does not always cover $\mathcal{A}_{n-1}$ because of the stochastic nature of estimates. When n is large enough, this region converges in an asymptotic fashion. Theorem 1 provides a bound for the difference between $y_{\min}$ and its estimate $\tilde{\mu}_{\min,n}$, which depends on $\mathcal{A}_n$. It implies that $\tilde{\mu}_{\min,n}$ will be in a small neighborhood of $y_{\min}$ with a relatively high probability. Furthermore, Corollary 1 states that the adaptive design region $\mathcal{A}_n$ will cover the true optimal setting $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} y(\boldsymbol{w})$ with a high probability.

Corollary 1. *Denote* $\boldsymbol{w}^*$ *be one of the optimal points that minimize* $y(\boldsymbol{w})$, *i.e.,* $y(\boldsymbol{w}^*) = \min_{\boldsymbol{w} \in \mathcal{A}} y(\boldsymbol{w})$. *Then for all* $\alpha \in (0,1)$, *we have*

$$P\left(\mathcal{A}_n \ni \boldsymbol{w}^*, \forall n \geq 1\right) \geq 1 - 3\alpha,$$

where $\mathcal{A}_n$ *is the design region defined in Theorem 1.*

4 Numerical Examples

In this section, we investigate the performance of the proposed ARSD method in comparison with the four benchmark methods defined as follows: (1) LCB: the method sequentially minimizes $\hat{\mu}_{0|n}(\boldsymbol{w}_0) - \sqrt{\beta_{0|n}}\hat{\sigma}_{0|n}(\boldsymbol{w})$. (2) EI: the method sequentially maximizes the expected improvement as the acquisition function, where the corresponding EI criterion is $E[I(\boldsymbol{w})] = \int_{y \in R} I(\boldsymbol{w}) f(y|\boldsymbol{w}, \boldsymbol{y}_n) dy$ with $f(y|\boldsymbol{w}, \boldsymbol{y}_n)$ to the predictive density and $I(\boldsymbol{w}) =$

$\max\{y_{\min,n} - Y(\boldsymbol{w}), 0\}$. Here $y_{\min,n}$ to be the minimum value of the obtained responses among the n runs. (3) MU: the method sequentially minimizes the prediction mean as the acquisition function, i.e., $\boldsymbol{w}_{n+1} = \text{argmin}_{\boldsymbol{w}_0}\ \hat{\mu}_{0|n}(\boldsymbol{w}_0)$; (4) SI: the method sequentially maximizes the prediction variance as the acquisition function, i.e., $\boldsymbol{w}_{n+1} = \text{argmax}_{\boldsymbol{w}_0}\ \hat{\sigma}_{0|n}(\boldsymbol{w}_0)$.

Note that these four benchmark approaches are sequential designs, each of which chooses the next design point by the given method, gets its response and updates the model estimation, and then continues to choose the next design point until the stopping criterion is met. Here we consider the methods in comparison have the same number of runs. In each numerical example, we will report the minimal values found by the five methods in comparison.

4.1 An Illustrative Example with one Qualitative Factor and one Quantitative Factor

Example 1. *Consider the simple case that there is only one quantitative factor $x \in [0, 1]$ and one qualitative factor z of three levels. The underlying function for the output response y is expressed as*

$$y = \begin{cases} 2 + \cos(6\pi x), & \text{if } z = 1, \\ 1 - \cos(4\pi x), & \text{if } z = 2, \\ \cos(2\pi x), & \text{if } z = 3. \end{cases} \tag{14}$$

It is easy to see that the minimum of the function in (14) *is obtained exactly at $z = 3$ and $x = 0.5$.*

To start the proposed ARSD method, we obtain an initial training data of three points, where a three-level full factorial design (Wu and Hamada 2009) is used for the qualitative factor and a random Latin hypercube design (McKay et al. 1979) is used for the quantitative factor. In each iteration of the sequential design, the corresponding output value of the chosen design point is calculated by (14), and the minimum of the obtained output values is regarded as the minimum of (14). For the proposed ARSD method, we choose $\rho = 2$. We have compared the proposed method with different values of $\rho = 0.5, 1, 2, 3$. The

results appear to have similar performance. When we choose $\rho = 2$, $\hat{\mu}_{0|n}(\boldsymbol{w}_0) - \rho\hat{\sigma}_{0|n}(\boldsymbol{w}_0) = \hat{\mu}_{0|n}(\boldsymbol{w}_0) - 2\hat{\sigma}_{0|n}(\boldsymbol{w}_0)$ can be viewed as the lower confidence limit of $Y_0|\boldsymbol{y}_n$ with confidence level around 0.95. In order to obtain the minimum of the response, the proposed method is to minimize the lower confidence limit of $Y_0|\boldsymbol{y}_n$. Thus it is more reasonable than minimizing the mean of $Y_0|\boldsymbol{y}_n$. Hereafter, we choose $\rho = 2$ in the simulation.

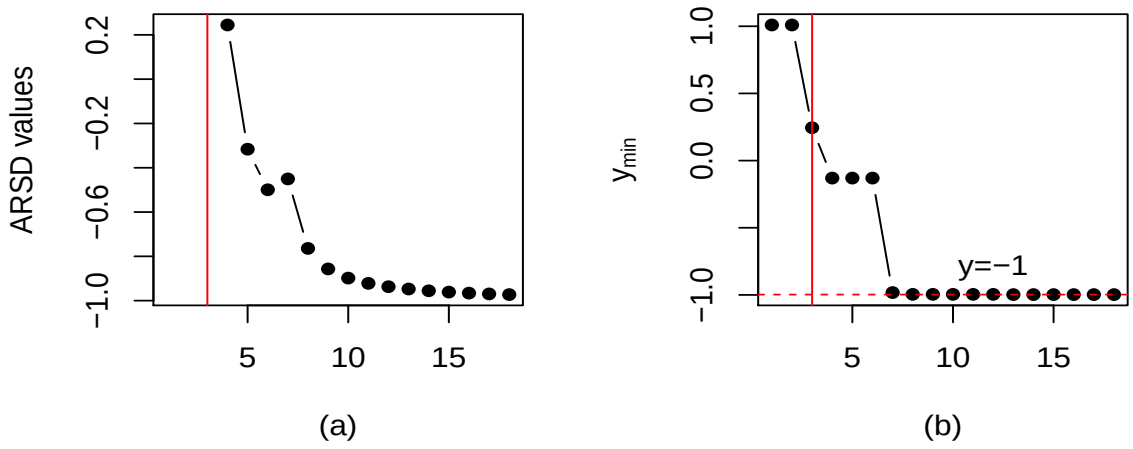


Figure 1: Results of the ARSD method in one simulation trial, (a) the ARSD value, (b) the obtained minimum of response, where 15 points are selected sequentially based on three initial runs.

Figure 1 shows the results of the ARSD sequential design in one simulation, where the three initial points and 15 sequentially added points are on the left and right of the red vertical line respectively. Here the ARSD value in Figure 1(a) represents the value of $\hat{\mu}_{0|n}(\boldsymbol{w}_{n+1}) - \rho\hat{\sigma}_{0|n}(\boldsymbol{w}_{n+1})$ in each iteration. It is seen that the ARSD value converges quickly within 10 iterations of the sequential runs. From Figure 1(b), it is clear that the estimated minimum of the responses drops sharply as the points are added sequentially by the proposed ARSD method.

