informs
http://pubsonline.informs.org/journal/ijoc/

INFORMS JOURNAL ON COMPUTING
Vol. 00, No. 0, Xxxxx 2017, pp. 1–17
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

Some Monotonicity Results for Stochastic Kriging Metamodels in Sequential Settings

A1 Bing W Jiaqiao Hu[a]

[a] Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, New York 11794

Contact: xzszwb@gmail.com (BW); jqhu@ams.sunysb.edu, http://orcid.org/0000-0002-9999-672X (JH)

Received: September 30, 2015
Revised: August 22, 2016; May 15, 2017; August 13, 2017
Accepted: August 15, 2017
Published Online:

https://doi.org/10.1287/ijoc.2017.0779

Abstract. Stochastic kriging (SK) and stochastic kriging with gradient estimators (SKG) are useful methods for effectively approximating the response surface of a simulation model. In this paper, we show that in a fully sequential setting when all model parameters are known, the mean squared errors of the optimal SK and SKG predictors are monotonically decreasing as the number of design points increases. In addition, we prove, under appropriate conditions, that the use of gradient information in the SKG framework generally improves the prediction performance of SK. Motivated by these findings, we propose a sequential procedure for adaptively choosing design points and simulation replications in obtaining SK (SKG) predictors with desired levels of fidelity. We justify the validity of the procedure and carry out numerical experiments to illustrate its performance.

History: Accepted by Marvin Nakayama, (former) Area Editor for Simulation.
Funding: This work was supported by the National Science Foundation [Grant CMMI-1634627].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/ijoc.2017.0779.

Keywords: A2 simulation • stochastic kriging • sequential sampling

1. Introduction

A3 Complex systems arising in supply chain management, financial engineering, and telecommunications frequently require simulation models for performance estimation. Correct interpretation of simulation outputs relies on statistical methods, which often involve a large number of simulation runs at each system configuration (engineering design). When the cost of simulation is high and the number of design alternatives is large, it is often desirable to compromise the accuracy of a simulation model in favor of good approximations by using metamodels (i.e., models of simulation models) to obtain an explicit approximation of simulation input-output relations. A variety of metamodels have been studied in the literature, ranging from simple low-order polynomial regressions to sophisticated models based on neural networks and radial-basis function approximations; see, e.g., Barton and Meckesheimer (2006), Barton (2009), Kleijnen (2007) and references therein for a review of different metamodeling techniques. Well-built metamodels consume less resources than those required by direct simulation and can be very helpful in expediting the analysis and decision making process (e.g., Wang 2005, Wang and Shan 2006, Yang et al. 2011, Chang et al. 2013, Xie et al. 2014).

Kriging is an interpolation-based technique widely used in developing metamodels. It uses the realization of a Gaussian random field to model the response surface of an unknown function and has been especially prominent in the design and analysis of deterministic simulation experiments; see, e.g., Sacks et al. (1989), Santner et al. (2003), Wang and Shan (2006), Kleijnen (2007, 2009), Viana et al. (2014), Ulaganathan et al. (2014). In Ankenman et al. (2010), kriging has been extended to the stochastic setting by explicitly taking into account the sampling noise inherent in random simulation. This gives rise to a novel framework called stochastic kriging (SK) for obtaining metamodels capable of capturing both intrinsic and extrinsic uncertainty in the design. More recently, it has been shown in Chen et al. (2013a), Qu and Fu (2014) that when additional gradient information is available, the accuracy of metamodels in SK may be further enhanced by incorporating gradient estimates, leading to significant improvement in surface prediction.

An important issue in the application of SK and stochastic kriging with gradient estimators (SKG) (Chen et al. 2013a) is the selection of design points in obtaining high-quality metamodels. In practice, this can be carried out either in a one-shot space-filling way or through a sequential sampling strategy. The latter allows the metamodel to be constructed sequentially as data accumulate so that new design points can be selected in an adaptive manner based on the updated metamodel. The benefits of sequential sampling have been shown in Jin et al. (2002) and Sacks et al. (1989), under the deterministic simulation setting. In stochastic simulation, however, response values are corrupted by sampling noise. Thus, a natural question is whether the added information implied by a new design point

will actually increase the overall quality of a metamodel. It is tempting to think that if a large amount of simulation effort is expended at each design, then a sequential strategy carefully designed in deterministic simulation should work equally well in the stochastic setting. But the key issue to address is how to effectively allocate simulation replications in such a sequential strategy.

In this paper, we investigate the performance of SK and SKG metamodels in a fully sequential setting, where design points are selected one at a time. One of our main contributions is to show that when all model parameters are known (or predetermined), the mean squared error (MSE) of an SK (SKG) predictor is monotonically decreasing as the number of design points increases. Thus, an interesting finding of our study is that the prediction performance of both SK and SKG models can always be improved by including additional design points, regardless of how these points and simulation replications are allocated. Note that since SK is an extension of kriging in the stochastic simulation setting, the same monotonicity results hold also for deterministic kriging models and, consequently, may have immediate applications in analyzing the convergence properties of existing sequential strategies based on, e.g., optimizing MSE or integrated MSE (IMSE).

In addition to the above monotonicity results, our analysis in the SKG case also provides important insight into the performance gained from using gradient information in the SKG framework. In particular, by comparing the MSEs of the respective SK and SKG predictors based on the same set of design points, we derive closed-form expressions for the reduction in MSE of SKG over SK under appropriate conditions on the simulation noises and covariance model. These lead us to conclude that, compared to SK, the use of gradient estimators in the SKG framework will in general have a positive impact on prediction performance, resulting in better predictors with smaller MSEs. To the best of our knowledge, this result has only been reported in Chen et al. (2013a) under a simplified setting, assuming that gradient information is available in one coordinate direction and spatial correlations between distinct design points are negligible.

Based on the monotonicity properties of SK and SKG, we further propose a design strategy called adaptive sequential kriging (ASK) for obtaining an SK (SKG) metamodel with a prescribed level of accuracy. The underlying idea is to select the new design point that achieves the maximum reduction in IMSE at each iteration. To address the simulation allocation issue, we make the key observation that the quality of an SK (SKG) predictor at a sampled design point is governed by the variance of the averaged intrinsic noise at the point. This suggests a simple rule for allocating simulation replications, which ensures the overall predictor IMSE to fall below a given threshold as the number of design points increases. Note that our proposed procedure for selecting design points resembles the sequential IMSE approach discussed in Jin et al. (2002) and Sacks et al. (1989). Thus, by addressing the additional simulation allocation issue, ASK can essentially be viewed as an extension of the IMSE approach in the stochastic simulation setting. Also relevant to our work is the adaptive exploration-exploitation search algorithm (AEES) introduced in Ajdari and Mahlooji (2014). AEES is also sequential in nature and allocates the simulation budget to design points based on their estimated intrinsic variances. However, the method uses a heuristic procedure to balance between exploration and exploitation in searching for new design points, which is different from ours. We justify the validity of ASK under the case when all model parameters are known. Specifically, by exploiting the monotonicity of SK and SKG, we show that if design points and simulation replications are allocated according to ASK, then the resulting predictor will reach a desired IMSE level in a finite number of iterations. Our preliminary experimental results indicate that our procedure is promising and may significantly outperform some of the existing approaches based on exploration-exploitation designs and maximizing MSE.

The rest of this paper is organized as follows. Section 2 introduces the notation used in the paper and outlines the basic mathematical frameworks of SK and SKG. In Section 3, we establish the monotonicity properties of SK and SKG predictors and show that SKG reduces MSE compared to SK. Based on these results, we introduce the proposed ASK algorithm and justify its validity in Section 4. We present computational comparison results in Section 5 and conclude the paper in Section 6.

2. Stochastic Kriging With and Without Gradient Estimators

Consider the problem of describing the response surface of an unknown function $f(\mathbf{x})$, $\mathbf{x} \in \mathscr{X}$, where $\mathbf{x}$ is a vector of design variables and $\mathscr{X}$ is a compact full-dimensional subset of $\mathbb{R}^d$. At each point $\mathbf{x}$, we assume that the true function value $f(\mathbf{x})$ cannot be evaluated exactly but can be estimated in a path-wise manner through stochastic simulation. For computational tractability, it is also assumed that $\mathscr{X}$ is characterized by simple constraints, e.g., box constraints with known bounding coordinates, so that we have complete knowledge of the underlying design space.

Given a set of design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, after we replicate n_i simulations at each point $\mathbf{x}_i$, $i = 1, \ldots, k$, the performance measures at these k design points can

be estimated by the vector $\bar{\mathbf{y}} = (\bar{y}(\mathbf{x}_1), \bar{y}(\mathbf{x}_2), \ldots, \bar{y}(\mathbf{x}_k))^\intercal$, where $\bar{y}(\mathbf{x}_i) = (1/n_i)\sum_{j=1}^{n_i} y_j(\mathbf{x}_i)$ and $y_j(\mathbf{x}_i)$ is the simulation output at $\mathbf{x}_i$ obtained on the jth replication run. In stochastic kriging, $y_j(\mathbf{x}_i)$ is assumed to take the following form:

$$\begin{aligned} y_j(\mathbf{x}_i) &= \mathbf{f}(\mathbf{x}_i)^\intercal \boldsymbol{\beta} + M(\mathbf{x}_i) + \epsilon_j(\mathbf{x}_i) \\ &= Y(\mathbf{x}_i) + \epsilon_j(\mathbf{x}_i), \end{aligned} \tag{1}$$

where $\mathbf{f}$: $\mathbb{R}^d \to \mathbb{R}^p$ is a vector of user specified basis functions, $\boldsymbol{\beta} \subseteq \mathbb{R}^p$ is an unknown parameter vector that needs to be estimated, and M is a realization of a zero mean second-order stationary random field. Thus, the response $Y(\mathbf{x}_i)$ is modeled using a trend term $\mathbf{f}(\mathbf{x}_i)^\intercal\boldsymbol{\beta}$ representing the mean response value and a noise term $M(\mathbf{x}_i)$ quantifying our uncertainty about the unknown true response at $\mathbf{x}_i$. The last term $\epsilon_j(\mathbf{x}_i)$ in Equation (1), often called the intrinsic noise, is primarily used in stochastic kriging to model the simulation noise in the jth replication run at $\mathbf{x}_i$. Throughout the paper, we assume that the noise $\epsilon_j(\mathbf{x}_i)$ at a design point $\mathbf{x}_i$ is independent and identically distributed (i.i.d.) across replications.

