Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

SEQUENTIAL EXPERIMENTAL DESIGNS FOR STOCHASTIC KRIGING

Xi Chen

Industrial and Systems Engineering
Virginia Tech
Blacksburg, VA 24601, USA

Qiang Zhou

Systems Engineering and Engineering Management
City University of Hong Kong
Kowloon Tong, HK

ABSTRACT

Recently the stochastic kriging (SK) methodology proposed by Ankenman et al. (2010) has emerged as an effective metamodeling tool for approximating a mean response surface implied by a stochastic simulation. Although fruitful results have been achieved through bridging applications and theoretical investigations of SK, there lacks a unified account of efficient simulation experimental design strategies for applying SK metamodeling techniques. In this paper, we propose a sequential experimental design framework for applying SK to predicting performance measures of complex stochastic systems. This framework is flexible; i.e., it can incorporate a variety of design criteria. We propose several novel design criteria under the proposed framework, and compare the performance with that of classic non-sequential designs. The evaluation uses illustrative test functions and the well-known M/M/1 and the (s,S) inventory system simulation models.

1 INTRODUCTION

To build a high-quality metamodel with a given computational budget to expend, a carefully designed simulation experiment is critical. The literature on experimental designs for deterministic computer experiments abounds and various design schemes have been proposed, for instance, Latin hypercube designs (LHDs) (McKay et al. 1979), orthogonal array based LHDs (Tang 1993), uniform designs (Fang et al. 2000), and maximum entropy designs (Shewry and Wynn 1987), to name a few. Despite earlier efforts made by some researchers (e.g., Ng and Yin 2012, van Beers and Kleijnen 2008), there has been no systematic account of experimental designs for stochastic simulation. In particular, very little work has been done on devising efficient experimental designs for building efficient SK metamodels.

As compared to designs for deterministic computer experiments (in such experiments the same output is produced if the simulation is run twice at the same design point), an efficient simulation design for SK or for stochastic simulation at large is much more challenging to construct, as one needs to determine not only the design-point locations to conduct simulation runs but also the amount of computational effort to be expended at each point. While non-sequential designs that choose all design points up front (e.g., most space-filling designs) may be considered sufficient for deterministic computer experiments, a sequential design strategy is arguably more efficient for stochastic simulation. Sequential designs offer a huge advantage over non-sequential ones in that they improve budget allocation efficiency and reduce waste of computing resources—they permit learning information from previous simulation runs and consequently allocate the remaining simulation budget more wisely. In this paper, we aim to provide the first step toward establishing a general sequential design framework for implementing SK techniques for stochastic simulation and show the benefit of using sequential designs over non-sequential ones.

The remainder of this paper is organized as follows. Section 2 provides a review on SK. In Section 3, a general sequential design framework and some design criteria are proposed for implementing SK techniques in the context of stochastic simulation. The predictive performances of SK using different sequential design

978-1-4799-7486-3/14/$31.00 ©2014 IEEE

criteria under the proposed framework are compared through some illustrative examples in Section 5. Section 6 concludes the paper.

2 A REVIEW ON STOCHASTIC KRIGING

The theoretical aspects of stochastic kriging (SK) prediction are briefly introduced in this subsection before it is applied to the research questions to follow. Standard SK models the simulation response estimate obtained at a design point $\mathbf{x} \in \mathscr{X} \subset \mathbb{R}^d$ on the jth simulation replication as

$$\mathscr{Y}_j(\mathbf{x}) = \mathsf{Y}(\mathbf{x}) + \varepsilon_j(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \beta + \mathsf{M}(\mathbf{x}) + \varepsilon_j(\mathbf{x}) \ , \tag{1}$$

where $\mathsf{Y}(\mathbf{x})$ represents the unknown true response that we intend to estimate at point $\mathbf{x} \in \mathscr{X}$, and the term $\varepsilon_j(\mathbf{x})$ represents the mean-zero simulation error realized on the jth replication.

The terms $\mathbf{f}(\cdot)$ and β in (1) are, respectively, a $p \times 1$ vector of known functions of $\mathbf{x}$ and a $p \times 1$ vector of unknown parameters. The term $\mathsf{M}(\cdot)$ represents a second-order stationary mean-zero Gaussian random field (Santner et al. 2003; Kleijnen 2008). That is, the spatial covariance between any two points in the random field is typically modeled as

$$\mathsf{Cov}(\mathsf{M}(\mathbf{x}), \mathsf{M}(\mathbf{y})) = \tau^2 \mathscr{R}(\mathbf{x} - \mathbf{y}; \theta) \ , \tag{2}$$

where τ^2 can be interpreted as the spatial variance of the random process $\mathsf{M}(\cdot)$ at all $\mathbf{x}$. The spatial correlation function $\mathscr{R}(\cdot;\theta)$ is assumed to be anisotropic; it determines the smoothness properties of $\mathsf{M}(\cdot)$ and it depends on $\mathbf{x}$ and $\mathbf{y}$ only through their difference. The parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_d)^\top$ controls how quickly the spatial correlation diminishes as the two points become farther apart, and it measures the roughness of the underlying response surface in each direction. Ankenman et al. (2010) refer to the stochastic nature of M as *extrinsic uncertainty*, in contrast to the *intrinsic uncertainty* represented by ε that is inherent in a stochastic simulation output, and they assume that M and ε are independent.

The simulation errors $\varepsilon_1(\mathbf{x}), \varepsilon_2(\mathbf{x}), \ldots$ are assumed to be independent and identically distributed across replications at a given design point, and the variance of $\varepsilon_j(\mathbf{x})$ may depend on $\mathbf{x}$. Notice that the simulation output $\mathscr{Y}_j(\mathbf{x})$ could be composed of a large number of more basic random variables obtained on the jth simulation replication. For instance, $\mathscr{Y}_j(\mathbf{x})$ represents the average of hundreds of individual simulated waiting times of incoming calls to a call center on the jth replication when the service rate is $\mathbf{x}$ calls per hour. Hence, the normality of $\varepsilon_j(\mathbf{x})$ could be anticipated.