Note that the true minimum of the function in (14) is -1. The proposed ARSD method achieves the minimum with four iterations of the sequential procedure. Moreover, when the minimum is achieved, the sequential inputs converge at the minimum point. Figure 2 marks

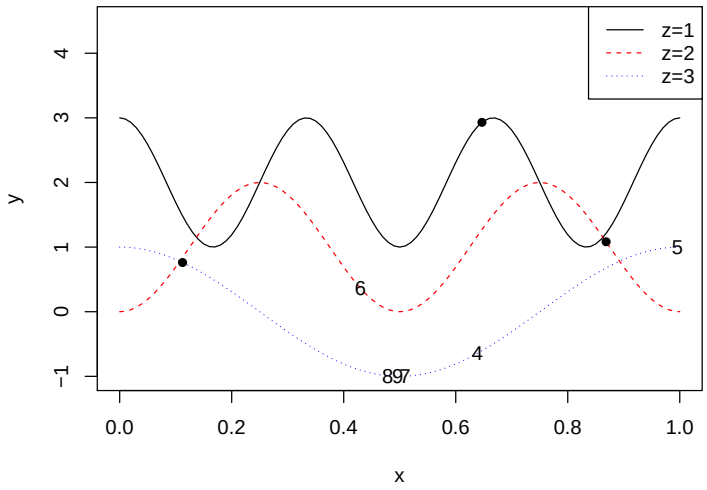


Figure 2: Illustration of the ARSD sequential inputs with three initial points and six sequentially added points.

the selected design points corresponding to the simulation in Figure 1. In Figure 2, the small solid dots are the initial three points, and the points which are labeled "4" to "9" are six sequential points. From Figure 2, the proposed method efficiently allocates the design point to the level $z = 3$ to seek the minimum of response. The points marked with "7", "8", "9" almost coincide at $z = 3$, with their responses values close to the true minimal response value of -1.

To further examine the performance of the proposed ARSD method, it is of interest to understand how the adaptive set of the feasible region behaves. Figure 3 reports the sequential $\mathcal{A}_n$ subsets corresponding to the sequential design in Figures 1 and 2. From Figure 3, one can clearly observe that $\mathcal{A}_n$ quickly converges to the set $\{(x, z) : x \in (0.44, 0.66), z = 3\}$. Note that the true minimum point $(0.5, 3)$ belongs to this set.

We further examine the theoretical results in Theorem 1 and Corollary 1 through obtaining the empirical probability from simulation. Specifically, we set for $\alpha = 0.05$ and conduct 100 replications to record the number of times the inequality in Theorem 1 being held. The empirical probability in Theorem 1 is calculated as the ratio of the number of times Theorem 1 holds to the number of replications. The empirical probability in Corollary 1 is calculated

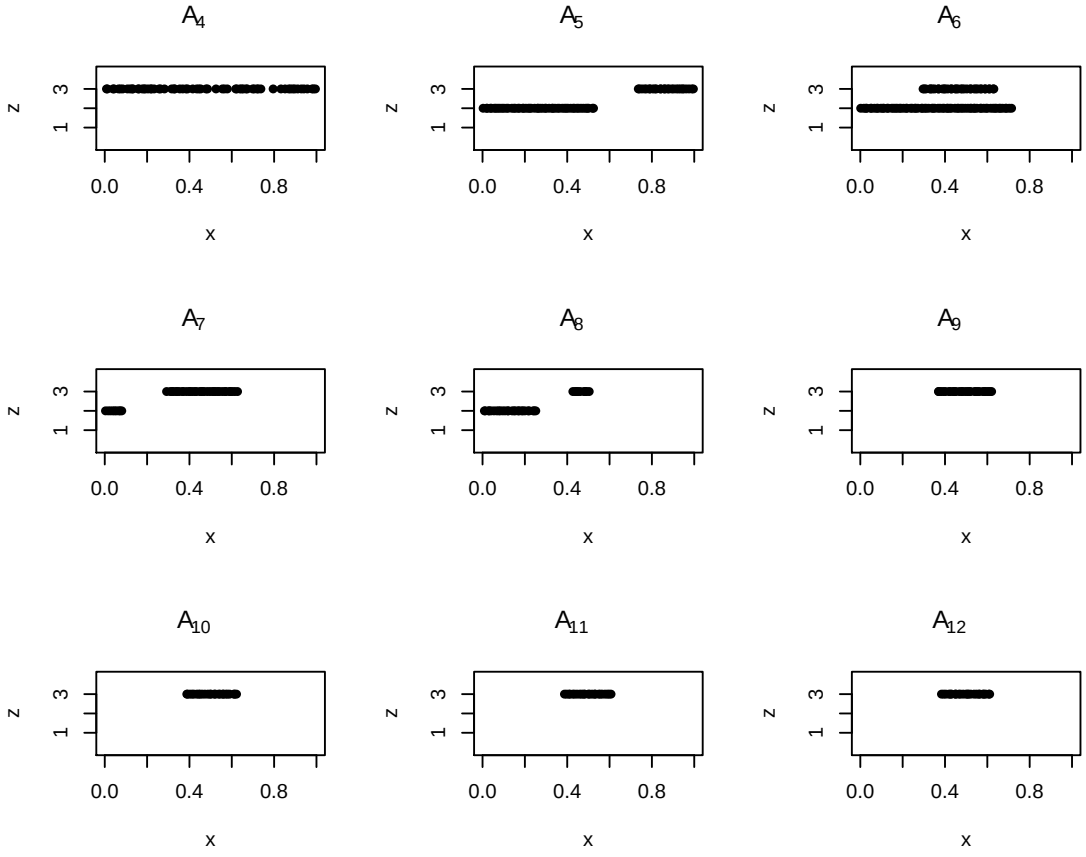


Figure 3: Illustration of the adaptive $\mathcal{A}_n$ design region from one simulation in Example 1.

similarly. The above simulation procedure is repeated 100 times. Figure 4 reports the boxplots of the empirical probabilities for Theorem 1 and Corollary 1. From Figure 4, it is seen that the empirical probability in Theorem 1 is greater than $1 - 4\alpha = 0.8$, and the empirical probability in Corollary 1 is greater than $1 - 3\alpha = 0.85$.

We would like to remark that the performance of sequential design methods will depend on parameter estimation. It is important to check the estimation of the variances of model errors σ_j^2's, correlation $\tau_{r,s}^j$ and other parameter values for the model presented in Section 2. Figure A1 in the Appendix reports the boxplots of the estimates of the parameters in Example 1. From Figure A1, one can see that the estimates tend to be stable when the fifth sequential points are added.

Now we compare the proposed ARSD method with the other four benchmark methods

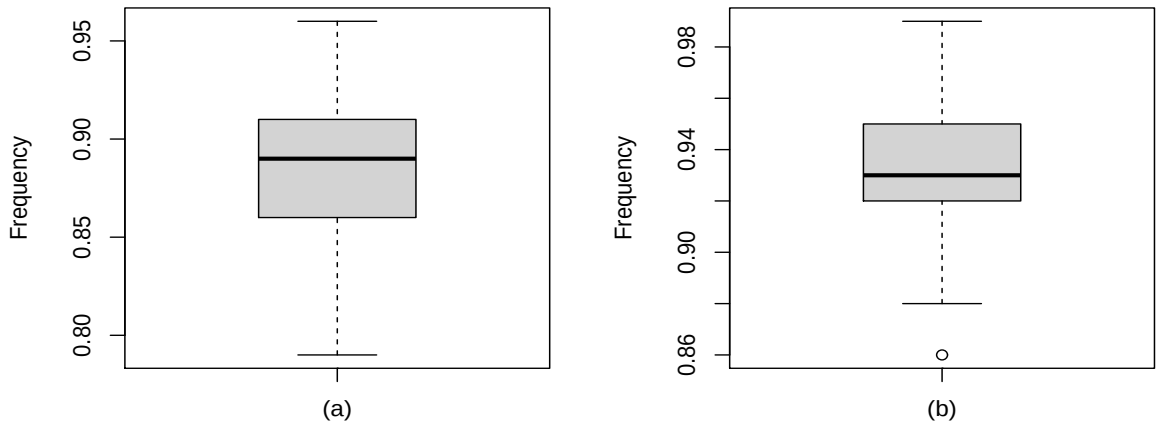


Figure 4: Boxplots of the empirical probability for (a) Theorem 1 and (b) Corollary 1 from 100 simulations of Example 1.

over 100 simulations. For each method, the same number of initial runs of size three is used with a three-level full factorial design for the qualitative factor and a random Latin hypercube design is for the quantitative factor. When the ARSD value is less than 1% of the current ARSD value (Jones et al., 1998) or the maximal number of iterations is achieved, we stop to search the next point.

Figure 5 reports the boxplots and the number of runs until stopping criterion with the maximal number of iterations to be 15. The histograms of the obtained minimums for methods in comparison can be found in the Appendix. From Figure 5, it is seen that the performance of the ARSD is much better than the LCB, EI, MU and SI methods. For the proposed ARSD method, most minimums are near -1. The ARSD method can obtain the true minimum with a higher probability than other methods. We also find out that ARSD generally need a smaller number of runs than other methods.