The goal of stochastic kriging is to construct a metamodel that predicts the response $Y(\mathbf{x}_0)$ at any $\mathbf{x}_0 \in \mathscr{X}$. Let Σ_M be a $k \times k$ covariance matrix across all design points $\mathbf{x}_1, \ldots, \mathbf{x}_k$ with its (i,j)th element given by $\mathrm{Cov}[M(\mathbf{x}_i), M(\mathbf{x}_j)]$. Let $\Sigma_M(\mathbf{x}_0, \cdot) = (\mathrm{Cov}[M(\mathbf{x}_0), M(\mathbf{x}_1)], \ldots, \mathrm{Cov}[M(\mathbf{x}_0), M(\mathbf{x}_k)])^\intercal$ represent the spatial covariances between (an un-sampled point) $\mathbf{x}_0$ and all design points. Let Σ_ϵ be the $k \times k$ covariance matrix associated with the intrinsic simulation noise with (i,j)th element $\mathrm{Cov}[\bar{\epsilon}(\mathbf{x}_i), \bar{\epsilon}(\mathbf{x}_j)]$, where $\bar{\epsilon}(\mathbf{x}_i) = (1/n_i)\sum_{j=1}^{n_i} \epsilon_j(\mathbf{x}_i)$ for all $i = 1, \ldots, k$. We also let $\mathbf{F} = (\mathbf{f}(\mathbf{x}_1), \ldots, \mathbf{f}(\mathbf{x}_k))^\intercal$ be the $k \times p$ matrix of user defined basis functions.

Under the above notation, it has been shown in Ankenman et al. (2010) that when $\boldsymbol{\beta}, \Sigma_M(\mathbf{x}_0, \cdot)$ and Σ_M are known, the MSE-optimal predictor is of the form

$$\hat{y}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)^\intercal\boldsymbol{\beta} + \Sigma_M(\mathbf{x}_0, \cdot)^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}(\bar{\mathbf{y}} - \mathbf{F}\boldsymbol{\beta}) \tag{2}$$

and the corresponding optimal MSE is given by

$$\begin{aligned} &\mathrm{MSE}(\hat{y}(\mathbf{x}_0)) \\ &\quad = \Sigma_M(\mathbf{x}_0, \mathbf{x}_0) - \Sigma_M(\mathbf{x}_0, \cdot)^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_M(\mathbf{x}_0, \cdot). \end{aligned} \tag{3}$$

On the other hand, when $\Sigma_M(\mathbf{x}_0, \cdot)$ and Σ_M are known, but $\boldsymbol{\beta}$ is estimated via the generalized least squares estimator, the MSE-optimal predictor becomes (see, e.g., Chen et al. 2013b)

$$\hat{y}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)^\intercal\hat{\boldsymbol{\beta}} + \Sigma_M(\mathbf{x}_0, \cdot)^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}(\bar{\mathbf{y}} - \mathbf{F}\hat{\boldsymbol{\beta}}), \tag{4}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{F}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{F})^{-1}\mathbf{F}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\bar{\mathbf{y}}$, and the optimal MSE is

$$\begin{aligned} \mathrm{MSE}(\hat{y}(\mathbf{x}_0)) = \Sigma_M(\mathbf{x}_0, \mathbf{x}_0) - \Sigma_M(\mathbf{x}_0, \cdot)^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1} \\ \cdot \Sigma_M(\mathbf{x}_0, \cdot) + \eta^\intercal(\mathbf{F}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{F})^{-1}\eta, \end{aligned} \tag{5}$$

where $\eta = \mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_M(\mathbf{x}_0, \cdot)$.

Suppose in addition to simulation outputs $y_j(\mathbf{x}_i)$, one can also obtain unbiased gradient estimates $\mathscr{D}_j(\mathbf{x}_i) = (\mathscr{D}_j^1(\mathbf{x}_i), \ldots, \mathscr{D}_j^d(\mathbf{x}_i))^\intercal$ of $f(\mathbf{x}_i)$ at $\mathbf{x}_i$ on the jth replication run. For such a setting, Chen et al. (2013a) have introduced an augmented kriging model called stochastic kriging with gradient estimators that explicitly incorporates gradient estimators in constructing an SK predictor. In particular, let $\bar{\mathbf{y}}^+$ be the augmented response vector containing the sample averages of simulation outputs and gradient estimates, i.e., $\bar{\mathbf{y}}^+ = (\bar{y}(\mathbf{x}_1), \ldots, \bar{y}(\mathbf{x}_k), \bar{\mathscr{D}}^1(\mathbf{x}_1), \ldots, \bar{\mathscr{D}}^1(\mathbf{x}_k), \ldots, \bar{\mathscr{D}}^d(\mathbf{x}_1), \ldots, \bar{\mathscr{D}}^d(\mathbf{x}_k))^\intercal$ with $\bar{\mathscr{D}}^\ell(\mathbf{x}_i) = (1/n_i)\sum_{j=1}^{n_i} \mathscr{D}_j^\ell(\mathbf{x}_i)$ for $\ell = 1, \ldots, d$. Define $\mathbf{F}^+ = \begin{pmatrix} \mathbf{F} \\ \mathbf{F}_d \end{pmatrix}$, where $\mathbf{F}_d$ is a $kd \times p$ matrix containing the partial derivatives of the basis functions. Let Σ_M^+ be a covariance matrix representing the spatial covariances between the random field M and partial derivatives of M at all design points, and let $\Sigma_M^+(\mathbf{x}, \cdot)$ be a vector containing the covariances between $\mathbf{x}$ and all design points. Let Σ_ϵ^+ be the covariance matrix of the intrinsic simulation noises involved in estimating true performance measures and gradients at all design points. It has been shown in Chen et al. (2013a) that when $\boldsymbol{\beta}$ is known, the MSE-optimal predictor and its corresponding MSE can be obtained under the SKG framework by replacing $\bar{\mathbf{y}}, \Sigma_M, \Sigma_\epsilon, \Sigma_M(\mathbf{x}, \cdot)$, and $\mathbf{F}$ in Equations (2) and (3) with $\bar{\mathbf{y}}^+, \Sigma_M^+, \Sigma_\epsilon^+, \Sigma_M^+(\mathbf{x}, \cdot)$, and $\mathbf{F}^+$, respectively, whereas when $\boldsymbol{\beta}$ is estimated, the optimal predictor and its MSE can be obtained by doing the same substitutions in Equations (4) and (5).

3. Monotonic Performance of SK and SKG Predictors

In this section, we analyze the performance of SK and SKG metamodels under the setting where design points are selected one at a time, e.g., via a sequential sampling strategy. Our main result is to show that under both the SK and SKG frameworks, and when the parameter vector $\boldsymbol{\beta}$ is either known or estimated (assuming all other parameters defining the covariance matrices of the intrinsic and extrinsic noise are fixed constants), the MSE of the corresponding predictor is monotonically decreasing as the number of design points increases. Our analysis proceeds in three steps. First, we prove the monotonicity of SK by connecting the MSEs of the respective SK predictors based on k and $k+1$ design points. We give explicit formulas showing the degree of reduction in MSE when a new design point is added to the model. Then, we show that essentially the same technique (i.e., by comparing the MSEs of SK and SKG predictors based on the same design points) can be applied to characterize the possible reduction in MSE of the SKG predictor over the standard SK predictor. Finally, we present the monotonicity results in the SKG case, the proofs of which

follow straightforwardly from the derivations used in the previous step.

The following condition, due to Ankenman et al. (2010), is assumed throughout our analysis:

Assumption 1. *The random field M is a zero mean second-order stationary Gaussian random field, and the intrinsic simulation noises $\epsilon_1(\mathbf{x}_i), \epsilon_2(\mathbf{x}_i), \ldots$ are i.i.d. $\mathcal{N}(0, V(\mathbf{x}_i))$, independent of $\epsilon_j(\mathbf{x}_h)$ for all j and $h \neq i$, and independent of M.*

The condition on the random field M implies that the covariance between $M(\mathbf{x}_i)$ and $M(\mathbf{x}_j)$ can be expressed in the form $\mathrm{Cov}[M(\mathbf{x}_i), M(\mathbf{x}_j)] = \tau^2 R_M(d(\mathbf{x}_i, \mathbf{x}_j); \boldsymbol{\theta})$, where $\tau^2 > 0$ is the bounded variance of $M(\mathbf{x})$ at all $\mathbf{x}$, and R_M is the correlation function that depends on the distance $d(\mathbf{x}_i, \mathbf{x}_j)$ between $\mathbf{x}_i$ and $\mathbf{x}_j$ and an unknown parameter vector $\boldsymbol{\theta}$ that needs to be estimated. The independence of the simulation noise across all design points excludes the use of common random numbers; it implies that the covariance matrix Σ_ϵ is a positive semi-definite diagonal matrix. We assume that the correlation function $R_M(d, \boldsymbol{\theta})$ is continuous in its first argument d and satisfies $R_M(0, \boldsymbol{\theta}) = 1$ and $\lim_{d\to\infty} R_M(d, \boldsymbol{\theta}) = 0$. In addition, we also assume that the variance function $V(\mathbf{x})$ is uniformly bounded for all $\mathbf{x} \in \mathscr{X}$.