An experimental design for SK consists a set of design points to run independent simulations and the corresponding numbers of replications to apply, i.e., $\mathscr{D} = \{(\mathbf{x}_i, n_i)_{i=1}^k\}$. Denote the $k \times 1$ vector of the sample averages of simulation responses by $\bar{\mathscr{Y}} = \left(\bar{\mathscr{Y}}(\mathbf{x}_1), \bar{\mathscr{Y}}(\mathbf{x}_2), \ldots, \bar{\mathscr{Y}}(\mathbf{x}_k)\right)^\top$, in which

$$\bar{\mathscr{Y}}(\mathbf{x}_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathscr{Y}_j(\mathbf{x}_i) = \mathsf{Y}(\mathbf{x}_i) + \bar{\varepsilon}(\mathbf{x}_i), \text{ and } \bar{\varepsilon}(\mathbf{x}_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} \varepsilon_j(\mathbf{x}_i) \quad i = 1, 2, \ldots, k \ . \tag{3}$$

That is, $\bar{\mathscr{Y}}(\mathbf{x}_i)$ is the resulting point estimate of the performance measure of interest obtained at design point $\mathbf{x}_i$ and $\bar{\varepsilon}(\mathbf{x}_i)$ is the simulation error associated with it. We write $\bar{\varepsilon}$ as a shorthand for the vector $(\bar{\varepsilon}(\mathbf{x}_1), \bar{\varepsilon}(\mathbf{x}_2), \ldots, \bar{\varepsilon}(\mathbf{x}_k))^\top$.

To do global prediction, standard SK prescribes using the the best linear unbiased predictor of $\mathsf{Y}(\mathbf{x}_0)$ that has the minimum mean squared error (MSE) among all linear unbiased predictors (see Appendix A.1 of Chen et al. 2012),

$$\widehat{\mathsf{Y}}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)^\top \widehat{\beta} + \Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot)^\top \Sigma^{-1} \left(\bar{\mathscr{Y}} - \mathbf{F}\widehat{\beta}\right), \tag{4}$$

where $\widehat{\beta} = \left(\mathbf{F}^\top \Sigma^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^\top \Sigma^{-1} \bar{\mathscr{Y}}$ is the generalized least squares estimator of β, $\Sigma = \Sigma_{\mathsf{M}} + \Sigma_{\varepsilon}$, and $\mathbf{F} = \left(\mathbf{f}(\mathbf{x}_1)^\top, \mathbf{f}(\mathbf{x}_2)^\top, \ldots, \mathbf{f}(\mathbf{x}_k)^\top\right)^\top$. The corresponding MSE follows as

$$\mathrm{MSE}(\widehat{\mathsf{Y}}(\mathbf{x}_0)) = \Sigma_{\mathsf{M}}(\mathbf{x}_0, \mathbf{x}_0) - \Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot)^\top \Sigma^{-1} \Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot) + \zeta^\top (\mathbf{F}^\top \Sigma^{-1} \mathbf{F})^{-1} \zeta, \tag{5}$$

with $\zeta = \mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top \Sigma^{-1} \Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot)$. MSE is a widely used predictive accuracy measure, and the sequential design criteria to be introduced in Section 4 are developed on it. We now elaborate on Σ_{M}, $\Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot)$ and Σ_ε in (4) and (5). The $k \times k$ matrix Σ_{M} records spatial covariances across the design points, i.e., its (i,h)th entry $\Sigma_{\mathsf{M}}(\mathbf{x}_i, \mathbf{x}_h)$ gives $\text{Cov}(\mathsf{M}(\mathbf{x}_i), \mathsf{M}(\mathbf{x}_h))$ as specified in (2). The $k \times 1$ vector $\Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot)$ contains the spatial covariances between the k design points and a given prediction point $\mathbf{x}_0$. The $k \times k$ matrix Σ_ε is the variance-covariance matrix of the vector of simulation errors associated with the vector of point estimates $\bar{\mathscr{Y}}$, $\bar{\varepsilon}$. Common random numbers, CRN, is a widely used variance reduction technique that tends to introduce positive correlation between simulation outputs obtained at a pair of distinctive design points on the same replication. As the use of CRN does not necessarily improve the predictive performance of SK (**?**), we consider only independent simulations at distinct design points in the proposed research. If a common number of simulation replications n is used at all k design points, then Σ_ε is reduced to a $k \times k$ diagonal matrix $n^{-1}\text{diag}\{\sigma_{10}^2, \sigma_{20}^2, \ldots, \sigma_{k0}^2\}$ with $\sigma_{i0}^2 := \text{Var}(\varepsilon_j(\mathbf{x}_i))$.

To implement SK for prediction, the standard practice is to first substitute $\widehat{\Sigma}_\varepsilon$ into $\Sigma = \Sigma_{\mathsf{M}} + \Sigma_\varepsilon$, with the ith diagonal entry of $\widehat{\Sigma}_\varepsilon$ specified by the simulation output sample variances for $i = 1, 2, \ldots, k$. Prediction then follows (4) and (5) upon obtaining estimates of β, θ and τ^2 through maximizing the log-likelihood function formed under the standard assumption stipulated by SK that $(\mathsf{Y}(\mathbf{x}_0), \bar{\mathscr{Y}}^\top)^\top$ follows a multivariate normal distribution (see, e.g., Ankenman et al. 2010 and Chen and Kim 2014).

3 A SEQUENTIAL DESIGN FRAMEWORK FOR STOCHASTIC KRIGING

With a given total computational budget of, say N, simulation replications to allocate, the ultimate goal of a sequential experimental design for SK is to find a design consists of not only the design-point locations but also the amount of simulation effort to be expended at each of them, i.e., $\{(\mathbf{x}_i, n_i), i = 1, 2, \ldots\}$, where n_i represents the number of simulation runs to be made at design point $\mathbf{x}_i$. To simplify our analysis, suppose that there is a relatively dense candidate design-point set, $\mathbf{X}^k$, consisting of k distinct points in $\mathscr{X}$, from which we consider selecting design points to run simulations. Notice that the value of k should not be small so that the k design points in $\mathbf{X}^k$ can cover the design space $\mathscr{X}$ adequately. The sequential design will begin with a pilot experiment. The pilot experiment is conducted at k_0 points chosen from $\mathbf{X}^k$ following a space-filling design such as maximin LHD (Morris and Mitchell 1995) with the ith point receiving n_0^i simulation replications ($k_0 \cdot \max n_0^i < N$). Denote this pilot design as $\mathscr{D}_0 = \{(\mathbf{x}_i, n_0^i), i = 1, 2, \ldots, k_0\}$. The purpose of this pilot experiment is to collect some information about the underlying response surface and to have an initial assessment of the sampling variability across $\mathscr{X}$. The resulting data can be used to fit an initial SK metamodel and an intrinsic variance metamodel across $\mathscr{X}$. Define Δn as a decision unit. In the subsequent iterations of the sequential design, each time we choose a design point (either an untried or an existing one from $\mathbf{X}^k$) according to a prespecified criterion and assign Δn replications to it. In fact, the pilot experiment determines the number of remaining iterations to go as $(N - \sum_{i=1}^{k_0} n_0^i)/\Delta n$. Call the design generated after the jth iteration $\mathscr{D}_j$. Although the total number of distinct design points selected is unpredictable, the total simulation budget allocated by the end of the jth iteration is guaranteed to be $\sum_{i=1}^{k_0} n_0^i + j \cdot \Delta n$.