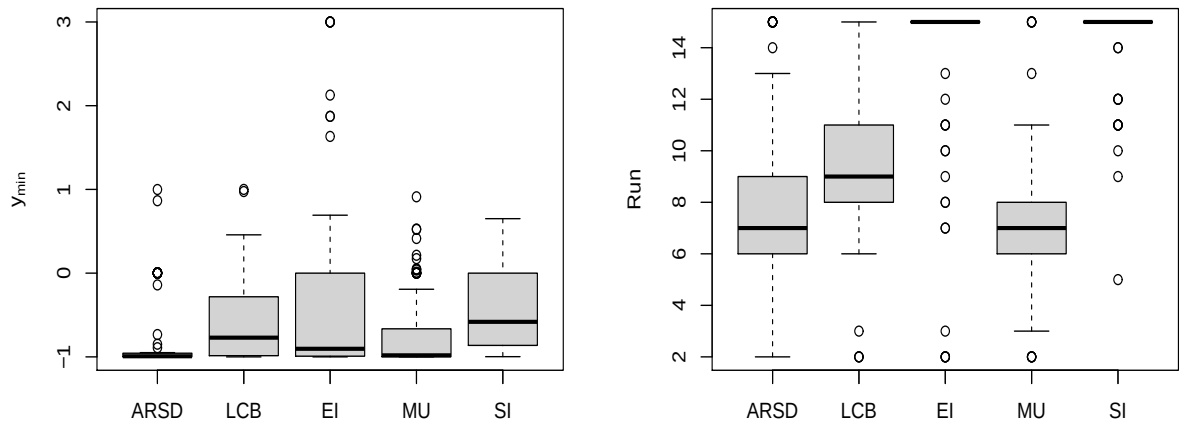


Figure 5: Boxplots of the obtained minimum values of response and boxplots of the number of runs until stopping criterion over 100 simulations in Example 1.

4.2 Examples with Multiple Quantitative and Qualitative Factors

Example 2. *This example is used in Deng et al. (2017) for a computer experiment with* $p = 3$ *quantitative factors and* $q = 3$ *qualitative factors. The response of the experiment has the following expression*

$$y = \sum_{i=1}^{3} \frac{x_i z_{4-i}}{4000} + \prod_{i=1}^{3} \cos\left(\frac{x_i}{\sqrt{i}}\right) \sin\left(\frac{z_{4-i}}{\sqrt{i}}\right), \tag{15}$$

where $-100 < x_i < 100$ *for* $i = 1, \cdots, p$ *and* $z_j = \{-50, 0, 50\}$ *for* $j = 1, \cdots, q$.

Note that the qualitative factors in this example behave as the ordinal factors. In each simulation, a 9-run initial design is adopted, where a three-level fractional factorial design is used for the qualitative factors and a random Latin hypercube design is used for the quantitative factors. It is easy to know that the true minimum of (15) is 3.75. The proposed ARSD method is compared with the LCB, EI, MU and SI methods. For the proposed ARSD design, we choose $\rho = 2$. For each method in comparison, it has the same number of initial runs and then nine follow-up points are obtained sequentially. Figure 6 displays the boxplots of obtained minimums over 100 simulations. The histograms of obtained minimums

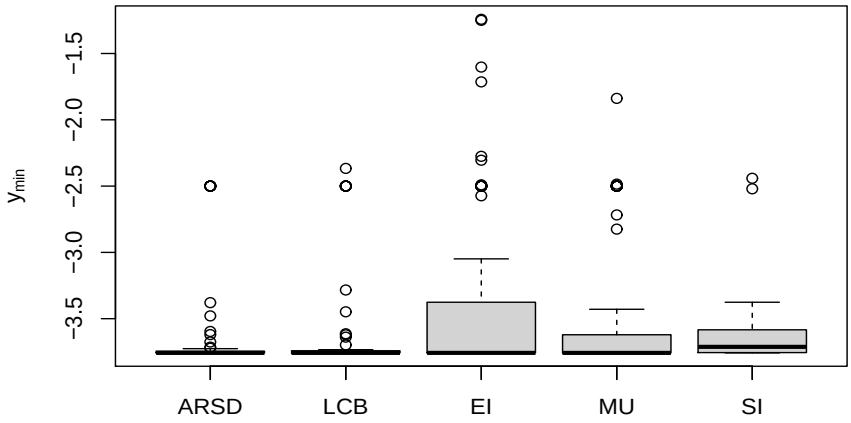


Figure 6: Boxplots of the obtained minimums of the response over 100 simulations in Example 2.

can be found from the Appendix. From Figure 6, it is seen that the proposed ARSD method outperforms the EI, MU and SI methods significantly in terms of the obtained minimal values. We also examine the computational time of the methods in comparison in Table 1. It is seen that that the computational time of the ARSD method is also efficient in comparison with other methods.

Table 1: Average computational time (in mins) over 100 simulations for methods in comparison in Example 2.

Methods	ARSD	LCB	EI	MU	SI
Time	1.03	1.10	1.90	1.18	1.72

5 Case Study of HPC Data

In this section, we apply the proposed ARSD sequential design for studying the HPC systems, which are important infrastructures to advance the industry 4.0. To enhance the performance of the HPC systems, a key step is to understand the HPC variability since there are run-to-run variation in the execution of a computing task. In particular, the input/output (IO) throughput (i.e., data transfer speed) is an important metric, which is affected by various system factors such as CPU frequency, the number of threads, IO operation mode, and IO scheduler. The relationship between the IO throughput (y) and these system factors can be quite complicated. Moreover, some of these system factors are quantitative and some are qualitative.

In this case study, our objective is to find an optimal level combination of system factors that optimizes the IO performance variability measure. Table 2 summarizes the input factors, of which the quantitative factors are the CPU clock frequency (x_1) and the number of threads (x_2), and the qualitative factors are the IO operation mode (z_1) with three levels, the IO scheduler (z_2) with three levels and the VM IO scheduler (z_3) with three levels. Here the IO scheduler is the method that computer operating systems use to decide in which order the block IO operations will be submitted to storage volumes. For the IO operation modes, Initialwrite measures the performance of writing a new file, Randomread measures the performance of reading a file with accesses being made to random locations within the file, and Fwrite measures the performance of writing a file using fwrite() function. The HPC server is configured with a dedicated 2TB HDD on a 2 socket, 4 core (2 hyperthreads/core) Intel Xeon E5-2623 v3 (Haswell) platform with 32 GB DDR4, using Linux operating system. The IOzone benchmark task (Norcott 2020) was used in this computer experiment (Xu et al. 2020).

For a given level combination of input factors as a configuration, the HPC server executes the IOzone benchmark task and the IO throughput (in kilobytes per second) is recorded. By executing for 40 replicates, the mean and the standard deviation (SD) of the 40-replicate IO throughput values are calculated (Cameral et al. 2009). Clearly, a smaller value of the SD indicates the robustness of the HPC system, and a large mean value indicates the

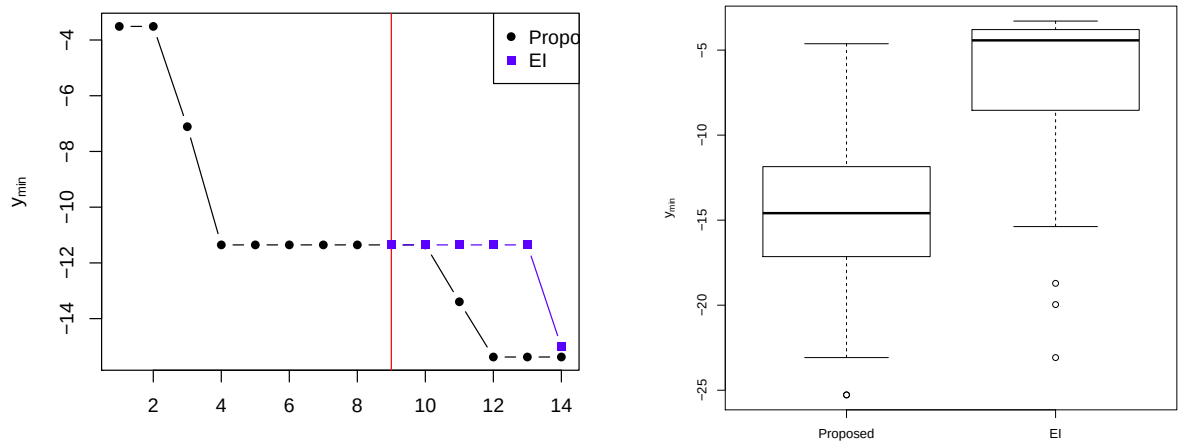
effectiveness of the HPC system. Hence, we consider to use the signal-to-noise ratio Y_{SN}, i.e., the ratio of the mean and SD of the throughput values, as the output response in the optimization.