Unless otherwise specified, we use the subscript k to signify the quantities obtained based on a given set of k design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. Similarly, if a new design point $\mathbf{x}_{k+1}$ is added to an SK (SKG) model, we will use the subscript $k+1$ to denote any quantity that applies to the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{x}_{k+1}\}$. In addition, any quantity related to SKG is further distinguished using the "+" symbol.

3.1. Stochastic Kriging

To show the monotonicity of SK predictors, we need the following intermediate result.

Lemma 1. *If Assumption 1 holds, then the matrix $(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}$ is positive definite for all k.*

Proof. See Section 1 of the online supplement. □

Let $\mathbf{x}_0$ be a prediction point, $\hat{y}_k(\mathbf{x}_0)$ be the SK predictor constructed using Equation (2) based on a set of k design points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$, and let $\hat{y}_{k+1}(\mathbf{x}_0)$ be the resulting predictor when a new design point $\mathbf{x}_{k+1}$ is included in the set. The following result shows that the MSE of $\hat{y}_{k+1}(\mathbf{x}_0)$ cannot be greater than the MSE of $\hat{y}_k(\mathbf{x}_0)$.

Theorem 1. *Suppose that $\mathbf{x}_{k+1} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. For any prediction point $\mathbf{x}_0 \in \mathscr{X}$, let $\mathrm{MSE}(\hat{y}_k(\mathbf{x}_0))$ and $\mathrm{MSE}(\hat{y}_{k+1}(\mathbf{x}_0))$ denote the MSEs of the predictors $\hat{y}_k(\mathbf{x}_0)$ and $\hat{y}_{k+1}(\mathbf{x}_0)$ constructed using Equation (2). If Assumption 1 holds, then $\mathrm{MSE}(\hat{y}_k(\mathbf{x}_0)) \geq \mathrm{MSE}(\hat{y}_{k+1}(\mathbf{x}_0))$.*

Proof. See Section 2 of the online supplement. □

The next result shows that the conclusion of Theorem 1 still holds true when the predictors are constructed using Equation (4). Its proof appears in Section 3 of the online supplement.

Theorem 2. *Suppose that $\mathbf{x}_{k+1} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. For any prediction point $\mathbf{x}_0 \in \mathscr{X}$, let $\mathrm{MSE}(\hat{y}_k(\mathbf{x}_0))$ and $\mathrm{MSE}(\hat{y}_{k+1}(\mathbf{x}_0))$ denote the MSEs of the predictors $\hat{y}_k(\mathbf{x}_0)$ and $\hat{y}_{k+1}(\mathbf{x}_0)$ constructed using Equation (4). If Assumption 1 holds and $\mathbf{F}_k$ has full column rank, then $\mathrm{MSE}(\hat{y}_k(\mathbf{x}_0)) \geq \mathrm{MSE}(\hat{y}_{k+1}(\mathbf{x}_0))$.*

3.2. Stochastic Kriging with Gradient Estimators

Let $\bar{\mathbf{y}}^+$ and $\bar{\mathscr{D}}^\ell(\mathbf{x}_i) = (1/n_i)\sum_{j=1}^{n_i} \mathscr{D}_j^\ell(\mathbf{x}_i)$ be defined as in Section 2. In the SKG framework proposed in Chen et al. (2013a), each (partial) gradient estimator $\mathscr{D}_j^\ell(\mathbf{x}_i)$, $\ell = 1, \ldots, d$, is assumed to take the form

$$\begin{aligned}\mathscr{D}_j^\ell(\mathbf{x}_i) &= \frac{\partial Y(\mathbf{x}_i)}{\partial x_\ell} + \zeta_j^\ell(\mathbf{x}_i) = \left(\frac{\partial \mathbf{f}(\mathbf{x}_i)}{\partial x_\ell}\right)^\intercal \boldsymbol{\beta} + \frac{\partial M(\mathbf{x}_i)}{\partial x_\ell} + \zeta_j^\ell(\mathbf{x}_i) \\ &= D^\ell(\mathbf{x}_i) + \zeta_j^\ell(\mathbf{x}_i), \end{aligned} \tag{6}$$

where $\zeta_j^\ell(\mathbf{x}_i)$, $\ell = 1, \ldots, d$ are the simulation noises associated with gradient estimators at $\mathbf{x}_i$. We separate the averaged simulation responses from gradient estimators in the augmented response vector $\bar{\mathbf{y}}^+$ and rewrite it as: $\bar{\mathbf{y}}^+ = (\bar{\mathbf{y}}^\intercal, \bar{\mathscr{D}}^\intercal)^\intercal$, where $\bar{\mathbf{y}} = (\bar{y}(\mathbf{x}_1), \ldots, \bar{y}(\mathbf{x}_k))^\intercal$ and $\bar{\mathscr{D}} = (\bar{\mathscr{D}}^1(\mathbf{x}_1), \ldots, \bar{\mathscr{D}}^1(\mathbf{x}_k), \ldots, \bar{\mathscr{D}}^d(\mathbf{x}_1), \ldots, \bar{\mathscr{D}}^d(\mathbf{x}_k))^\intercal$. It is easily seen that when gradient information is incorporated, the sum of the spatial covariance matrix Σ_M^+ (between M and its partial derivatives) and the simulation noise covariance matrix Σ_ϵ^+ is simply the covariance matrix of the augmented response vector $\bar{\mathbf{y}}^+$, i.e., $\Sigma_M^+ + \Sigma_\epsilon^+ = E[(\bar{\mathbf{y}}^+ - E[\bar{\mathbf{y}}^+]) \times (\bar{\mathbf{y}}^+ - E[\bar{\mathbf{y}}^+])^\intercal]$. Let $\Sigma_{y,d} = E[(\bar{\mathbf{y}} - E[\bar{\mathbf{y}}])(\bar{\mathscr{D}} - E[\bar{\mathscr{D}}])^\intercal]$ be the $k \times kd$ cross-covariance matrix between the averaged simulation responses and gradient estimators. Denote by $\Sigma_{d,d} = E[(\bar{\mathscr{D}} - E[\bar{\mathscr{D}}])(\bar{\mathscr{D}} - E[\bar{\mathscr{D}}])^\intercal]$ the $kd \times kd$ covariance matrix of the averaged gradient estimators at all design points. For a given prediction point $\mathbf{x}$, we define the covariance vector between $\mathbf{x}$ and all design points as $\Sigma_M^+(\mathbf{x}, \cdot) = (\Sigma_M(\mathbf{x}, \cdot)^\intercal, \Sigma_{M,d}(\mathbf{x}, \cdot)^\intercal)^\intercal$, where $\Sigma_{M,d}(\mathbf{x}, \cdot) = (\mathrm{Cov}[Y(\mathbf{x}), D^1(\mathbf{x}_1)], \ldots, \mathrm{Cov}[Y(\mathbf{x}), D^1(\mathbf{x}_k)], \ldots, \mathrm{Cov}[Y(\mathbf{x}), D^d(\mathbf{x}_1)], \ldots, \mathrm{Cov}[Y(\mathbf{x}), D^d(\mathbf{x}_k)])^\intercal$ and $\Sigma_M(\mathbf{x}, \cdot) = (\mathrm{Cov}[Y(\mathbf{x}), Y(\mathbf{x}_1)], \ldots, \mathrm{Cov}[Y(\mathbf{x}), Y(\mathbf{x}_k)])^\intercal$. We also let $\mathbf{F}^+ = (\mathbf{F}^\intercal, \mathbf{F}_d^\intercal)^\intercal$, where $\mathbf{F}_d = (\partial \mathbf{f}(\mathbf{x}_1)/\partial x_{11}, \ldots, \partial \mathbf{f}(\mathbf{x}_k)/\partial x_{k1}, \ldots, \partial \mathbf{f}(\mathbf{x}_1)/\partial x_{1d}, \ldots, \partial \mathbf{f}(\mathbf{x}_k)/\partial x_{kd})^\intercal$. We make the following assumptions in our subsequent analysis.

Assumption 2. *The simulation noises associated with the gradient estimators $\zeta_1^\ell(\mathbf{x}_i), \ldots, \zeta_{n_i}^\ell(\mathbf{x}_i)$ are i.i.d. with mean zero and variance $V_\zeta^\ell(\mathbf{x}_i) \triangleq Var(\zeta_j^\ell(\mathbf{x}_i))$ for $\ell = 1, \ldots, d$, independent of the random process M and its derivative processes. In addition, $V_\zeta^\ell(\mathbf{x})$ is uniformly bounded on $\mathscr{X}$ for all ℓ, and the noises $\epsilon_l(\mathbf{x}_i)$ and $\zeta_h^\ell(\mathbf{x}_j)$ are independent for all $i \neq j$ and $l \neq h$.*

Assumption 3. *The mean function* $\mathbf{f}(\mathbf{x})^\intercal\boldsymbol{\beta}$ *is differentiable and the second-order mixed derivative of* $R_M(d(\mathbf{x}_i,\mathbf{x}_j),\boldsymbol{\theta})$ *exists and is continuous.*

Note that under Assumptions 1 and 2, noise terms with different replication indices or design points are assumed to be independent, and correlation only exists between $\epsilon_j(\mathbf{x}_i)$ and $\zeta_j^\ell(\mathbf{x}_i)$ at the same design point $\mathbf{x}_i$ within the same replication j. Assumption 3 guarantees that the Gaussian process M has differentiable sample paths and ensures the validity of (6). By the linearity of the differential operator, the first order partial derivative process $D^\ell(\mathbf{x}_i)$, $\ell=1,\dots,d$ of $Y(\mathbf{x}_i)$ is also Gaussian; see e.g., Chen et al. (2013a). Thus, it is natural to also assume the following condition.