Suppose that our goal of the sequential design is to minimize a performance measure $C(\cdot)$. Then for the jth iteration we set up the following myopic optimization problem seeking the optimal design point $\mathbf{x}^*$ to assign the Δn simulation runs:

$$\begin{aligned}
&\text{maximize}_{\mathbf{x}^* \in \mathbf{X}^k} \ \Delta C(\{\mathscr{D}_{j-1}, (\mathbf{x}^*, \Delta n)\}) \qquad (6)\\
&\text{subject to}\\
&\sum_{i=1}^{k_0} n_0^i + j \cdot \Delta n \le N\\
&n_0^i, \Delta n \text{ nonnegative integers}
\end{aligned}$$

where $\Delta C(\{\mathscr{D}_{j-1},(\mathbf{x}^*,\Delta n)\}) = C(\{\mathscr{D}_{j-1}\}) - C(\{\mathscr{D}_{j-1},(\mathbf{x}^*,\Delta n)\})$ is the incremental gain in terms of the chosen performance measure by assigning Δn runs to $\mathbf{x}^*$. Since upon finishing the $(j-1)$st iteration $C(\{\mathscr{D}_{j-1}\})$ becomes fixed, the optimization program (6) can be regarded as equivalent to the alternative optimization program whose objective function is changed to minimize$_{\mathbf{x}^*\in\mathbf{X}^k}$ $C(\{\mathscr{D}_{j-1},(\mathbf{x}^*,\Delta n)\})$, and the latter is subject to the same set of constraints as defined in (6). We will work with this version instead.

4 SEQUENTIAL DESIGN CRITERIA

In this section we propose several ad hoc design criteria as candidates for $C(\cdot)$ in (6) and provide the rationale for them. All these criteria are to some extent built on the integrated mean squared error (IMSE), which directly quantifies the predictive accuracy achieved by the fitted metamodel built on a given design $\mathscr{D}$, $\text{IMSE}(\mathscr{D}) := \int_{\mathscr{X}} \text{MSE}(\widehat{\mathsf{Y}}(\mathbf{x}_0);\mathscr{D})d\mathbf{x}_0$. Below we introduce the standard design criterion IMSE first and develop the other criteria as its more sophisticated extensions. In contrast to IMSE, the common feature shared by the other criteria is to create different measures to evaluate the need of allocating more simulation budget at simulated versus unsimulated points, in the hope of striking a better balance of exploration and exploitation.

- The IMSE based criterion (**IMSE**). This criterion chooses the next point from the candidate design-point set $\mathbf{X}^k$ that leads to the minimum approximated integrated mean squared error of prediction, i.e., $\mathbf{x}^* := \text{argmin}_{\mathbf{x}_i\in\mathbf{X}^k}\widehat{\text{IMSE}}(\{\mathscr{D}_{j-1},(\mathbf{x}_i,\Delta n)\})$.
- The Comparison index based criterion (**Comp**). This criterion chooses the next point that gives the minimum comparison index, $\mathbf{x}^* := \text{argmin}_{\mathbf{x}_i\in\mathbf{X}^k}\text{Comp}(\{\mathscr{D}_{j-1},(\mathbf{x}_i,\Delta n)\})$. The comparison index for each point $\mathbf{x}_i \in \mathbf{X}^k$ is defined by

$$\text{Comp}(\{\mathscr{D}_{j-1},(\mathbf{x}_i,\Delta n)\}) = \begin{cases} \frac{\widehat{\text{IMSE}}(\{\mathscr{D}_{j-1},(\mathbf{x}_i,\Delta n)\})}{\sum_{i=1}^k \widehat{\text{IMSE}}(\{\mathscr{D}_{j-1},(\mathbf{x}_i,\Delta n)\})} & \text{if } \mathbf{x}_i \notin \mathscr{D}_{j-1} \\ \\ \frac{|\widehat{\mathsf{Y}}(\mathbf{x}_i)|/\max\{\gamma|\widehat{\mathsf{Y}}(\mathbf{x}_i)|,\sqrt{\widehat{\sigma}_i^2/n_i}\}}{\sum_{i=1}^k |\widehat{\mathsf{Y}}(\mathbf{x}_i)|/\max\{\gamma|\widehat{\mathsf{Y}}(\mathbf{x}_i)|,\sqrt{\widehat{\sigma}_i^2/n_i}\}} & \text{if } \mathbf{x}_i \in \mathscr{D}_{j-1} \end{cases} \tag{7}$$

  where $\widehat{\mathsf{Y}}(\mathbf{x}_i)$ and $\widehat{\sigma}_i^2$ represent the predicted response and intrinsic variance given by respective SK metamodels at point $\mathbf{x}_i \in \mathbf{X}^k$. Notice that the value of $\text{Comp}(\cdot)$ lies in $(0,1)$. A nutshell interpretation can be given as follows. If the point is yet to be simulated (i.e., $\mathbf{x}_i \notin \mathscr{D}_{j-1}$), then we focus on evaluating its potential in bringing down the IMSE if Δn replications are assigned there. Its comparison index is calculated based on $\widehat{\text{IMSE}}(\{\mathscr{D}_{j-1},(\mathbf{x}_i,\Delta n)\})$ as given for the IMSE based criteria. If the point has already been simulated (i.e., $\mathbf{x}_i \in \mathscr{D}_{j-1}$), then this criterion focuses on the estimated impact of the sampling variability as compared to the magnitude of the response at that point. The parameter $\gamma \in (0,1)$ is a user defined quantity that reflects the relative magnitude of sampling error that the user considers tolerable. Notice that if either the estimated sampling variability $\widehat{\sigma}_i^2$ is large or the number of runs already assigned to $\mathbf{x}_i$, n_i, is small, the ratio $|\widehat{\mathsf{Y}}(\mathbf{x}_i)|/\max\{\gamma|\widehat{\mathsf{Y}}(\mathbf{x}_i)|,\sqrt{\widehat{\sigma}_i^2/n_i}\}$ tends to be small. In this case, $\mathbf{x}_i$ is more likely to be chosen as the next point among all simulated design points to run simulations. When n_i is sufficiently large, the ratio will take a constant value $1/\gamma$, hence $\mathbf{x}_i$ is considered to have been exploited sufficiently, and it is less likely to be selected. In summary, $\text{Comp}(\{\mathscr{D}_{j-1},(\mathbf{x}_i,\Delta n)\})$ is calculated as an exploration and exploitation index depending on whether the $\mathbf{x}_i$ has been simulated or not. We note that $\gamma = 0.005$ is used for all the numerical examples presented in Section 5.
- The modified IMSE criteria (**MIMSE**). These criteria emerge by combining the strengths of the aforementioned two criteria, **IMSE** and **Comp**. Three versions are considered here, namely, **MIMSE-1, MIMSE-2**, and **MIMSE-3**. All of them reweight the approximated IMSEs calculated at the simulated design points to balance the resource allocation among the simulated design points