Table 2: A summary of input factors in the IO throughput experiment of the HPC system.

Category	Variable	No. of levels	Values
Hardware	x_1: CPU clock frequency (GHz)	continuous	1.2, 1.4, 1.5, 1.6, 1.8, 1.9, 2.0, 2.1 2.3, 2.4, 2.5, 2.7, 2.8, 2.9, 3.0
Operating System	z_2: IO scheduler	3	CFQ, DEAD, NOOP
	z_3: VM IO scheduler	3	CFQ, DEAD, NOOP
Application	z_1: IO operation mode	3	Fwrite, Initialwrite, Randomread
	x_2: number of threads	continuous	1, 2, 4, 8, 16, 32, 64, 128, 256

We apply the proposed ARSD sequential design to find the optimal configuration to achieve the maximum of Y_{SN}, the ratio between the mean and SD of the throughput values. It appears that there is little domain knowledge on the configuration (the setting of input factors) to maximize the ratio of the mean and standard deviation of the throughput values. Note that maximizing Y_{SN} is equivalent to minimizing $-Y_{SN}$. Thus we use $-Y_{SN}$ as the response for our proposed ARSD sequential design. For the initial experiment, we consider a 9-run design with a three-level fractional factorial design for the qualitative factors and a random Latin hypercube design for quantitative factors. Then five design points are obtained sequentially by each method in comparison. Here, we focus on the comparison between the proposed ARSD sequential and the EI method. For the proposed ARSD design, we choose $\rho = 2$. Figure 7(a) reports the obtained minimums of $-Y_{SN}$ in one simulation trial with the same initial runs from the ARSD method and the EI method, respectively. One can see that the ARSD sequential design performs much better than the EI design. The ARSD method obtains a smaller value of response than the EI method under the same number of runs. In the case of Figure 7(a), the maximum of Y_{SN} obtained by the proposed ARSD method is

20.198 at the setting $x_1 = 1.2$, $x_2 = 2$, $z_1 =$ "Initialwrite", $z_2 =$ "NOOP", $z_3 =$ "NOOP". It is interesting to note that the number of threads in this optimal setting of maximizing Y_{SN} is $x_2 = 2$. For different initial runs, it happens that the ARSD method performs better than the EI method in most cases. Moreover, we also compare the ARSD sequential design with the EI design in 100 simulations with different initial runs. Figure 7(b) reports the boxplots of minimal response $(-Y_{SN})$ values obtained by the ARSD method in comparison with the EI method. Clearly, the proposed ARSD sequential design is much more efficient than the EI design in finding the maximal value of Y_{SN}.



(a) (b)

Figure 7: The performance of the proposed ARSD method and EI methods for the HPC case study: (a) the sequential procedure in one simulation; (b) the boxplot of the obtained minimums of the response $(-Y_{SN})$ over 100 simulations.

6 Discussion

In this work, we propose an adaptive-region sequential design for optimization of computer experiments with qualitative and quantitative factors. Here we have focused on finding the optimal level combination of factors to minimize (or maximize) the response output.

The proposed adaptive-region sequential design combines the predictive mean and standard derivation to achieve the balance between exploitation and exploration with meaningful interpretation. Moreover, the adaptive design region varies with the collected design points with a theoretical justification based on the bound between the true optimal response and its estimate.

Currently the proposed ARSD criterion is built based on the AGP in Deng et al. (2017) for computer experiments with qualitative and quantitative factors. The proposed methodology can also be applicable for other Gaussian process models for computer experiments with qualitative and quantitative factors. It is worth pointing out that the theoretical justification of the adaptive design region can be extend to the case of contour estimation for computer experiment with qualitative and quantitative factors.

There are several directions for the further development of the proposed method. First, when the number of input variables or the levels of qualitative factors is large, the use of AGP can involve a large number of parameters and consequently the sequential procedure can be slow due to the parameter estimation and the large number of possible design points. One possible solution to overcome this drawback is to adopt a more parsimonious Gaussian process model (Zhang et al. 2020) or construct a sensible initial design. To address the intensive search over a large design region in the optimization in (7), more advanced optimization technique is needed such as the mixed integer programming used in Xie and Deng (2020). Second, there is a tuning parameter ρ in the proposed ARSD criterion in (7). In the numerical study, there seems to be no significant difference on the performance of the proposed method under $\rho = 0.5, 1, 2, 3$. It will be interesting to understand the sensitivity on the choice of the tuning parameter. It will also be interesting to understand how the adaptive design region behaves at different stages of sequential design. Third, when the response is rugged, i.e., there are several local optima, the performance of the ARSD may not be as good as the case of unique optima. We will consider several robust approach to better balance the trade-off between exploration and exploitation. It would be even more interesting to investigate some specific metrics to quantify the balance between exploration and exploitation. Fourth, it will be interesting to investigate more theoretical properties on

the convergence of the estimated optimum. A key challenge lies in the design space to be semi-continuous and semi-discrete. Finally, the proposed method is not limited to the continuous response. It will be interesting to investigate on how to extend the proposed method for computer experiments with non-continuous output such as binary responses (Sung et al. 2020).

Acknowledgement

The work by Cai is supported by the National Natural Science Foundation of China (Grant No. 12001155), and the Natural Science Foundation of Hebei Province of China (Grant No. A2022208001). The work by Lin is supported by the Natural Sciences and Engineering Research Council of Canada. We are grateful to the editor and the referees for their constructive comments that have helped improve the article significantly.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Auer, P (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3, 397-422.

Bichon, B., Mahadevan, S., and Eldred, M. (2009). Reliability-based design optimization using efficient global reliability analysis. *50th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics, and materials conference, AIAA, Palm Springs, California*, 2009-2261.

Bingham, D., Ranjan, P., and Welch, W. J. (2014). Design of computer experiments for

optimization, estimation of function contours, and related objectives. *Statistics in Action: A Canadian Outlook*, 109.

Cameron, K.W., Anwar, A., Cheng, Y., et al. (2019). MOANA: modeling and analyzing I/O variability in parallel system experimental design. *IEEE Transactions on Parallel and Distributed Systems*, 30(8), 1843-1856.

Chen, Z., Mak, S., and Wu, C. F. J. (2019) A hierarchical expected improvement method for Bayesian optimization. *arXiv preprint arXiv*: 1911.07285.

Cortes, C., DeSalvo, G., Gentile, C., Mohri, M., and Zhang, N. (2020). Adaptive region-based active learning. *In International Conference on Machine Learning*, 2144-2153, PMLR.

Deng, X., Hung, Y. and Lin, C. D. (2015). Design for computer experiments with qualitative and quantitative factors. *Statistica Sinica*, 25(4), 1567-1581.

Deng, X., Lin, C. D., Liu, K. W., and Rowe, R. K. (2017). Additive Gaussian process for computer models with qualitative and quantitative factors. *Technometrics*, 59, 283-292.

Fang, K. T., Li, R., and Sudjianto, A. (2005). *Design and Modeling for Computer Experiments*, New York: Chapman & Hall/CRC Press.

Frazier, P. I., Powell, W. B., and Dayanik, S. (2008). A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5), 2410-2439.

Gramacy, R. B. (2020). *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*, New York: Chapman & Hall/CRC Press.

He, Y., Lin, C.D., Sun, F., and Lv, B. (2019). Construction of marginally coupled designs by subspace theory. *Bernoulli*, 25, 2163-2182.

Jala, M., Levy-Leduc, C., Moulines, E., Conil, E., and Wiart, J. (2016). Sequential design of computer experiments for the assessment of fetus exposure to electromagnetic fields. *Technometrics*, 58, 30-42.

Joseph, V. R. (2016). Space-filling designs for computer experiments: A review. *Quality Engineering*, 28(1), 28-35.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13, 455-492.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21, 239-245.