Assumption 4. *The vector* $(Y(\mathbf{x}_1),\dots,Y(\mathbf{x}_k),D^1(\mathbf{x}_1),\dots,D^1(\mathbf{x}_k),\dots,D^d(\mathbf{x}_1),\dots,D^d(\mathbf{x}_k))^\intercal$ *has a joint normal distribution with covariance matrix* $\Sigma^+_{M_k}$.

The main results of this section are given in Theorems 3 and 4, which state that a k-point predictor constructed under the SKG framework performs at least as well as the standard SK predictor based on the same design points. Since the comparison is made with respect to the same set of points, we omit the subscript k in these theorems for simplicity.

Theorem 3. *Let* $\hat{y}(\mathbf{x})$ *be the SK predictor constructed using* (2) *and let* $\hat{y}^+(\mathbf{x})$ *be the SKG predictor obtained by substituting* $\bar{\mathbf{y}}^+$, Σ^+_M, Σ^+_ϵ, $\Sigma^+_M(\mathbf{x},\cdot)$, *and* $\mathbf{F}^+$ *for* $\bar{\mathbf{y}}$, Σ_M, Σ_ϵ, $\Sigma_M(\mathbf{x},\cdot)$, *and* $\mathbf{F}$ *in Equation* (2). *If Assumptions 1–4 hold, then* $\mathrm{MSE}(\hat{y}(\mathbf{x}))\ge\mathrm{MSE}(\hat{y}^+(\mathbf{x}))$ *for any* $\mathbf{x}\in\mathscr{X}$.

Proof. See Section 4 of the online supplement. □

The following result shows that the conclusion of Theorem 3 still holds true when SKG predictors are constructed using Equation (4). Its proof is given in Section 5 of the online supplement.

Theorem 4. *Let* $\hat{y}(\mathbf{x})$ *be the SK predictor constructed using* (4) *and* $\hat{y}^+(\mathbf{x})$ *be the SKG predictor obtained by substituting* $\bar{\mathbf{y}}^+,\Sigma^+_M,\Sigma^+_\epsilon,\Sigma^+_M(\mathbf{x},\cdot)$, *and* $\mathbf{F}^+$ *for* $\bar{\mathbf{y}},\Sigma_M,\Sigma_\epsilon,\Sigma_M(\mathbf{x},\cdot)$, *and* $\mathbf{F}$ *in Equation* (4). *If* $\mathbf{F}$ *has full column rank and Assumptions 1–4 hold, then* $\mathrm{MSE}(\hat{y}(\mathbf{x}))\ge\mathrm{MSE}(\hat{y}^+(\mathbf{x}))$ *for any* $\mathbf{x}\in\mathscr{X}$.

Given a set of design points $\{\mathbf{x}_1,\dots,\mathbf{x}_k\}$, consider a prediction point $\mathbf{x}_0\notin\{\mathbf{x}_1,\dots,\mathbf{x}_k\}$. Let $\hat{y}^+_k(\mathbf{x}_0)$ be the SKG predictor constructed using Equation (2) based on the set of k design points $\{\mathbf{x}_1,\dots,\mathbf{x}_k\}$ and $\hat{y}^+_{k+1}(\mathbf{x}_0)$ be the resulting predictor when a new design point $\mathbf{x}_{k+1}$ is included in the set. We now establish the monotonic performance of the SKG predictor by showing that $\hat{y}^+_{k+1}(\mathbf{x}_0)$ has a smaller MSE than $\hat{y}^+_k(\mathbf{x}_0)$.

Similar to Section 3.1, let $\bar{\mathbf{y}}^+_{k+1}$ be the augmented response vector when $k+1$ design points are available. In $\bar{\mathbf{y}}^+_{k+1}$, we separate the averaged response and gradient estimators at $\mathbf{x}_{k+1}$ from those of the rest k design points. In particular, we define $\bar{\mathbf{y}}^+_1=(\bar{y}(\mathbf{x}_{k+1}),\bar{\mathscr{D}}^1(\mathbf{x}_{k+1}),\dots,\bar{\mathscr{D}}^d(\mathbf{x}_{k+1}))^\intercal$ and write

$$\begin{aligned}\bar{\mathbf{y}}^+_{k+1}&=\big(\bar{y}(\mathbf{x}_1),\dots,\bar{y}(\mathbf{x}_{k+1}),\bar{\mathscr{D}}^1(\mathbf{x}_1),\dots,\bar{\mathscr{D}}^1(\mathbf{x}_{k+1}),\dots,\\&\qquad\bar{\mathscr{D}}^d(\mathbf{x}_1),\dots,\bar{\mathscr{D}}^d(\mathbf{x}_{k+1})\big)^\intercal\\&=\big(\bar{y}(\mathbf{x}_1),\dots,\bar{y}(\mathbf{x}_k),\bar{\mathscr{D}}^1(\mathbf{x}_1),\dots,\bar{\mathscr{D}}^d(\mathbf{x}_k),\\&\qquad\bar{y}(\mathbf{x}_{k+1}),\bar{\mathscr{D}}^1(\mathbf{x}_{k+1}),\dots,\bar{\mathscr{D}}^d(\mathbf{x}_{k+1})\big)^\intercal\\&=\big(\bar{\mathbf{y}}_k^{+\intercal},\bar{\mathbf{y}}_1^{+\intercal}\big)^\intercal.\end{aligned}$$

For brevity, we also define the following notation:

$$\begin{aligned}\Sigma^+_{M_{k+1}}+\Sigma^+_{\epsilon_{k+1}}&=E[(\bar{\mathbf{y}}^+_{k+1}-E[\bar{\mathbf{y}}^+_{k+1}])(\bar{\mathbf{y}}^+_{k+1}-E[\bar{\mathbf{y}}^+_{k+1}])^\intercal],\\\Sigma^+_{M_k}+\Sigma^+_{\epsilon_k}&=E[(\bar{\mathbf{y}}^+_k-E[\bar{\mathbf{y}}^+_k])(\bar{\mathbf{y}}^+_k-E[\bar{\mathbf{y}}^+_k])^\intercal],\\\Sigma^+_{k,1}&=E[(\bar{\mathbf{y}}^+_k-E[\bar{\mathbf{y}}^+_k])(\bar{\mathbf{y}}^+_1-E[\bar{\mathbf{y}}^+_1])^\intercal],\\\Sigma^+_{1,1}&=E[(\bar{\mathbf{y}}^+_1-E[\bar{\mathbf{y}}^+_1])(\bar{\mathbf{y}}^+_1-E[\bar{\mathbf{y}}^+_1])^\intercal],\end{aligned}$$

$$\begin{aligned}&\Sigma^+_{M_{k+1}}(\mathbf{x},\cdot)\\&=\big(\mathrm{Cov}(Y(\mathbf{x}),Y(\mathbf{x}_1)),\dots,\mathrm{Cov}(Y(\mathbf{x}),Y(\mathbf{x}_k)),\\&\quad\mathrm{Cov}(Y(\mathbf{x}),D^1(\mathbf{x}_1)),\dots,\mathrm{Cov}(Y(\mathbf{x}),D^1(\mathbf{x}_k)),\\&\quad\dots,\mathrm{Cov}(Y(\mathbf{x}),D^d(\mathbf{x}_1)),\dots,\mathrm{Cov}(Y(\mathbf{x}),D^d(\mathbf{x}_k)),\\&\quad\mathrm{Cov}(Y(\mathbf{x}),Y(\mathbf{x}_{k+1})),\mathrm{Cov}(Y(\mathbf{x}),D^1(\mathbf{x}_{k+1})),\\&\qquad\dots,\mathrm{Cov}(Y(\mathbf{x}),D^d(\mathbf{x}_{k+1}))\big)^\intercal\\&\triangleq\big(\Sigma^+_{M_k}(\mathbf{x},\cdot)^\intercal,\Sigma^+_{M_1}(\mathbf{x},\cdot)^\intercal\big)^\intercal,\end{aligned}\tag{7}$$

$$\begin{aligned}\mathbf{F}^+_{k+1}&=\left(\mathbf{f}(\mathbf{x}_1),\dots,\mathbf{f}(\mathbf{x}_{k+1}),\frac{\partial\mathbf{f}(\mathbf{x}_1)}{\partial x_1},\dots,\frac{\partial\mathbf{f}(\mathbf{x}_{k+1})}{\partial x_1},\dots,\right.\\&\qquad\left.\frac{\partial\mathbf{f}(\mathbf{x}_1)}{\partial x_d},\dots,\frac{\partial\mathbf{f}(\mathbf{x}_{k+1})}{\partial x_d}\right)^\intercal\\&=\left(\mathbf{f}(\mathbf{x}_1),\dots,\mathbf{f}(\mathbf{x}_k),\frac{\partial\mathbf{f}(\mathbf{x}_1)}{\partial x_1},\dots,\frac{\partial\mathbf{f}(\mathbf{x}_k)}{\partial x_1},\dots,\frac{\partial\mathbf{f}(\mathbf{x}_1)}{\partial x_d},\right.\\&\qquad\left.\dots,\frac{\partial\mathbf{f}(\mathbf{x}_k)}{\partial x_d},\mathbf{f}(\mathbf{x}_{k+1}),\frac{\partial\mathbf{f}(\mathbf{x}_{k+1})}{\partial x_1},\dots,\frac{\partial\mathbf{f}(\mathbf{x}_{k+1})}{\partial x_d}\right)^\intercal\\&\triangleq(\mathbf{F}_k^{+\intercal},\mathbf{F}_1^{+\intercal})^\intercal.\end{aligned}\tag{8}$$

Note that in (7), $\Sigma^+_{M_k}(\mathbf{x},\cdot)$ is a $k(d+1)\times1$ vector and $\Sigma^+_{M_1}(\mathbf{x},\cdot)$ is a $(d+1)\times1$ vector, and in (8), $\mathbf{F}^+_k$ is a $k(d+1)\times p$ matrix and $\mathbf{F}^+_1$ is a $(d+1)\times p$ matrix.