and those unsimulated ones. Specifically, at each potential design point $\mathbf{x}_i \in \mathbf{X}^k$ define its MIMSE index as $\text{MIMSE}(\{\mathscr{D}_{j-1}, (\mathbf{x}_i, \Delta n)\}) = f_i \times \widehat{\text{IMSE}}(\{\mathscr{D}_{j-1}, (\mathbf{x}_i, \Delta n)\})$. Then the next point to run simulations is chosen as $\mathbf{x}^* := \text{argmin}_{\mathbf{x}_i \in \mathbf{X}^k} \text{MIMSE}(\{\mathscr{D}_{j-1}, (\mathbf{x}_i, \Delta n)\})$. Here f_i specifies the weight assigned to $\mathbf{x}_i$: If the point is yet to be simulated (i.e., $\mathbf{x}_i \notin \mathscr{D}_{j-1}$), $f_i = 1$. Otherwise, f_i is defined as a monotonically decreasing function of $u(\mathbf{x}_i, n_i) - \gamma$, where $u(\mathbf{x}_i; n_i)$ represents a measure of remaining uncertainty at $\mathbf{x}_i$ given that n_i simulation runs having already been assigned there; the pre-specified parameter $\gamma \in (0, 1)$ reflects the amount of uncertainty at a simulated design point that the user considers tolerable. If $u(\mathbf{x}_i; n_i) - \gamma \gg 0$, then f_i takes a value much smaller than 1. In this case, the MIMSE index calculated will be depreciated significantly by the weight f_i from $\widehat{\text{IMSE}}(\{\mathscr{D}_{j-1}, (\mathbf{x}_i, \Delta n)\})$ such that a greater chance is presented for $\mathbf{x}_i$ to be the point chosen next. On the other hand, if $u(\mathbf{x}_i; n_i) - \gamma < 0$, then f_i tends to take a value greater than 1 so that the MIMSE index calculated will be inflated by f_i as compared to $\widehat{\text{IMSE}}(\{\mathscr{D}_{j-1}, (\mathbf{x}_i, \Delta n)\})$ to prevent $\mathbf{x}_i$ from getting even more runs. When the amount of uncertainty at $\mathbf{x}_i$ hits the threshold level, i.e., $u(\mathbf{x}_i; n_i) - \gamma = 0$, the value of f_i is set to a value slightly greater than 1 to penalize the simulated points slightly from getting more simulation budget. Mathematically speaking, the weight f_i for the modified IMSE criteria is given by

$$f_i = \begin{cases} 1 & \text{if } \mathbf{x}_i \notin \mathscr{D}_{j-1} \ , \\ p \cdot [1 + \eta\,(u(\mathbf{x}_i; n_i) - \gamma)]^{-1} & \text{if } \mathbf{x}_i \in \mathscr{D}_{j-1} \end{cases} \tag{8}$$

where $p > 1$ is a constant slightly greater than 1 which specifies the penalty given to $\mathbf{x}_i$ when $u(\mathbf{x}_i; n_i) = \gamma$; the value of η is set as $\eta = (p-1)/(4\gamma)$ such that $f_i = 1$ (hence no penalization is given to $\mathbf{x}_i$) when the uncertainty measure $u(\mathbf{x}_i; n_i) = 5\gamma$. We note that $p = 1.1$ and $\gamma = 0.01$ are used for all the numerical examples presented in Section 5. **MIMSE-1, MIMSE-2**, and **MIMSE-3** are different in their ways of defining the uncertainty measure, $u(\mathbf{x}_i; n_i)$ at $\mathbf{x}_i \in \mathscr{D}_{j-1}$. Specifically,

MIMSE-1: $u(\mathbf{x}_i; n_i) = \sqrt{\widehat{\sigma}^2/n_i}/|\widehat{\mathsf{Y}}(\mathbf{x}_i)|$; the uncertainty measure is specified as the estimated relative error at $\mathbf{x}_i \in \mathscr{D}_{j-1}$.

MIMSE-2: $u(\mathbf{x}_i; n_i) = \sqrt{\widehat{\sigma}^2/n_i}/\overline{|\widehat{\mathsf{Y}}|}$; the uncertainty measure is defined in a similar fashion as that for **MIMSE-1**, but $|\widehat{\mathsf{Y}}(\mathbf{x}_i)|$ is replaced by $\overline{|\widehat{\mathsf{Y}}|}$, the average magnitude of the predicted responses at all the simulated design points in $\mathscr{D}_{j-1}$.

MIMSE-3: $u(\mathbf{x}_i; n_i) = |\bar{\mathscr{Y}}(\mathbf{x}_i) - \widehat{\mathsf{Y}}(\mathbf{x}_i)|/|\widehat{\mathsf{Y}}(\mathbf{x}_i)|$, where $\bar{\mathscr{Y}}(\mathbf{x}_i)$ represents the average simulated response at $\mathbf{x}_i$. In this case, the uncertainty measure is defined as the relative discrepancy between the averaged simulation response and the predicted value at $\mathbf{x}_i$.

5 NUMERICAL EXAMPLES

In this section, we compare the predictive performances of SK with the design criteria proposed in Section 4 through four illustrative examples. The first two are stylized one-dimensional problems that show different features of each design criterion under consideration. The last two examples, namely, an M/M/1 queue and a simple (s, S) inventory system, further demonstrate the advantages of applying SK with sequential design schemes over commonly used non-sequential designs.

5.1 Example 1

Consider the following one-dimensional problem in which we try to approximate a simple non-increasing function $\mathsf{Y}(x)$ given by

$$\mathsf{Y}(x) = 2 + 3/x, \ x \in \mathscr{X} = [0.5, 7] \ . \tag{9}$$

Specifically, the simulation output at design point x on the jth replication is generated according to

$$\mathscr{Y}_j(x) = \mathsf{Y}(x) + \varepsilon_j(x) \ . \tag{10}$$

That is, the simulation error $\varepsilon_j(x)$ on the jth replication is sampled from a normal distribution with mean zero and standard deviation as a function of x, $g(x)$, i.e., $\varepsilon_j(x) \sim \mathscr{N}(0, [g(x)]^2)$. Hence the sampling errors are heterogeneous across the design space $\mathscr{X}$. The following two scenarios of sampling variability are taken into account:

E.g. (1.1) $g(x) = x^{-1.5}$, i.e., the intrinsic variance decreases as x increases;
E.g. (1.2) $g(x) = x/3$, i.e., the intrinsic variance increases as x increases.