Norcott, W. D. (2020). IOzone filesystem benchmark. http://www.iozone.org/.

Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R. T., and Kim, N. (2010). Adaptive designs of experiments for accurate approximation of a target region. *Journal of Mechanical Design*: 132(7): 071008.

Picheny, V., Gramacy, R. B., Wild, S. M., and Digabel S. L. (2016). Bayesian optimization under mixed constraints with a slack-variable augmented Lagrangian. *arXiv preprint arXiv*: 1605.09466.

Ponweiser, W., Wagner, T., and Vincze, M. (2008). Clustered multiple generalized expected improvement: A novel infill sampling criterion for surrogate models. *2008 IEEE Congress on Evolutionary Computation*: 3515-3522.

Qian, P. Z. (2012). Sliced Latin hypercube designs. *Journal of the American Statistical Association*, 107(497), 393-399.

Ranjan, P., Bingham, D., and Michailidis, G. (2008). Sequential experiment design for contour estimation from complex computer codes. *Technometrics*, 50(4), 527-541.

Ranjan, P., Haynes, R., and Karsten, R. (2011). A computationally stable approach to Gaussian process interpolation of deterministic computer simulation data. *Technometrics*, 53(4), 366-378.

Roy, S. (2008). *Sequential-adaptive design of computer experiments for the estimation of percentiles*. Ohio State University. Ph.D. Thesis.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn., H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, 4, 409-423.

Santner, T. J., Williams, B. J., and Notz,W. I. (2003). *The Design and Analysis of Computer Experiments.* New York: Springer.

Sakellariou, R., Buenabad-Chávez, J., Kavakli, E., Spais, I., and Tountopoulos, V. (2018). High performance computing and industry 4.0: experiences from the DISRUP-T project. *SAMOS '18: Proceedings of the 18th International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation*, Pages 218âĂŞ219, DOI: 10.1145/3229631.3264660.

Sauer A., Gramacy R. B., and Higdon D. (2020). Active learning for deep Gaussian process surrogates. *arXiv preprint arXiv*: 2012.08015.

Scott, W., Frazier, P., and Powell, W. (2011). The correlated knowledge gradient for simulation optimization of continuous parameters using Gaussian process regression. *SIAM Journal on Optimization*, 21(3), 996-1026.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: no regret and experimental design. *arXiv preprint arXiv*: 0912.3995.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2012). Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions of Information Theory*, 58(5), 3250-3265.

Sung, C. L., Hung, Y., Rittase, W., Zhu, C., and Wu, C. F. J. (2020). A generalized Gaussian process model for computer experiments with binary time series. *Journal of the American Statistical Association*, 115(530), 945-956.

Wang, Y., Peng, Z., and Zhang, R., and Xiao, Q. (2021). Robust sequential design for piecewise-stationary multi-armed bandit problem in the presence of outliers. *Statistical Theory and Related Fields*, 5(2), 122-133.

Wang, L., Xiao, Q., and Xu, H. (2018). Optimal maximin L_1 distance Latin hypercube designs based on good lattice point designs. *The Annals of Statistics*, 46, 3741-3766.

Wu, C. F. J., and Hamada, M. (2009). *Experiments: Planning, Analysis, and Optimization*. New York: Wiley.

Xie, W., and Deng, X. (2020). Scalable algorithms for the sparse ridge regression. *SIAM Journal on Optimization*, 30(4), 3359-3386.

Xu, L., Lux, T., Chang, T., Li, B., Hong, Y., Watson, L., Butt, A., Yao, D., and Cameron, K. (2020). Prediction of high-performance computing input/output variability and its application to optimization for system configurations. *Quality Engineering*, DOI:10.1080/08982112.2020.1866203.

Yang, F., Lin, C. D., and Ranjan, P. (2020). Global fitting of the response surface via estimating multiple contours of a simulator. *Journal of Statistical Theory and Practice*, 14, 1-21.

Zhang, Q., Chien, P., Liu Q., Xu L., and Hong, Y. (2020). Mixed-input Gaussian process emulators for computer experiments with a large number of categorical levels. *Journal of Quality Technology*, DOI: 10.1080/00224065.2020.1778431.

Zhou, Q., Qian, P .Z. G., and Zhou, S. (2011). A simple approach to emulation for computer models with qualitative and quantitative factors. *Technometrics*, 53, 266-273.

Appendix

Details of the Maximum Likelihood Estimation of AGP

Under the AGP model (1), the log-likelihood function is

$$l(\mu, \boldsymbol{\sigma}^2, \boldsymbol{T}, \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Phi}| - \frac{1}{2}(\boldsymbol{y}_n - \mu\mathbf{1}_n)^T\boldsymbol{\Phi}^{-1}(\boldsymbol{y}_n - \mu\mathbf{1}_n). \tag{16}$$

Setting the derivative of $l(\mu, \boldsymbol{\sigma}^2, \boldsymbol{T}, \boldsymbol{\theta})$ with respect to μ to be zero, we have the maximum likelihood estimator of μ is

$$\hat{\mu} = \frac{\mathbf{1}_n^T \boldsymbol{\Phi}^{-1} \boldsymbol{y}_n}{\mathbf{1}_n^T \boldsymbol{\Phi}^{-1} \mathbf{1}_n}. \tag{17}$$

Substituting (17) into (16), we obtain

$$\begin{aligned} l(\hat{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{T}, \boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Phi}| - \frac{1}{2}(\boldsymbol{y}_n - \hat{\mu}\mathbf{1}_n)^T \boldsymbol{\Phi}^{-1}(\boldsymbol{y}_n - \hat{\mu}\mathbf{1}_n) \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Phi}| - \frac{1}{2}\boldsymbol{y}_n^T \boldsymbol{\Phi}^{-1}\boldsymbol{y}_n + \frac{1}{2}\frac{(\mathbf{1}_n^T \boldsymbol{\Phi}^{-1}\boldsymbol{y}_n)^2}{\mathbf{1}_n^T \boldsymbol{\Phi}^{-1}\mathbf{1}_n}. \end{aligned}$$

The estimators of $\boldsymbol{\sigma}^2, \boldsymbol{T}, \boldsymbol{\theta}$ can be obtained as

$$(\hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{T}}, \hat{\boldsymbol{\theta}}) = \operatorname{argmin}\left\{\log|\boldsymbol{\Phi}| + \boldsymbol{y}_n^T \boldsymbol{\Phi}^{-1}\boldsymbol{y}_n - \frac{(\mathbf{1}_n^T \boldsymbol{\Phi}^{-1}\boldsymbol{y}_n)^2}{\mathbf{1}_n^T \boldsymbol{\Phi}^{-1}\mathbf{1}_n}\right\}. \tag{18}$$

To ensure the $m_j \times m_j$ matrix $\boldsymbol{T}^{(j)}$ is a valid correlation function, $\boldsymbol{T}^{(j)} = (\tau_{r,s}^{(j)})_{m_j \times m_j}$ must be a positive definite matrix with unit diagonal elements. We apply the hypersphere parameterization approach in Zhou et al. (2011) to quantify the correlations of qualitative factors. The key of this approach is to find a lower triangular matrix $\boldsymbol{L}^{(j)} = (l_{r,s}^{(j)})$ with strictly positive diagonal entries such that

$$\boldsymbol{T}^{(j)} = \boldsymbol{L}^{(j)} \boldsymbol{L}^{(j)T}.$$

For $r = 1$, let $l_{1,1} = 1$. For $r = 2, \cdots, m_j$, the entries of the rth row of $\boldsymbol{L}^{(j)}$ can be constructed as follows:

$$\begin{aligned} &l_{r,1}^{(j)} = \cos(\theta_{r,1}), \\ &l_{r,2}^{(j)} = \sin(\theta_{r,1})\cos(\theta_{r,2}), \\ &\cdots, \\ &l_{r,r-1}^{(j)} = \sin(\theta_{r,1})\cdots\sin(\theta_{r,r-2})\cos(\theta_{r,r-1}), \\ &l_{r,r}^{(j)} = \sin(\theta_{r,1})\cdots\sin(\theta_{r,r-2})\sin(\theta_{r,r-1}), \end{aligned}$$

where $\theta_{r,s} \in (0, \pi)$.