Corollary 1. *Suppose that* $\mathbf{x}_{k+1}\notin\{\mathbf{x}_1,\dots,\mathbf{x}_k\}$. *For any prediction point* $\mathbf{x}_0\in\mathscr{X}$, *let* $\hat{y}^+_k(\mathbf{x}_0)$ *and* $\hat{y}^+_{k+1}(\mathbf{x}_0)$ *be SKG predictors constructed using Equation* (2). *If Assumptions 1–4 hold, then* $\mathrm{MSE}(\hat{y}^+_k(\mathbf{x}_0))\ge\mathrm{MSE}(\hat{y}^+_{k+1}(\mathbf{x}_0))$.

Proof. The proof follows straightforwardly by replacing $\mathrm{MSE}(\hat{y}(\mathbf{x}))$, $\mathrm{MSE}(\hat{y}^+(\mathbf{x}))$, $(\Sigma^+_M+\Sigma^+_\epsilon)$, $(\Sigma_M+\Sigma_\epsilon)$, $\Sigma_{y,d}$, $\Sigma_{d,d}$, $\Sigma^+_M(\mathbf{x},\cdot)$, $\Sigma_M(\mathbf{x},\cdot)$, and $\Sigma_{M,d}(\mathbf{x},\cdot)$ in the proof of Theorem 3 with $\mathrm{MSE}(\hat{y}^+_k(\mathbf{x}_0))$, $\mathrm{MSE}(\hat{y}^+_{k+1}(\mathbf{x}_0))$, $(\Sigma^+_{M_{k+1}}+\Sigma^+_{\epsilon_{k+1}})$, $(\Sigma^+_{M_k}+\Sigma^+_{\epsilon_k})$, $\Sigma^+_{k,1}$, $\Sigma^+_{1,1}$, $\Sigma^+_{M_{k+1}}(\mathbf{x}_0,\cdot)$, $\Sigma^+_{M_k}(\mathbf{x}_0,\cdot)$, and $\Sigma^+_{M_1}(\mathbf{x}_0,\cdot)$, respectively. We omit the details. □

Similarly, when SKG predictors are constructed by using Equation (4), we have the same monotonicity property.

Corollary 2. *Suppose that* $\mathbf{x}_{k+1} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. *For any prediction point* $\mathbf{x}_0 \in \mathscr{X}$, *let* $\hat{y}_k^+(\mathbf{x}_0)$ *and* $\hat{y}_{k+1}^+(\mathbf{x}_0)$ *be SKG predictors constructed using Equation* (4). *If Assumptions* 1–4 *hold and* $\mathbf{F}_k^+$ *has full column rank, then* $\mathrm{MSE}(\hat{y}_k^+(\mathbf{x}_0)) \geq \mathrm{MSE}(\hat{y}_{k+1}^+(\mathbf{x}_0))$.

Proof. The proof is identical to the proof of Theorem 4 with $\mathrm{MSE}(\hat{y}_k^+(\mathbf{x}_0))$, $\mathrm{MSE}(\hat{y}_{k+1}^+(\mathbf{x}_0))$, $(\Sigma_{M_{k+1}}^+ + \Sigma_{\epsilon_{k+1}}^+)$, $(\Sigma_{M_k}^+ + \Sigma_{\epsilon_k}^+)$, $\Sigma_{k,1}^+$, $\Sigma_{1,1}^+$, $\Sigma_{M_{k+1}}^+(\mathbf{x}_0, \cdot)$, $\Sigma_{M_k}^+(\mathbf{x}_0, \cdot)$, $\Sigma_{M_1}^+(\mathbf{x}_0, \cdot)$, $\mathbf{F}_{k+1}^+$, $\mathbf{F}_k^+$, $\mathbf{F}_1^+$ replacing $\mathrm{MSE}(\hat{y}(\mathbf{x}))$, $\mathrm{MSE}(\hat{y}^+(\mathbf{x}))$, $(\Sigma_M^+ + \Sigma_\epsilon^+)$, $(\Sigma_M + \Sigma_\epsilon)$, $\Sigma_{y,d}$, $\Sigma_{d,d}$, $\Sigma_M^+(\mathbf{x}, \cdot)$, $\Sigma_M(\mathbf{x}, \cdot)$, $\Sigma_{M,d}(\mathbf{x}, \cdot)$, $\mathbf{F}^+$, $\mathbf{F}$, $\mathbf{F}_d$. □

4. An Adaptive Sequential Kriging Method

Given a set of design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ and the numbers of simulation replications $\{n_1, \ldots, n_k\}$ allocated to each point, let $\tilde{y}_k(\mathbf{x})$ be the SK predictor (i.e., $\tilde{y}_k(\mathbf{x}) \triangleq \hat{y}_k(\mathbf{x})$) or SKG predictor (i.e., $\tilde{y}_k(\mathbf{x}) \triangleq \hat{y}_k^+(\mathbf{x})$) constructed using (4). Since $\tilde{y}_k(\mathbf{x})$ aims to provide a global fit of an unknown response surface, the integrated MSE (IMSE)

$$\mathrm{IMSE}(k) \triangleq \int_{\mathbf{x} \in \mathscr{X}} \mathrm{MSE}(\tilde{y}_k(\mathbf{x}))\, d\mathbf{x} \tag{9}$$

serves as a useful criterion to measure the quality of the model over the design space. In the setting when the total simulation budget N is fixed and the number of design points K is specified, Ankenman et al. (2010) propose a two-stage design strategy that uses a small number m of predetermined design points to estimate model parameters (e.g., $V(\mathbf{x})$, τ^2, and $\boldsymbol{\theta}$) and then jointly selects the rest of the $K - m$ design points and allocation of replications to minimize an estimated IMSE. To avoid solving a high-dimensional non-linear optimization problem at the second stage, the authors recommend choosing the remaining design points in a space-filling way. Thus, the emphasis of the strategy is focused on efficient allocation of simulation replications.

In this section, we consider the alternative setting of sequentially selecting design points and simulation replications in order to attain a desired IMSE target $\varepsilon > 0$. To address the simulation allocation issue, we show in Section 4.1 that under both the SK and SKG frameworks, the optimal predictor MSE at a design point $\mathbf{x}_i$ is dominated by the variance of the averaged intrinsic noise at $\mathbf{x}_i$

$$\mathrm{MSE}(\tilde{y}_k(\mathbf{x}_i)) \leq \frac{V(\mathbf{x}_i)}{n_i}. \tag{10}$$

In addition, the MSE is continuous in the sense that $\mathrm{MSE}(\tilde{y}_k(\mathbf{x}))$ will stay close to $\mathrm{MSE}(\tilde{y}_k(\mathbf{x}_i))$ for any new location $\mathbf{x}$ sufficiently close to $\mathbf{x}_i$. Consequently, to ensure the overall IMSE to fall below a given threshold ε (as the number of design points becomes large), it is sufficient to have $\mathrm{MSE}(\tilde{y}_k(\mathbf{x}_i)) < \varepsilon/|\mathscr{X}|$ at all sampled design points $\mathbf{x}_i$, where $|\mathscr{X}|$ is the volume of $\mathscr{X}$. This, together with (10), leads to the condition $n_i = \lceil (V(\mathbf{x}_i)|\mathscr{X}|)/\varepsilon \rceil$, suggesting that n_i should be chosen proportionally to the intrinsic variance at $\mathbf{x}_i$, where $\lceil a \rceil$ indicates the smallest integer that is greater than a.

Motivated by the monotonicity results derived in Section 3, once the number of simulation replications at a design point is determined, we consider an adaptive sequential approach that maximizes the difference between successive IMSEs at each iteration. In particular, given $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ and $\{n_1, \ldots, n_k\}$, we consider selecting the next design point $\mathbf{x}_{k+1}$ with $n_{k+1} = \lceil (V(\mathbf{x}_{k+1})|\mathscr{X}|)/\varepsilon \rceil$ replications allocated to it that achieves the maximum reduction in IMSE at each iteration:

$$\begin{aligned}
\mathbf{x}_{k+1} &= \arg\max_{\mathbf{x} \in \mathscr{X}} (\mathrm{IMSE}(k) - \mathrm{IMSE}(k+1)) \\
&= \arg\max_{\mathbf{x} \in \mathscr{X}} \int_{\mathbf{x}_0 \in \mathscr{X}} (\mathrm{MSE}(\tilde{y}_k(\mathbf{x}_0)) - \mathrm{MSE}(\tilde{y}_{k+1}(\mathbf{x}_0)))\, d\mathbf{x}_0 \\
&= \arg\min_{\mathbf{x} \in \mathscr{X}} \int_{\mathbf{x}_0 \in \mathscr{X}} \mathrm{MSE}(\tilde{y}_{k+1}(\mathbf{x}_0))\, d\mathbf{x}_0,
\end{aligned} \tag{11}$$

where $\mathrm{MSE}(\tilde{y}_{k+1}(\mathbf{x}_0))$ (and hence $\mathrm{IMSE}(k+1)$) is viewed as a function of the new location $\mathbf{x}$. The last equality follows because $\mathrm{MSE}(\tilde{y}_k(\mathbf{x}_0))$ is constant with respect to the decision variable $\mathbf{x}$.