Experimental setup. Given a total budget of N simulation replications, we start with a pilot design $\mathscr{D}_0$ consisting of 4 equally spaced design points with 20 simulation runs assigned to each one of them. In each subsequent iteration $\Delta n = 20$ runs are allocated to a design point selected from the set of potential design points, $\mathbf{X}^k$, which consists of $k = 193$ equally spaced points in $\mathscr{X}$. A check-point set $\mathbf{X}^K$ with K (typically $K \geq k$) points densely located in $\mathscr{X}$ is constructed to evaluate the predictive performances of SK. Notice that typically $\mathbf{X}^k \subseteq \mathbf{X}^K$; here for simplicity we set them equal to each other. We use a total simulation budget of $N = 500$ and 5000 runs to investigate the finite-sample and large-sample performances of SK using different design criteria. Since a grid design of sufficient number of design points are typically adequate for one-dimensional problems, we also consider a grid design of equally spaced 25 (resp. 193) design points in $\mathscr{X}$ when $N = 500$ (resp. $N = 5000$) and use its corresponding predictive performance of SK as a benchmark. The two grid designs used when $N = 500$ and 5000 are denoted by **Grid25** and **Grid193**, respectively. The entire experiment is repeated for 100 independent macro-replications, and the corresponding performance measure, the empirical root mean squared errors (ERMSE), is calculated as follows,

$$\text{ERMSE}_\ell = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(\widehat{\mathsf{Y}}(x_i) - \mathsf{Y}(x_i)\right)^2}, \quad \ell = 1, 2, \ldots, 100, \tag{11}$$

where $\widehat{\mathsf{Y}}(\cdot)$ represents the prediction given by SK in a given macro-replication.

Results. Table 1 summarizes the ERMSEs obtained using different designs with a total budget of $N = 500$ and 5000 simulation replications. We observe that all the sequential designs lead to better predictive performances as compared to the corresponding grid designs. For E.g. (1.1), it is interesting to see that among all sequential design criteria **IMSE** is the least efficient in terms of the ERMSEs achieved at the end of sequential iterations. Nevertheless, Figures 1 (a) and (b) show that **IMSE** tends to lead to relatively faster decreasing ERMSEs than those other criteria do in the early iterations. This finite-sample property is especially desirable when the total budget N is not large. In comparison, some of the better performing criteria such as **Comp**, **MIMSE-1** and **MIMSE-2** lead to slowly decreasing ERMSEs initially but tend to produce more aggressive reduction in ERMSEs as the iterations proceed even further.

Unlike E.g. (1.1), E.g. (1.2) is more difficult to handle as the intrinsic variance is increasing while the true response is decreasing as x increases. The ratio of the sampling variability relative to the mean response becomes particularly large when x is large. For instance, the true function $\mathsf{Y}(x)$ at $x = 7$ is about 2.43 but the sampling error at that point has a standard deviation of about 2.33. In this case, the sequential designs try to reconcile the competing interests of reducing spatial uncertainty in the region where x is small versus diminishing the intrinsic variability at simulated design points with large x values. On the other hand, **Grid25** with evenly allocated simulation budget seem to help SK achieve quite robust predictive performance for E.g. (1.2) with a moderate sample size. As we observe from Table 1, the predictive performances of SK using the sequential design criteria are very close to those corresponding to **Grid25** with a budget of $N = 500$. However, as the total budget grows even further to $N = 5000$, the advantage of using sequential designs starts to emerge.

Table 1: A summary of the ERMSEs for Example 1 obtained by different criteria with a total budget of $N = 500$ and 5000 simulation replications.

N		E.g. (1.1)				E.g. (1.2)			
	Percentiles	25th	50th	75th	97.5th	25th	50th	75th	97.5th
500	**IMSE**	0.106	0.140	0.184	0.316	0.180	0.208	0.244	0.304
	Comp	0.074	0.083	0.103	0.142	0.177	0.200	0.233	0.328
	MIMSE-1	0.092	0.110	0.129	0.186	0.193	0.229	0.266	0.343
	MIMSE-2	0.076	0.087	0.103	0.136	0.166	0.190	0.224	0.340
	MIMSE-3	0.076	0.087	0.098	0.132	0.194	0.221	0.252	0.339
	Grid25	0.162	0.221	0.260	0.359	0.168	0.198	0.252	0.346
5000	**IMSE**	0.031	0.038	0.053	0.078	0.055	0.060	0.072	0.094
	Comp	0.030	0.036	0.041	0.052	0.056	0.064	0.072	0.090
	MIMSE-1	0.039	0.048	0.054	0.071	0.055	0.063	0.071	0.094
	MIMSE-2	0.033	0.038	0.044	0.056	0.055	0.061	0.068	0.088
	MIMSE-3	0.032	0.037	0.046	0.060	0.055	0.065	0.074	0.093
	Grid193	0.074	0.095	0.109	0.145	0.059	0.073	0.089	0.125

5.2 Example 2

Following Subsection 5.1, we consider using SK to approximate a function that is not monotonic in its domain. Specifically,

$$\mathsf{Y}(x) = 10 + x \cdot \sin(3x), \quad x \in [0.5, 7]\ . \tag{12}$$

The simulation output at design point x on the jth replication is generated according to (10), where the simulation error $\varepsilon_j(x) \sim \mathcal{N}(0, [g(x)]^2)$. We consider the following three sampling variability scenarios:

E.g. (2.1) $g(x) = x^{1/2}$, i.e., the intrinsic variance increases as x increases;

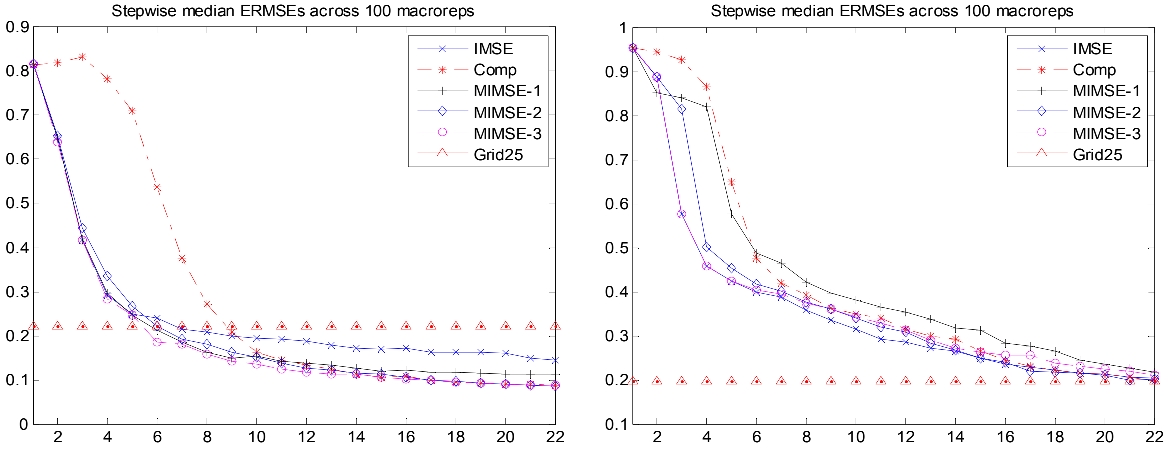
E.g. (2.2) $g(x) = |x \cdot \sin(3x)|^{1/2}$, i.e., the intrinsic variance changes together with the true function value, and its magnitude depends on the deviation of the true response $\mathsf{Y}(x)$ from the constant 10;

E.g. (2.3) $g(x) = x^{-1/2}$, i.e., the intrinsic variance decreases as x increases.