Proof of Lemma 2

Proof. Recall the definition of $y_{\min}, \tilde{\mu}_{\min,n}, \tilde{\mu}^L_{\min,n}, \tilde{\mu}^U_{\min,n}$ as follows

$$y_{\min} = \min_{\boldsymbol{w}_0 \in \mathcal{A}} y(\boldsymbol{w}_0) = \sup_{v \in R}\{v : \frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \int_{\boldsymbol{x}_0 \in \boldsymbol{X}} \mathbf{1}_{\{y(\boldsymbol{w}_0) < v\}} \mathrm{d}\boldsymbol{x}_0 < \epsilon\}, \tag{19}$$

$$\tilde{\mu}_{\min,n} = \min_{\boldsymbol{w}_0 \in \mathcal{A}} \mu_{0|n}(\boldsymbol{w}_0) = \sup_{v \in R}\{v : \frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \int_{\boldsymbol{x}_0 \in \boldsymbol{X}} \mathbf{1}_{\{\mu_{0|n}(\boldsymbol{w}_0) < v\}} \mathrm{d}\boldsymbol{x}_0 < \epsilon\}, \tag{20}$$

$$\tilde{\mu}^L_{\min,n} = \min_{\boldsymbol{w}_0 \in \mathcal{A}} \mu^L_{0|n}(\boldsymbol{w}_0) = \sup_{v \in R}\{v : \frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \int_{\boldsymbol{x}_0 \in \boldsymbol{X}} \mathbf{1}_{\{\mu^L_{0|n}(\boldsymbol{w}_0) < v\}} \mathrm{d}\boldsymbol{x}_0 < \epsilon\}, \tag{21}$$

$$\tilde{\mu}^U_{\min,n} = \min_{\boldsymbol{w}_0 \in \mathcal{A}} \mu^U_{0|n}(\boldsymbol{w}_0) = \sup_{v \in R}\{v : \frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \int_{\boldsymbol{x}_0 \in \boldsymbol{X}} \mathbf{1}_{\{\mu^U_{0|n}(\boldsymbol{w}_0) < v\}} \mathrm{d}\boldsymbol{x}_0 < \epsilon\}, \tag{22}$$

where $\epsilon > 0$ is a small positive number, and $\mathbf{1}_{\{\cdot\}}$ is the indicator function.

By Lemma 1, for a given $\boldsymbol{x}_0 \in \boldsymbol{X}$, we have

$$P\left(\mu^L_{0|n}(\boldsymbol{w}_0) \le y(\boldsymbol{w}_0) \le \mu^U_{0|n}(\boldsymbol{w}_0), \quad \forall \boldsymbol{z}_0 \in \boldsymbol{Z}, \forall n \ge 1\right) \ge 1 - \alpha.$$

When $\mu^L_{0|n}(\boldsymbol{w}_0) \le y(\boldsymbol{w}_0) \le \mu^U_{0|n}(\boldsymbol{w}_0)$, for given $\boldsymbol{x}_0 \in \boldsymbol{X}$ and all $v \in R$,

$$\frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \mathbf{1}_{\{\mu^U_{0|n}(\boldsymbol{w}_0) < v\}} \le \frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \mathbf{1}_{\{y(\boldsymbol{w}_0) < v\}} \le \frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \mathbf{1}_{\{\mu^L_{0|n}(\boldsymbol{w}_0) < v\}}, \forall n \ge 1.$$

Integrating each term over $\boldsymbol{x}_0$ in $\boldsymbol{X}$, we have

$$\begin{aligned} \int_{\boldsymbol{x}_0 \in \boldsymbol{X}} \frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \mathbf{1}_{\{\mu^U_{0|n}(\boldsymbol{w}_0) < v\}} \mathrm{d}\boldsymbol{x}_0 &\le \int_{\boldsymbol{x}_0 \in \boldsymbol{X}} \frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \mathbf{1}_{\{y(\boldsymbol{w}_0) < v\}} \mathrm{d}\boldsymbol{x}_0 \\ &\le \int_{\boldsymbol{x}_0 \in \boldsymbol{X}} \frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \mathbf{1}_{\{\mu^L_{0|n}(\boldsymbol{w}_0) < v\}} \mathrm{d}\boldsymbol{x}_0, \forall n \ge 1. \end{aligned} \tag{23}$$

Let $v = \tilde{\mu}^L_{\min,n}$ given in (21), we get

$$\int_{\boldsymbol{x}_0 \in \boldsymbol{X}} \frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \mathbf{1}_{\{y(\boldsymbol{w}_0) < \tilde{\mu}^L_{\min,n}\}} \mathrm{d}\boldsymbol{x}_0 < \epsilon.$$

By (19), we obtain that $\tilde{\mu}^L_{\min,n} \le y_{\min}$. Thus

$$\begin{aligned} & P\left(\tilde{\mu}^L_{\min,n} \le y_{\min}, \forall n \ge 1\right) \\ \ge & P\left(\mu^L_{0|n}(\boldsymbol{w}_0) \le y(\boldsymbol{w}_0) \le \mu^U_{0|n}(\boldsymbol{w}_0), \forall \boldsymbol{z}_0 \in \boldsymbol{Z}, \forall n \ge 1\right) \\ \ge & 1 - \alpha. \end{aligned} \tag{24}$$

Similarly, in (23), let $v = y_{\min}$, we get that

$$\int_{\boldsymbol{x}_0 \in \boldsymbol{X}} \frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \mathbf{1}_{\{\mu^U_{0|n}(\boldsymbol{w}_0) < y_{\min}\}} \mathrm{d}\boldsymbol{x}_0 < \epsilon.$$

By (22), we obtain that $y_{\min} \leq \tilde{\mu}^U_{\min,n}$, thus

$$\begin{aligned} & P\left(y_{\min} \leq \tilde{\mu}^U_{\min,n}, \forall n \geq 1\right) \\ \geq & P\left(\mu^L_{0|n}(\boldsymbol{w}_0) \leq y(\boldsymbol{w}_0) \leq \mu^U_{0|n}(\boldsymbol{w}_0), \forall \boldsymbol{z}_0 \in \boldsymbol{Z}, \forall n \geq 1\right) \\ \geq & 1 - \alpha. \end{aligned} \tag{25}$$

By (24) and (25), we have

$$\begin{aligned} & P\left(y_{\min} \in [\tilde{\mu}^L_{\min,n}, \tilde{\mu}^U_{\min,n}], \forall n \geq 1\right) \\ = & P\left(\tilde{\mu}^L_{\min,n} \leq y_{\min} \leq \tilde{\mu}^U_{\min,n}, \forall n \geq 1\right) \\ \geq & 1 - 2\alpha. \end{aligned}$$

□

Proof of Theorem 1

Proof. By (22), for any $n \geq 1$,

$$\frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \int_{\boldsymbol{x}_0 \in \boldsymbol{X}} \mathbf{1}_{\{\mu^U_{0|n}(\boldsymbol{w}_0) < \tilde{\mu}^U_{\min,n}\}} \mathrm{d}\boldsymbol{x}_0 < \epsilon.$$

Define

$$\mathcal{A}_n(\boldsymbol{z}_0) = \{\boldsymbol{x}_0 \in \boldsymbol{X} : \mu^L_{0|n}(\boldsymbol{x}_0, \boldsymbol{z}_0) \leq \tilde{\mu}^U_{\min,n}\}.$$

It is clear that $\mathcal{A}_n = \{(\boldsymbol{x}_0, \boldsymbol{z}_0) : \boldsymbol{z}_0 \in \boldsymbol{Z}, \boldsymbol{x}_0 \in \mathcal{A}_n(\boldsymbol{z}_0)\}$, then we have

$$\frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \int_{\boldsymbol{x}_0 \in \mathcal{A}_n(\boldsymbol{z}_0)} \mathbf{1}_{\{\mu^U_{0|n}(\boldsymbol{w}_0) < \tilde{\mu}^U_{\min,n}\}} \mathrm{d}\boldsymbol{x}_0 = \frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \int_{\boldsymbol{x}_0 \in \boldsymbol{X}} \mathbf{1}_{\{\mu^U_{0|n}(\boldsymbol{w}_0) < \tilde{\mu}^U_{\min,n}\}} \mathrm{d}\boldsymbol{x}_0 < \epsilon.$$