The above discussion assumes that the intrinsic variance function $V(\mathbf{x})$ is known. In practice, it can be estimated by applying the approach outlined in Ankenman et al. (2010), which uses the standard (deterministic) kriging method to fit a spatial correlation model of the form $V(\mathbf{x}) = \sigma^2 + Z(\mathbf{x})$ with Z being another mean zero stationary random field that is independent of M. Specifically, since $V(\mathbf{x})$ is not directly observable, the intrinsic variance $V(\mathbf{x}_i)$ at a design point $\mathbf{x}_i$ is replaced by its sample variance calculated using the n_i simulation observations at $\mathbf{x}_i$. The estimates of $V(\mathbf{x}_i)$ are then used in (4) to construct an optimal MSE predictor $\hat{V}(\mathbf{x})$ by simply ignoring the intrinsic noise. Once $\hat{V}(\mathbf{x})$ is obtained, the rest of the parameters $(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta})$ can be estimated in the way described in Ankenman et al. (2010) by constructing the log-likelihood function and then applying a standard non-linear optimization algorithm to search for the maximum likelihood estimators of $(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta})$. The same techniques can also be used to estimate the model parameters in the SKG framework; we refer the reader to Chen et al. (2013a) for more details.

To summarize, we proposed the following strategy, which we refer to as adaptive sequential kriging (ASK), for obtaining experiment designs in constructing an SK (SKG) predictor with a predefined level of accuracy:

Step 0. Specify an IMSE target $\varepsilon > 0$, a set of initial space-filling design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, and numbers of simulation replications $\{n_1, \ldots, n_m\}$. Collect output performance measures (including gradient information if using SKG) at each $\mathbf{x}_i$. Set $k = m$.

Step 1. Fit $\hat{V}$ (and $\hat{V}_\zeta^\ell$ for all ℓ) and construct MLEs $(\hat{\beta}_k, \hat{\tau}_k^2, \hat{\theta}_k)$ using all performance measures collected thus far.

Step 2. Choose the next design point $\mathbf{x}_{k+1}$ as

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}\in\mathscr{X}} \int_{\mathbf{x}_0\in\mathscr{X}} \widehat{\mathrm{MSE}}(\tilde{y}_{k+1}(\mathbf{x}_0))\, d\mathbf{x}_0,$$

where $\widehat{\mathrm{MSE}}(\tilde{y}_{k+1}(\mathbf{x}_0))$ is an estimator of $\mathrm{MSE}(\tilde{y}_{k+1}(\mathbf{x}_0))$; see Remark 3.

Step 3. Allocate $n_{k+1} = \lceil (\hat{V}(\mathbf{x}_{k+1})|\mathscr{X}|)/\varepsilon \rceil$ replications to $\mathbf{x}_{k+1}$, collect output performance measures at $\mathbf{x}_{k+1}$, and calculate $\widehat{\mathrm{MSE}}(\tilde{y}_{k+1}(\mathbf{x}_0))$ by incorporating the new design point $\mathbf{x}_{k+1}$ into the current model.

Step 4. If $\widehat{\mathrm{IMSE}}(k+1) \triangleq \int_{\mathbf{x}_0\in\mathscr{X}} \widehat{\mathrm{MSE}}(\tilde{y}_{k+1}(\mathbf{x}_0))\, d\mathbf{x}_0 \le \varepsilon$, then terminate; otherwise set $k = k+1$ and go to step 1.

Remark 1. When IMSE is used as an error measure, the prediction performance depends on the volume of the design space (see (9)). However, in practice, it might be desirable to have a performance measure that does not rely on $\mathscr{X}$. One simple choice of such a measure is to consider the average IMSE (AIMSE) obtained by normalizing the IMSE with respect to the volume of the design space $|\mathscr{X}|$. It is easy to see that if the desired accuracy ε in the initialization step of ASK is instead specified using AIMSE, then the choice of n_{k+1} at Step 3 simplifies to $n_{k+1} = \lceil \hat{V}(\mathbf{x}_{k+1})/\varepsilon \rceil$, which no longer depends on $|\mathscr{X}|$. The two performance measures IMSE and AIMSE are only different in terms of their implementations, but are equivalent in theory.

Remark 2. In (deterministic) kriging, the estimation of model parameters requires the calculation of the determinant and inverse of the covariance matrix Σ_M. Sometimes, e.g., when some design points are very close to each other, Σ_M may become ill-conditioned or near-singular, leading to numerical instability or a significant loss of accuracy in parameter estimation. Note that since $\Sigma_\epsilon(\mathbf{x}_{k+1}, \mathbf{x}_{k+1}) = V(\mathbf{x}_{k+1})/n_{k+1}$, the choice of n_{k+1} at Step 3 leads to an almost constant simulation precision at all sampled design points. In view of (3) and (5), this effect is equivalent to introducing a nugget ε in the model, where ε corresponds to a prescribed AIMSE threshold (see Remark 1). Thus, to avoid the parameter estimation issue, the value of ε can be selected to ensure that the matrix $\Sigma_M + \varepsilon I$ is well-conditioned. For example, one approach as discussed in Ranjan et al. (2011) is to choose ε such that the condition number of $\Sigma_M + \epsilon I$ is smaller than a certain prescribed (very large) threshold. In our numerical experiments conducted in Section 5, we set ε to 0.01, which keeps the matrix $\Sigma_M + \varepsilon I$ far from being ill-conditioned.

Remark 3. Recall from (11) that at any candidate design point $\mathbf{x}$, the number of simulation replications at $\mathbf{x}$ is predetermined by $\lceil (V(\mathbf{x})|\mathscr{X}|)/\varepsilon \rceil$ so that $\mathrm{MSE}(\tilde{y}_{k+1}(\mathbf{x}_0))$, as well as $\mathrm{IMSE}(k+1)$, can be viewed as a function of $\mathbf{x}$. However, since the model parameters are estimated, the $\mathrm{IMSE}(k+1)$ in (11) is estimated at Step 2 of the algorithm via the integral $\int_{\mathbf{x}_0\in\mathscr{X}} \widehat{\mathrm{MSE}}(\tilde{y}_{k+1}(\mathbf{x}_0))\, d\mathbf{x}_0$, where at a given point $\mathbf{x}$, $\widehat{\mathrm{MSE}}(\tilde{y}_{k+1}(\mathbf{x}_0))$ is the plug-in estimator of $\mathrm{MSE}(\tilde{y}_{k+1}(\mathbf{x}_0))$ obtained by replacing in (5) the true intrinsic variance $V(\mathbf{x})$ with $\hat{V}(\mathbf{x})$ and the true model parameters $\boldsymbol{\beta}$, τ^2, and $\boldsymbol{\theta}$ with their corresponding MLEs $\hat{\beta}_k$, $\hat{\tau}_k^2$, and $\hat{\theta}_k$. The value of the integral can then be approximated by numerical integration (e.g., Gaussian quadrature; see Section 5) or Monte Carlo methods.

Remark 4. Note that although the plug-in estimator $\widehat{\mathrm{MSE}}(\tilde{y}_{k+1}(\mathbf{x}_0))$ is easy to implement, a potential drawback of the estimator is that it tends to underestimate the true predictor MSE (when taking into account the fact that model parameters are estimated; see, e.g., den Hertog et al. 2006). Consequently, the true predictor IMSE may also be underestimated by $\widehat{\mathrm{IMSE}}(k+1)$ at Step 4, causing early stopping of the algorithm. It is very difficult to theoretically quantify the estimation error of $\widehat{\mathrm{IMSE}}(k+1)$. If the early stopping issue is of practical concern, a parameterized bootstrapping method can be employed to obtain an empirical estimate of the correct predictor MSE at Step 4; we refer the reader to den Hertog et al. (2006) for more details.

4.1. Validity of ASK

In this section, we justify the validity of the proposed ASK algorithm under the idealized setting when all model parameters are known. In particular, by exploiting the monotonicity properties derived in Section 3, we show that if the design points and simulation replications are sequentially determined according to the ASK procedure, then the IMSE of the resulting predictor can be made smaller than a given threshold by increasing the number of design points.

To establish our main result, we state a list of lemmas and corollaries, which show that under both the SK and SKG frameworks and when either $\boldsymbol{\beta}$ is known or estimated, the optimal predictor MSE at a sampled design point $\mathbf{x}_i$ is always upper bounded by the right-hand-side of (10); moreover, for a design point $\mathbf{x}_i$ with a small MSE value, the MSE at any point in the vicinity of $\mathbf{x}_i$ will also stay small. The proofs of these results can be found in the online supplement. Not surprisingly, these results are consistent with our understanding of deterministic kriging models where the predictor variance is zero at all design locations; see, e.g., Vazquez and Bect (2010). Note that for simplicity, we have assumed a constant trend model $\mathbf{f}(\mathbf{x})^\top\boldsymbol{\beta} = \beta$ (i.e., $p = 1$, $\mathbf{f}(\mathbf{x}) = 1\ \forall\, \mathbf{x}$ and $\mathbf{F} = \mathbf{1}_k$) in our analysis.