Table 2: A summary of the ERMSEs for Example 2 obtained using different designs with a total budget of $N = 500$ and 5000 simulation replications.

N		E.g. (2.1)				E.g. (2.2)				E.g. (2.3)			
	Percentiles	25th	50th	75th	97.5th	25th	50th	75th	97.5th	25th	50th	75th	97.5th
500	**IMSE**	0.259	0.307	0.364	0.440	0.202	0.233	0.279	0.365	0.090	0.105	0.118	0.156
	Comp	0.263	0.301	0.346	0.434	0.212	0.244	0.285	0.356	0.094	0.105	0.119	0.148
	MIMSE-1	0.270	0.311	0.369	0.458	0.203	0.248	0.275	0.369	0.095	0.106	0.120	0.156
	MIMSE-2	0.261	0.300	0.352	0.452	0.202	0.234	0.265	0.366	0.094	0.105	0.119	0.148
	MIMSE-3	0.269	0.317	0.362	0.438	0.200	0.237	0.274	0.357	0.094	0.105	0.119	0.148
	Grid25	0.257	0.302	0.350	0.436	0.199	0.234	0.283	0.361	0.091	0.102	0.120	0.151
5000	**IMSE**	0.087	0.099	0.111	0.135	0.061	0.070	0.082	0.112	0.029	0.033	0.036	0.045
	Comp	0.089	0.105	0.120	0.149	0.060	0.072	0.083	0.106	0.029	0.032	0.037	0.046
	MIMSE-1	0.092	0.109	0.128	0.155	0.071	0.080	0.093	0.124	0.028	0.033	0.038	0.047
	MIMSE-2	0.088	0.098	0.114	0.146	0.059	0.068	0.079	0.109	0.028	0.032	0.038	0.047
	MIMSE-3	0.092	0.101	0.124	0.156	0.068	0.075	0.088	0.116	0.029	0.033	0.037	0.056
	Grid193	0.093	0.109	0.128	0.161	0.062	0.075	0.094	0.128	0.030	0.036	0.043	0.055

Results. The experimental setup is as described in Subsection 5.1. The ERMSEs obtained using different criteria with a total budget of $N = 500$ and 5000 simulation replications are given in Table 2. Since the



(a) E.g. (1.1)

(b) E.g. (1.2)

Figure 1: Stepwise median ERMSEs across 100 macro-replications obtained by different designs with a total budget of $N = 500$ runs. The sequential designs all start with a pilot design $\mathscr{D}_0 = \{(0.5, 20), (2.67, 20), (4.83, 20), (7, 20)\}$ and adopt the step size $\Delta n = 20$ runs.

modeling difficulty decreases from E.g. (2.1) to E.g. (2.3), the corresponding ERMSEs obtained using all criteria are observed to decrease correspondingly, as expected. Performances due to all the designs are very close to each other for E.g. (2.1) to (2.3), including the grid designs. Table 3 shows the number of design points used by each criterion given a fixed budget N. We observe that for all three scenarios of Example 2, all sequential design criteria end up choosing to simulate at relatively large numbers of design points scattered in $\mathscr{X}$; this behavior is very similar to that of a grid design, which helps explain the resemblance in the predictive performances of SK using different designs.

Table 3: Example 2. A summary of the total numbers of design points used by different sequential criteria under the proposed sequential framework given a total budget of $N = 500$ and 5000 simulation runs.

		E.g. (2.1)			E.g. (2.2)			E.g. (2.3)		
		min	median	max	min	median	max	min	median	max
N 500	**IMSE**	23	25	25	22	25	25	22	24	25
	Comp	24	25	25	16	21	25	24	25	25
	MIMSE-1	17	20	23	20	20	22	25	25	25
	MIMSE-2	19	23	25	22	24	25	25	25	25
	MIMSE-3	18	23	25	20	23	25	24	25	25
	Grid25	25								
5000	**IMSE**	131	141	152	118	139	154	122	133	149
	Comp	143	147	151	114	128	147	129	133	136
	MIMSE-1	170	176	182	180	187	189	193	193	193
	MIMSE-2	182	193	193	193	193	193	193	193	193
	MIMSE-3	147	169	191	168	185	193	193	193	193
	Grid193	193								

5.3 Example 3: M/M/1 Queue

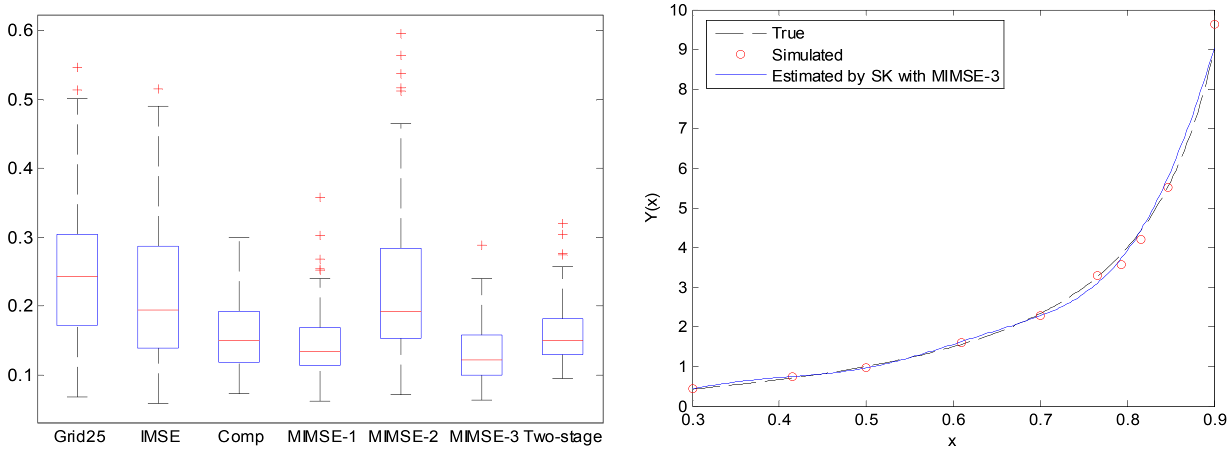
In this subsection, we are interested in estimating the steady-state mean number in an M/M/1 queue with service rate 1 and arrival rate x varying in $\mathscr{X} = [0.3, 0.9]$ via SK metamodeling. It is known that the underlying true response surface to model is $\mathsf{Y}(x) = x/(1-x)$. The goal is to efficiently allocate a total budget $N = 500$ replications to a set of design points for minimizing IMSE over $\mathscr{X}$. This example has been considered by Ankenman et al. (2010) who propose to use a two-stage design. Their design starts with simulating 20 replications of length $T = 1000$ time units at each of the initial set of 4 design points $\{0.3, 0.5, 0.7, 0.9\}$, and in the second stage it allocates the remaining budget to a set of 7 preselected design points $\{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. The second stage allocation is made according to the approximately optimal allocation rule derived from minimizing the approximated IMSE based on the initial stage design. The two-stage design is intended for bringing down the IMSE and hence it performs well. However, we note that calculating their two-stage budget allocation rule is a nontrivial exercise even for 7 pre-selected design points. Thanks to the proposed sequential design framework, we can automate the budget allocation under the proposed framework by using the criteria considered. We choose to set the step size to $\Delta n = 20$ runs. To make a fair comparison, the pilot design is made the same as the initial stage of the two-stage design given by Ankenman et al. (2010). For convenience, the design space $\mathscr{X}$ is discretized into a dense grid of $k = 193$ equally spaced points $\mathbf{X}^k = \{x_i\}_{i=1}^{193}$, which is used as both the candidate design-point set and the check-point set. A grid design of equally spaced 25 design points in $\mathscr{X} = [0.3, 0.9]$ is also considered to provide the benchmark performance. We repeat the entire experiment for 100 macro-replications, and obtain 100 ERMSEs as defined in (11) for each design.