Since $\mu^U_{0|n}(\boldsymbol{w}_0) = \mu_{0|n}(\boldsymbol{w}_0) + \sqrt{\beta_{0|n}}\sigma_{0|n}(\boldsymbol{w}_0)$, we get that

$$\begin{aligned} & \frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \int_{\boldsymbol{x}_0 \in \mathcal{A}_n(\boldsymbol{z}_0)} \mathbf{1}_{\{\mu_{0|n}(\boldsymbol{w}_0) < \tilde{\mu}^U_{\min,n} - \sqrt{\beta_{0|n}} \sup_{\boldsymbol{w}_0 \in \mathcal{A}_n} \sigma_{0|n}(\boldsymbol{w}_0)\}} \mathrm{d}\boldsymbol{x}_0 \\ \leq & \frac{1}{M} \sum_{\boldsymbol{z}_0 \in \boldsymbol{Z}} \int_{\boldsymbol{x}_0 \in \mathcal{A}_n(\boldsymbol{z}_0)} \mathbf{1}_{\{\mu_{0|n}(\boldsymbol{w}_0) < \tilde{\mu}^U_{\min,n} - \sqrt{\beta_{0|n}}\sigma_{0|n}(\boldsymbol{w}_0)\}} \mathrm{d}\boldsymbol{x}_0 \\ < & \epsilon. \end{aligned}$$

By the definition of $\mathcal{A}_n(\boldsymbol{z}_0)$,

$$\begin{aligned}
&\frac{1}{M}\sum_{\boldsymbol{z}_0\in\boldsymbol{Z}}\int_{\boldsymbol{x}_0\in\boldsymbol{X}}\mathbf{1}_{\{\mu_{0|n}(\boldsymbol{w}_0)<\tilde{\mu}^U_{\min,n}-\sqrt{\beta_{0|n}}\sup_{\boldsymbol{w}_0\in\mathcal{A}_n}\sigma_{0|n}(\boldsymbol{w}_0)\}}\mathrm{d}\boldsymbol{x}_0\\
=&\frac{1}{M}\sum_{\boldsymbol{z}_0\in\boldsymbol{Z}}\int_{\boldsymbol{x}_0\in\mathcal{A}_n(\boldsymbol{z}_0)}\mathbf{1}_{\{\mu_{0|n}(\boldsymbol{w}_0)<\tilde{\mu}^U_{\min,n}-\sqrt{\beta_{0|n}}\sup_{\boldsymbol{w}_0\in\mathcal{A}_n}\sigma_{0|n}(\boldsymbol{w}_0)\}}\mathrm{d}\boldsymbol{x}_0\\
<&\epsilon.
\end{aligned}$$

Thus,

$$\frac{1}{M}\sum_{\boldsymbol{z}_0\in\boldsymbol{Z}}\int_{\boldsymbol{x}_0\in\boldsymbol{X}}\mathbf{1}_{\{\mu_{0|n}(\boldsymbol{w}_0)<\tilde{\mu}^U_{\min,n}-\sqrt{\beta_{0|n}}\sup_{\boldsymbol{w}_0\in\mathcal{A}_n}\sigma_{0|n}(\boldsymbol{w}_0)\}}\mathrm{d}\boldsymbol{x}_0<\epsilon.$$

By (20), we obtain that

$$\tilde{\mu}_{\min,n}\geq\tilde{\mu}^U_{\min,n}-\sqrt{\beta_{0|n}}\sup_{\boldsymbol{w}_0\in\mathcal{A}_n}\sigma_{0|n}(\boldsymbol{w}_0).$$

By Lemma 2, we have $P\left(\tilde{\mu}^U_{\min,n}\geq y_{\min},\forall n\geq 1\right)\geq 1-2\alpha$. Thus,

$$P\left(\tilde{\mu}_{\min,n}\geq y_{\min}-\sqrt{\beta_{0|n}}\sup_{\boldsymbol{w}_0\in\mathcal{A}_n}\sigma_{0|n}(\boldsymbol{w}_0),\forall n\geq 1\right)\geq 1-2\alpha, \tag{26}$$

where $\beta_{0|n}=2\log\frac{\pi^2n^2M}{6\alpha'}$.

On the other hand, by (20), for any $n\geq 1$,

$$\frac{1}{M}\sum_{\boldsymbol{z}_0\in\boldsymbol{Z}}\int_{\boldsymbol{x}_0\in\boldsymbol{X}}\mathbf{1}_{\{\mu_{0|n}(\boldsymbol{w}_0)<\tilde{\mu}_{\min,n}\}}\mathrm{d}\boldsymbol{x}_0<\epsilon.$$

By the definition of $\mathcal{A}_n(\boldsymbol{z}_0)$ and $\tilde{\mu}^{\mathcal{A}}_n\leq\tilde{\mu}_{\min,n}\leq\tilde{\mu}^U_{\min,n}$, we get that

$$\frac{1}{M}\sum_{\boldsymbol{z}_0\in\boldsymbol{Z}}\int_{\boldsymbol{x}_0\in\mathcal{A}_n(\boldsymbol{z}_0)}\mathbf{1}_{\{\mu_{0|n}(\boldsymbol{w}_0)<\tilde{\mu}_{\min,n}\}}\mathrm{d}\boldsymbol{x}_0=\frac{1}{M}\sum_{\boldsymbol{z}_0\in\boldsymbol{Z}}\int_{\boldsymbol{x}_0\in\boldsymbol{X}}\mathbf{1}_{\{\mu_{0|n}(\boldsymbol{w}_0)<\tilde{\mu}_{\min,n}\}}\mathrm{d}\boldsymbol{x}_0<\epsilon.$$

Since $\mu^L_{0|n}(\boldsymbol{w}_0)=\mu_{0|n}(\boldsymbol{w}_0)-\sqrt{\beta_{0|n}}\sigma_{0|n}(\boldsymbol{w}_0)$, we get that

$$\begin{aligned}
&\frac{1}{M}\sum_{\boldsymbol{z}_0\in\boldsymbol{Z}}\int_{\boldsymbol{x}_0\in\mathcal{A}_n(\boldsymbol{z}_0)}\mathbf{1}_{\{\mu^L_{0|n}(\boldsymbol{w}_0)<\tilde{\mu}_{\min,n}-\sqrt{\beta_{0|n}}\sup_{\boldsymbol{w}_0\in\mathcal{A}_n}\sigma_{0|n}(\boldsymbol{w}_0)\}}\mathrm{d}\boldsymbol{x}_0\\
=&\frac{1}{M}\sum_{\boldsymbol{z}_0\in\boldsymbol{Z}}\int_{\boldsymbol{x}_0\in\mathcal{A}_n(\boldsymbol{z}_0)}\mathbf{1}_{\{\mu_{0|n}(\boldsymbol{w}_0)-\sqrt{\beta_{0|n}}\sigma_{0|n}(\boldsymbol{w}_0)<\tilde{\mu}_{\min,n}-\sqrt{\beta_{0|n}}\sup_{\boldsymbol{w}_0\in\mathcal{A}_n}\sigma_{0|n}(\boldsymbol{w}_0)\}}\mathrm{d}\boldsymbol{x}_0\\
\leq&\frac{1}{M}\sum_{\boldsymbol{z}_0\in\boldsymbol{Z}}\int_{\boldsymbol{x}_0\in\mathcal{A}_n(\boldsymbol{z}_0)}\mathbf{1}_{\{\mu_{0|n}(\boldsymbol{w}_0)-\sqrt{\beta_{0|n}}\sup_{\boldsymbol{w}_0\in\mathcal{A}_n}\sigma_{0|n}(\boldsymbol{w}_0)<\tilde{\mu}_{\min,n}-\sqrt{\beta_{0|n}}\sup_{\boldsymbol{w}_0\in\mathcal{A}_n}\sigma_{0|n}(\boldsymbol{w}_0)\}}\mathrm{d}\boldsymbol{x}_0\\
<&\epsilon.
\end{aligned}$$