Lemma 2. *Given a set of design points* $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, *let* $\hat{y}(\mathbf{x})$ *be the SK predictor constructed using* (2). *Let* $B_r(\mathbf{x})$ *be an open ball centered at* $\mathbf{x}$ *with radius* $r > 0$. *For any*

$\mathbf{x}_i \in \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ *and* $\varepsilon > 0$, *if Assumption 1 holds and* $n_i > V(\mathbf{x}_i)/\varepsilon$, *then* (a) $\mathrm{MSE}(\hat{y}(\mathbf{x}_i)) < \varepsilon$; (b) *there exists an* $r_i > 0$ *such that* $\mathrm{MSE}(\hat{y}(\mathbf{x})) < \varepsilon$ *for all* $\mathbf{x} \in B_{r_i}(\mathbf{x}_i) \cap \mathscr{X}$.

Corollary 3. *Given a set of design points* $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, *let* $\hat{y}(\mathbf{x})$ *be the SK predictor constructed using* (4). *For any* $\mathbf{x}_i \in \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ *and* $\varepsilon > 0$, *if Assumption 1 holds and* $n_i > V(\mathbf{x}_i)/\varepsilon$, *then* (a) $\mathrm{MSE}(\hat{y}(\mathbf{x}_i)) < \varepsilon$; (b) *there exists an* $r_i > 0$ *such that* $\mathrm{MSE}(\hat{y}(\mathbf{x})) < \varepsilon$ *for all* $\mathbf{x} \in B_{r_i}(\mathbf{x}_i) \cap \mathscr{X}$.

Lemma 3. *Given a set of design points* $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, *let* $\hat{y}^+(\mathbf{x})$ *be the SKG predictor constructed using* (2). *For any* $\mathbf{x}_i \in \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ *and* $\varepsilon > 0$, *if Assumptions 1–4 hold and* $n_i > V(\mathbf{x}_i)/\varepsilon$, *then* (a) $\mathrm{MSE}(\hat{y}^+(\mathbf{x}_i)) < \varepsilon$; (b) *there exists an* $r_i > 0$ *such that* $\mathrm{MSE}(\hat{y}^+(\mathbf{x})) < \varepsilon$ *for all* $\mathbf{x} \in B_{r_i}(\mathbf{x}_i) \cap \mathscr{X}$.

Corollary 4. *Given a set of design points* $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, *let* $\hat{y}^+(\mathbf{x})$ *be the SKG predictor constructed using* (4). *For any* $\mathbf{x}_i \in \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ *and* $\varepsilon > 0$, *if Assumptions 1–4 hold and* $n_i > V(\mathbf{x}_i)/\varepsilon$, *then* (a) $\mathrm{MSE}(\hat{y}^+(\mathbf{x}_i)) < \varepsilon$; (b) *there exists an* $r_i > 0$ *such that* $\mathrm{MSE}(\hat{y}^+(\mathbf{x})) < \varepsilon$ *for all* $\mathbf{x} \in B_{r_i}(\mathbf{x}_i) \cap \mathscr{X}$.

Proof. Follows directly from Corollary 3 and Theorem 4. □

The previous results indicate that for every design point $\mathbf{x}_i$ generated by ASK, there exists an open ball $B_{r_i}(\mathbf{x}_i)$ so that the MSE at any point in the ball can be made small. Intuitively, since ASK minimizes IMSE at each step, new design points should be chosen in the complement of the union of these open balls. Thus, as new points are generated, the collection of open balls increases and will cover the entire (compact) design space in finite time, at which point the desired IMSE target is attained. This intuition leads to the following main theorem of this section. Its proof appears in Section 9 of the online supplement.

Theorem 5. *Let* $\mathbf{x}_1, \mathbf{x}_2, \ldots$ *be the sequence of design points generated by the ASK algorithm and* $\varepsilon > 0$ *be a given tolerance. Suppose that Assumptions 1–4 hold and the number of simulation replications* $n_i > (V(\mathbf{x}_i)|\mathscr{X}|)/\varepsilon$ *for all* i, *then* $\lim_{k\to\infty} \mathrm{IMSE}(k) \le \varepsilon$.

5. Numerical Examples

In this section, we conduct some computational experiments to illustrate the performance of the proposed ASK procedure. We consider two sets of examples: an M/M/1 queue and five deterministic functions with added noise. In both cases, the performance of ASK is compared with those of AEES and the sequential MSE (SMSE) approach adapted from deterministic simulation. AEES uses an adaptive weighing approach to maintain a balance between explorative and exploitative searches of the design space and allocates a simulation budget to design points proportionate to their estimated intrinsic variances. SMSE employs the same simulation allocation rule as in ASK but adaptively selects the design point that achieves the maximum MSE at each step. In both ASK and SMSE, the optimal design point selection problem is solved using a global optimization algorithm proposed in Hu and Hu (2011).

In all comparison procedures, the SK (SKG) predictors are constructed using an (unknown) constant trend model with a Gaussian correlation function $R_M(d(\mathbf{x}_i, \mathbf{x}_j), \boldsymbol{\theta}) = \exp(-\Sigma_{\ell=1}^{d} \theta_\ell (x_{i\ell} - x_{j\ell})^2)$, where θ_ℓ and $x_{i\ell}$ are the respective ℓth components of $\boldsymbol{\theta}$ and $\mathbf{x}_i$. There are many other choices of the correlation function; the interested reader is referred to Xie et al. (2010) for an empirical analysis on how the choice of correlation function may affect the prediction quality of a kriging model. The variance functions V and V_ζ^ℓ are fitted using the standard kriging models assuming the same correlation structure. Since the true response curves of all test functions are known, the performance of a predictor is measured by the average integrated squared error (AISE), defined by

$$AISE = \frac{\int_{\mathbf{x}\in\mathscr{X}} (f(\mathbf{x}) - \hat{y}(\mathbf{x}))^2 \, d\mathbf{x}}{|\mathscr{X}|}, \tag{12}$$

which compares the true response value with the metamodel-predicted response value. Throughout our experiments, the integral in (12) and those in Steps 2 and 4 of the algorithm are evaluated by a 7-point Gaussian quadrature rule. For example, a one dimensional integral $\int_a^b g(x)\,dx$ is approximated by the weighted sum $\frac{1}{2}(b-a)\sum_{i=1}^{7} w_i g(\frac{1}{2}(b-a)x_i + \frac{1}{2}(a+b))$, whereas in two dimensions, $\int_a^b \int_c^d g(x,y)\,dx\,dy \approx \frac{1}{4}(b-a)\times(d-c)\sum_{i=1}^{7}\sum_{j=1}^{7} w_i w_j g(\frac{1}{2}(b-a)x_i + \frac{1}{2}(a+b), \frac{1}{2}(d-c)y_i + \frac{1}{2}(c+d))$, where the quadrature points x_i, y_i and weights $w_i, i = 1, \ldots, 7$ can be found in, e.g., Abramowitz and Stegun (1972). A similar rule can be given for a 3-dimensional region.

5.1. An M/M/1 Queueing Example

This example is taken from Ankenman et al. (2010). Consider an M/M/1 queue with service rate 1 and arrival rate $x \in (0,1)$. Let $f(x)$ be the long run expected number of customers in system. Elementary queueing theory shows that $f(x) = x/(1-x)$. Our goal is to model the response surface $f(x)$ over the domain $[0.05, 0.95]$ in a stochastic simulation setting. For a given arrival rate x, the response value $f(x)$ can be estimated via the time-average $\bar{f}(x) = (1/t)\int_0^t N_s(x)\,ds$ by performing a single simulation run, where N_s is the number of observed customers in system at time s.

In the implementation of ASK, we have used an AIMSE target of 0.01 and 5 initial space-filling design points over $[0.05, 0.95]$. The same 5 initial points are also used in both AEES and SMSE. At each initial design point, we perform 100 independent simulation runs of length $t = 1{,}000$ time units. Each simulation run is initiated by sampling from the steady-state distribution to eliminate the initial transience. Thus, it is reasonable to assume that the estimator $\bar{f}(x)$ is unbiased.

Table 1. AISEs (Mean ± SE) Obtained and the Respective Numbers of Simulation Replications and Design Points (Mean ± SE) Required by ASK, AEES, and SMSE on the M/M/1 Queueing Example When the AIMSE Reaches 0.01

	ASK	AEES	SMSE
AISE	9.4e−3 ± 5.8e−4	4.4e−3 ± 2.5e−4	1.3e−2 ± 1.6e−3
# replications	754.3 ± 78.9	1,984.5 ± 32.2	924.2 ± 13.7
# design points	13 ± 0.3	25 ± 0.5	19 ± 0.5

Note. All results are based on 100 independent runs.

Table 2. AISEs (Mean ± SE) Obtained and the Respective Numbers of Simulation Replications and Design Points (Mean ± SE) Required by ASK, AEES, and SMSE on the M/M/1 Queueing Example When the AISE Reaches 0.01

	ASK	AEES	SMSE
AISE	7.3e−3 ± 5.7e−4	9.4e−3 ± 5.5e−4	8.9e−3 ± 4.6e−4
# replications	786.3 ± 69.4	1,674.2 ± 23.4	1,239.8 ± 20.3
# design points	14 ± 0.5	23 ± 0.4	23 ± 0.5

Note. All results are based on 100 independent runs.

Each comparison algorithm is then repeated independently 10[illegible]nes.