Results. The resulting ERMSEs due to different criteria are shown in boxplots in Figure 2(a). As we can see from Figure 2(a), **MIMSE-1** and **MIMSE-3** achieve the best predictive accuracy, outperforming the two-stage allocation scheme proposed by Ankenman et al. (2010). As is known that the variance explodes in the "heavy traffic" region, so intuitively many replications are then needed to achieve good predictive performance. In particular, we found that **MIMSE-1** and **MIMSE-3** choose to use a moderate number of design points (between 10 to 15) and dedicated sufficient runs to the region of x that corresponds to the M/M/1 queue in "heavy traffic," whereas in comparison the two-stage design scheme tends to undersample that region. For the purpose of clarity, this feature is illustrated by the results obtained using **MIMSE-3** only in Figure 2(b). It is worth noting that despite the fact that grid designs seem rather efficient in Example 2, here **Grid25** leads to the worst performance among all designs. Similar to **Grid25**, we note that **IMSE** leads to uncompetitive performance; this is largely due to its resulting allocation of simulation runs to a relatively scattered set of design points, leaving the "heavy traffic" region undersampled.

5.4 Example 4: A Periodic Review (s, S) Inventory System

In this section, we consider a two-dimensional problem, which is based on a simple periodic review single-commodity (s, S) inventory system that supplies external demands and receives stock from a production facility. The system is assumed to have i.i.d. continuous demands, zero lead times, full backlogging, and linear ordering, holding and shortage costs. The scenario considered here is similar to that discussed in Fu and Healy (1997), upon which much of this example is constructed. Let X_i be the inventory position (inventory level plus outstanding orders in period i) and W_i be the inventory level (on-hand minus on backorder). The assumption of zero lead times gives $X_i = W_i$. The (s, S) inventory system works as follows. If X_i is below s units, an order of amount $(S - X_i)$ will be made and a fixed ordering cost K_o and a purchase cost $c(S - X_i)$ will be incurred. The inventory holding cost and shortage cost are also taken into account. The demand in period i, D_i, has distribution function $F_D(\cdot)$, which is absolutely continuous with density function $f_D(\cdot)$; denote the mean demand by $\mathrm{E}[D]$. The one-period cost is the sum of ordering, holding and backorder costs as follows

$$J_i = \underline{1}\{X_i < s\}(K_o + c(S - X_i)) + h\max\{0, W_i\} + p\max\{0, -W_i\},$$



(a)

(b)

Figure 2: Estimating M/M/1 queue steady-state mean number in the system with a total budget of $N = 500$ replications: (a) provides a summary of 100 ERMSEs achieved using different designs; (b) shows the true response surface of the steady-state mean number in an M/M/1 queue $\mathsf{Y}(x)$, the simulated responses at design points chosen by **MIMSE-3**, and the resulting fitted response surface by SK.

Table 4: Parameters for the (s,S) inventory system

$\mathrm{E}[D]$	c	p	K_o	h
20	0.05	0.5	5	0.05

and the quantity of interest is the long-run average cost per period

$$J = \lim_{n\to\infty} J_n, \text{ and } J_n = \frac{1}{n}\sum_{i=1}^{n} J_i.$$

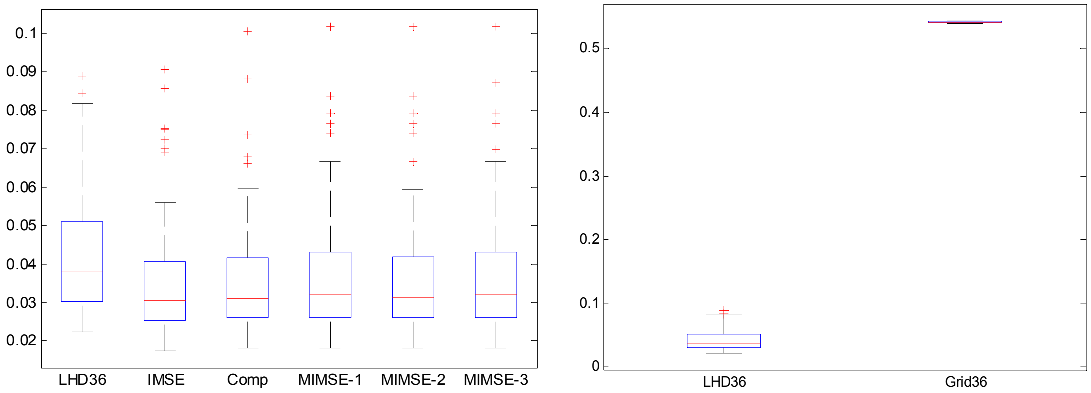
Define $\delta = S - s$ and let $\lambda = 1/\mathrm{E}[D]$. Consider the two-dimensional problem of estimating the unknown response $J(\delta, s)$ at a given point (δ, s). If the demands are i.i.d. exponentially distributed with mean $\mathrm{E}[D]$, then the analytic expression for $J(\delta, s)$ can be given as follows

$$J(\delta,s) = c\mathrm{E}[D] + \frac{K_o + h(s - \mathrm{E}[D] + \lambda\delta(s+\delta/2)) + (h+p)\mathrm{E}[D]e^{-\lambda s}}{1+\lambda\delta}.$$

The list of parameters involved are given in Table 4. The goal of this experiment is to compare the performances of SK in predicting $J(\delta, s)$ over the design space $\mathscr{X} = \Omega_\delta \times \Omega_s = [10,40] \times [10,50]$, by adopting different designs given a total budget of $N = 1800$ simulation replications to expend. In each simulation replication the run length of $T = 1000$ is used to estimate the response $J(\delta, s)$ at a given design point.