By $\tilde{\mu}_n^{\mathcal{A}} \leq \tilde{\mu}_{\min,n} \leq \tilde{\mu}_{\min,n}^U$ and the definition of $\mathcal{A}_n(\boldsymbol{z}_0)$,

$$\begin{aligned}
&\frac{1}{M}\sum_{\boldsymbol{z}_0\in\boldsymbol{Z}}\int_{\boldsymbol{x}_0\in\boldsymbol{X}} \mathbf{1}_{\{\mu_{0|n}^L(\boldsymbol{w}_0)<\tilde{\mu}_{\min,n}-\sqrt{\beta_{0|n}}\sup_{\boldsymbol{w}_0\in\mathcal{A}_n}\sigma_{0|n}(\boldsymbol{w}_0)\}}\mathrm{d}\boldsymbol{x}_0\\
=&\frac{1}{M}\sum_{\boldsymbol{z}_0\in\boldsymbol{Z}}\int_{\boldsymbol{x}_0\in\mathcal{A}_n(\boldsymbol{z}_0)} \mathbf{1}_{\{\mu_{0|n}^L(\boldsymbol{w}_0)<\tilde{\mu}_{\min,n}-\sqrt{\beta_{0|n}}\sup_{\boldsymbol{w}_0\in\mathcal{A}_n}\sigma_{0|n}(\boldsymbol{w}_0)\}}\mathrm{d}\boldsymbol{x}_0\\
<&\epsilon.
\end{aligned}$$

Thus,

$$\frac{1}{M}\sum_{\boldsymbol{z}_0\in\boldsymbol{Z}}\int_{\boldsymbol{x}_0\in\boldsymbol{X}} \mathbf{1}_{\{\mu_{0|n}^L(\boldsymbol{w}_0)<y_{\min}-\sqrt{\beta_{0|n}}\sup_{\boldsymbol{w}_0\in\mathcal{A}_n}\sigma_{0|n}(\boldsymbol{w}_0)\}}\mathrm{d}\boldsymbol{x}_0<\epsilon.$$

By (21), we obtain that

$$\tilde{\mu}_{\min,n}^L \geq \tilde{\mu}_{\min,n} - \sqrt{\beta_{0|n}}\sup_{\boldsymbol{w}_0\in\mathcal{A}_n}\sigma_{0|n}(\boldsymbol{w}_0).$$

By Lemma 2, we have

$$P\left(y_{\min} \geq \tilde{\mu}_{\min,n} - \sqrt{\beta_{0|n}}\sup_{\boldsymbol{w}_0\in\mathcal{A}_n}\sigma_{0|n}(\boldsymbol{w}_0), \forall n\geq 1\right)\geq 1-2\alpha. \tag{27}$$

Using (26) and (27), the conclusion of Theorem 1 is proved. □

Proof of Corollary 1

Proof. By Lemma 1, for the minimum point $\boldsymbol{w}^*\in\mathcal{A}$, we have

$$P\left(\mu_{0|n}^L(\boldsymbol{w}^*)\leq y(\boldsymbol{w}^*)\leq\mu_{0|n}^U(\boldsymbol{w}^*),\forall n\geq 1\right)\geq 1-\alpha.$$

By the definition of $y_{\min}$ in (19), we have $y_{\min}=y(\boldsymbol{w}^*)$. Thus

$$P\left(\mu_{0|n}^L(\boldsymbol{w}^*)\leq y_{\min}\leq\mu_{0|n}^U(\boldsymbol{w}^*),\forall n\geq 1\right)\geq 1-\alpha.$$

By Lemma 2, we have

$$P\left(\tilde{\mu}_{\min,n}^L\leq y_{\min}\leq\tilde{\mu}_{\min,n}^U,\forall n\geq 1\right)\geq 1-2\alpha.$$

Thus

$$
\begin{aligned}
&P\left(\mu^L_{0|n}(\boldsymbol{w}^*) \le \tilde{\mu}^U_{\min,n}, \forall n \ge 1\right) \\
\ge &P\left(\mu^L_{0|n}(\boldsymbol{w}^*) \le y_{\min} \le \mu^U_{0|n}(\boldsymbol{w}^*), \tilde{\mu}^L_{\min,n} \le y_{\min} \le \tilde{\mu}^U_{\min,n}, \forall n \ge 1\right) \\
\ge &1 - 3\alpha.
\end{aligned}
$$

It implies the conclusion of Corollary 1. □

Additional Results in Simulation

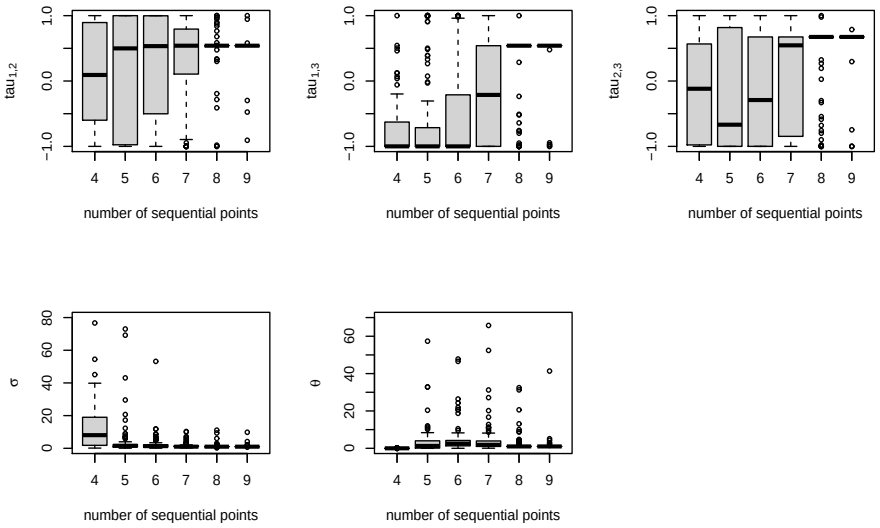


Figure A1: Boxplots of parameter estimates over 100 simulations in Example 1.

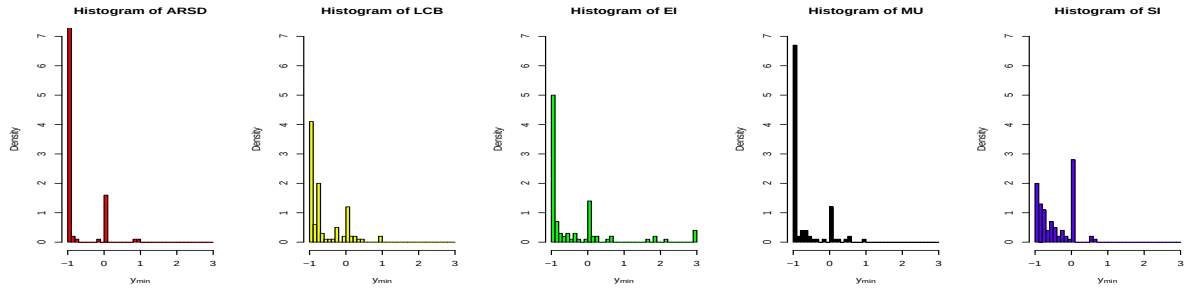


Figure A2: Histograms of the obtained minimum values of response over 100 simulations in Example 1.

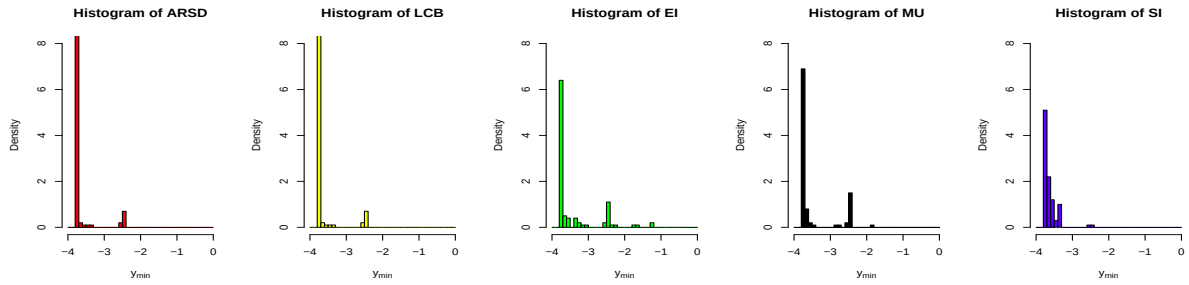


Figure A3: Histograms of the obtained minimums of the response over 100 simulations in Example 2.