We are interested in comparing the sequential design criteria with commonly used grid and Latin hypercube sampling designs (LHD). A two-layer nested LHD of $k = 36$ points (see Qian (2009)) is constructed as the full candidate design set $\mathbf{X}^k$: The first layer D_1 is a LHD with 12 points and the full set D_2 is a LHD with 36 points; and the name of the design follows from the fact that $D_1 \subset D_2 = \mathbf{X}^k$. In each iteration we consider selecting the next point from D_2 according to a given sequential design criterion to allocate $\Delta n = 50$ simulation runs. The initial design is conveniently chosen as D_1, a LHD of 12 points itself, with $n_0 = 50$ simulation runs allocated at each point. For comparison purposes, we also consider using a grid design and a LHD of 36 points, denoted by **Grid36** and **LHD36**, respectively. **Grid36** assigns 1800 runs evenly to 36 regularly spaced points in $\mathscr{X} = [10,40] \times [10,50]$, while **LHD36** assigns 1800 runs evenly to the 36 points in D_2. In contrast, although our proposed sequential design criteria choose points from the full set D_2, the resulting design formed by the end of a sequential procedure is very likely to be an unequal allocation of budget on a subset of points in D_2. To evaluate predictive accuracy of SK, we construct the check-point set $\mathbf{X}^K$ consisting of $K = 2536$ check-points, i.e., a grid design of regularly spaced 2500 points in $\mathscr{X}$ plus the 36 design points in D_2 for sequential designs and **LHD36** or those 36 grid points in **Grid36**.

The experiment is repeated for 100 independent macro-replications and the resulting ERMSEs are shown in Figures 3(a) and (b). Notice that for each macro-replication a distinct two-layer nested LHD is generated; hence we are using 100 independently generated pairs of initial and full design-point sets (D_1, D_2) in the 100 macro-replications. From Figure 3(a), we see that all five sequential criteria perform similarly; in particular, their corresponding ERMSEs are smaller than those due to **LHD36**. This demonstrates the advantage of applying SK with a sequential design as opposed to with a non-sequential one. Furthermore, a striking difference has been identified from Figure 3(b) between the two non-sequential designs: **LHD36** and **Grid36**: With the same number of design points and amount of simulation budget to use, the SK metamodel built on **LHD36** achieved much better predictive accuracy as compared to the one built on **Grid36**. This example to some extent reveals the known problem of grid designs in their diminishing efficiency as the dimension of the problem grows.



(a)

(b)

Figure 3: Estimating (s,S) Inventory System long-run average cost with a total budget of $N = 1800$ simulation replications to expend: (a) provides a summary of respective 100 ERMSEs achieved by using different sequential design criteria as compared to those obtained by 36-point LHD ; (b) compares the resulting 100 ERMSEs obtained by using 100 different 36-point LHDs (i.e., 100 D_2's) and those resulted from a fixed 36-point grid design.

6 CONCLUSION

In this paper we have established a general sequential design framework and proposed several design criteria to apply SK for stochastic simulation. The advantages of using the proposed sequential designs over the commonly used non-sequential ones have been demonstrated through some illustrative examples. In particular, these advantages are observed to become more evident as the problem dimension increases and/or as the simulation budget increases. Future research topics include exploring the framework structure to establish in-depth theoretical treatments, and constructing candidate design-point set and devising sequential design criteria for general higher-dimensional problems, to name a few.

ACKNOWLEDGMENTS

Dr. Qiang Zhou's work was supported by the National Natural Science Foundation of China under grant No. 11301441.

REFERENCES

Ankenman, B. E., B. L. Nelson, and J. Staum. 2010. "Stochastic Kriging for Simulation Metamodeling". *Operations Research* 58:371–382.

Chen, X., B. E. Ankenman, and B. L. Nelson. 2012. "The Effects of Common Random Numbers on Stochastic Kriging Metamodels". *ACM Transactions on Modeling and Computer Simulation* 22:7/1–7/20.

Chen, X., and K.-K. Kim. 2014. "Stochastic Kriging with Biased Sample Estimates". *ACM Transactions on Modeling and Computer Simulation* 24:8/1–8/23.

Fang, K. T., D. K. J. Lin, P. Winker, and Y. Zhang. 2000. "Uniform Design: Theory and Application". *Technometrics* 42:237–248.

Fu, M. C., and K. J. Healy. 1997. "Techniques for Simulation Optimization: An Experimental Study on an (s,S) Inventory System". *IIE Transactions* 29:191–199.

Kleijnen, J. P. C. 2008. *Design and Analysis of Simulation Experiments*. New York: Springer.

McKay, M. D., R. J. Beckman, and W. J. Conover. 1979. "Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". *Technometrics* 21:239–245.

Morris, M., and T. Mitchell. 1995. "Exploratory Designs for Computer Experiments". *Journal of Statistical Planning and Inference* 43:381–402.

Ng, S. H., and J. Yin. 2012. "Bayesian Kriging Analysis and Design for Stochastic Simulations". *ACM Transactions on Modeling and Computer Simulation* 22:1–26.

Qian, P. Z. G. 2009. "Nested Latin Hypercube Designs". *Biometrika* 96:957–970.

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

Shewry, M. C., and H. P. Wynn. 1987. "Maximum Entropy Sampling". *Journal of Applied Statistics* 14:165–170.

Tang, B. 1993. "Orthogonal Array-Based Latin Hypercubes". *Journal of the American Statistical Association* 88:1392–1397.

van Beers, W. C. M., and J. P. C. Kleijnen. 2008. "Customized Sequential Designs for Random Simulation Experiments: Kriging Metamodeling and Bootstrapping". *European Journal of Operational Research* 186:1099–1113.

AUTHOR BIOGRAPHIES

XI CHEN is an Assistant Professor in the Grado Department of Industrial and Systems Engineering at Virginia Tech. Her research interests include stochastic modeling and simulation, applied probability and statistics, computer experiment design and analysis, and simulation optimization. Her email address is xchen6@vt.edu.

QIANG ZHOU is an Assistant Professor in the Department of Systems Engineering and Engineering Management at City University of Hong Kong. His research focuses on statistical modeling, monitoring and analysis of complex processes/systems for the purpose of quality control, improvement of productivity and operational performance. His email address is q.zhou@cityu.edu.hk.