Table 1 shows the AISE values obtained, along with the numbers of simulation replications and design points required by these procedures upon termination. From the table, we see that the mean AISE obtained by ASK upon stopping is very close to the AIMSE target, whereas those obtained by AEES and SMSE are slightly different from the AIMSE target. Therefore, in order to have a fair efficiency comparison, we have performed another experiment that runs each algorithm until the true AISE drops below 0.01. The performances of the three algorithms are reported in Table 2. In Figure 1(a), we also plot the mean AISEs (in logarithmic scale) obtained by these procedures as functions of the number of new design points selected. The fitted response surfaces (averaged over 100 runs) are plotted in Figure 1(b).

From the figure, we observe that all three algorithms show a similar initial improvement in AISE. As the number of design points increases, the performance of ASK is consistently improving and surpasses the other two competing algorithms after about five design points. AEES shows a very slow improvement in AISE when the number of design points is less than 13. We conjecture that this is related to the rapid increase of the response surface $f(x)$ when the arrival rate x becomes large (say exceeds 0.9). The search for new design points in AEES relies on a discretization method that resembles the space-filling strategy. Thus, when the number of design points is small, a coarse discretization of the design space may cause the algorithm to fail to capture the extreme trend of $f(x)$ over small intervals (when x is large), leading to poor initial performance. Table 2 shows that ASK terminates with the smallest mean AISE value by using the least numbers of simulation replications and design points. Note that both ASK and SMSE use the same simulation allocation rule and differ only in the way the design points are selected. This indicates that, in addition to the allocation of simulation replications, careful selection of design points is equally important and may have a significant impact on predictor quality.

It is also interesting to observe in Figure 1(a) that the mean AISE curve of AEES does not exhibit a monotone decrease as the number of design points increases. This is because in all three algorithms, model parameters are constantly adjusted based on available information obtained from sampled design points. Thus, a change in the model parameter estimates may lead to a temporary increase in IMSE. This does not contradict the monotonicity results of Section 3, which are established under fixed model parameters.

Since in ASK, all model parameters are sequentially estimated, an experiment is also performed on the queueing example to examine the robustness of the performance of ASK with respect to the parameter estimation error. In particular, we approximate the true model parameters by sampling 50 space-filling design points over the design space (again 100 independent simulation replications are assigned to each point generated) and then constructing the MLEs of model parameters using all performance measures collected. We then fix the parameter estimates (treating them as true model parameters) and run the ASK algorithm until the true AISE drops below 0.01. Figure 2 shows the performance of ASK as the number of design points increases. The numbers of total simulation replications consumed and design points added are also reported in Table 3. Figure 2 clearly indicates the monotonicity of the AISE of the SK predictor with respect to the number of design points added, which conforms well with our theoretical findings obtained in Section 3. In addition, we see from Table 3 that both the number of replication runs and number of design points added are smaller but very close to the corresponding entries shown in Table 2. This suggests that use of true model parameters can result in additional computational efficiency gains, but the consequence of using estimated parameters in the proposed ASK algorithm is not significant.

5.2. Deterministic Examples with Added Noise

We consider the following set of benchmark functions (see, e.g., Ajdari and Mahlooji 2014, Qu and Fu 2014) in our experiments. Problems (1)–(4) are highly nonlinear with multiple extreme values, so predicting the true response surfaces of these functions is difficult; whereas problem (5) is a smooth function aimed at

Figure 1. An M/M/1 Queueing Example

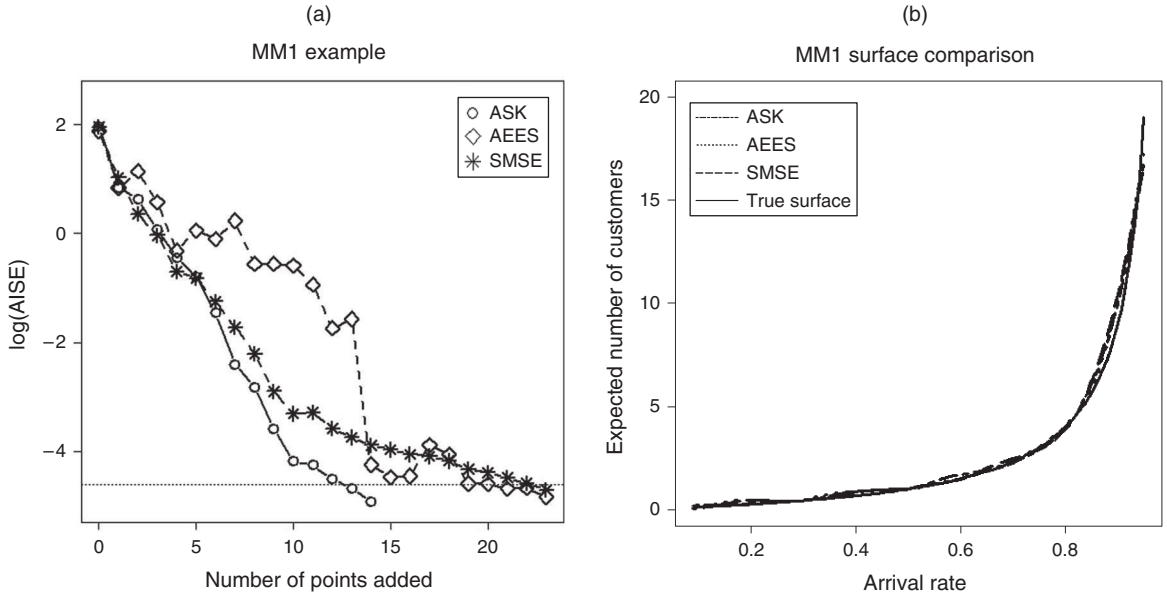
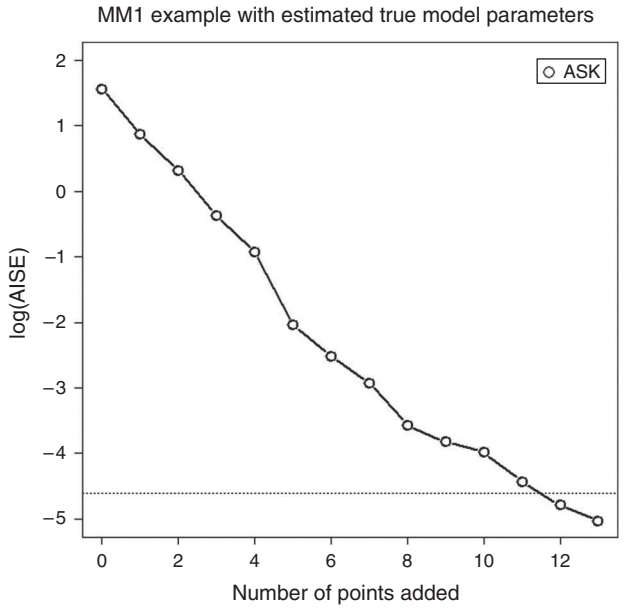


Figure 2. Performance of ASK with Estimated True Model Parameters



illustrating the performance of the algorithms in three dimensions.

(1) $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x})$, $\mathbf{x} = (x_1, x_2)^\intercal \in [-1,1] \times [-1,1]$, where $f(\mathbf{x}) = 4x_1^2 - 2.1x_1^4 + x_1^6/3 + x_1x_2 - 4x_2^2 + 4x_2^4$ and $\epsilon(\mathbf{x}) \sim \mathcal{N}(0, V(\mathbf{x}))$.

(2) $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x})$, $\mathbf{x} = (x_1, x_2)^\intercal \in [-1,1] \times [-1,1]$, where $f(\mathbf{x}) = x_1 \sin(\pi x_2) + x_2 \sin(\pi x_1)$ and $\epsilon(\mathbf{x}) \sim \mathcal{N}(0, V(\mathbf{x}))$.

(3) $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x})$, $\mathbf{x} = (x_1, x_2)^\intercal \in [-1,1] \times [-1,1]$, where $f(\mathbf{x}) = 3(1-x_1)^2 \exp(-x_1^2 - (x_2+1)^2) - 10(x_1/5 - x_1^3 - x_2^5)\exp(-x_1^2 - x_2^2) - \frac{1}{3}\exp(-(x_1+1)^2 - x_2^2)$ and $\epsilon(\mathbf{x}) \sim \mathcal{N}(0, V(\mathbf{x}))$.

(4) $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x})$, $\mathbf{x} = (x_1, x_2)^\intercal \in [-10,10] \times [-10,10]$, where $f(\mathbf{x}) = 1 + x_1^2/4{,}000 + x_2^2/4{,}000 - \cos(x_1)\cos(x_2/\sqrt{2})$ and $\epsilon(\mathbf{x}) \sim \mathcal{N}(0, V(\mathbf{x}))$.

(5) $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x})$, $\mathbf{x} = (x_1, x_2, x_3)^\intercal \in [-1,1] \times [-1,1] \times [-1,1]$, where $f(\mathbf{x}) = x_1^2 + x_2x_3$ and $\epsilon(\mathbf{x}) \sim \mathcal{N}(0, V(\mathbf{x}))$.

We implement ASK, AEES, and SMSE under both the SK and SKG frameworks on examples (1)–(5). In the SKG case, we assume that noisy gradient estimates $\mathcal{D}_j^\ell(\mathbf{x}) = \partial f(\mathbf{x})/\partial x_\ell + \zeta_j^\ell(\mathbf{x})$ are available at $\mathbf{x}$ on the jth simulation replication, where $\zeta_j^\ell(\mathbf{x}) \sim \mathcal{N}(0, V_\zeta(\mathbf{x}))$ for $\ell = 1, \ldots, d$. To investigate the impact of different variance functions on the performance of ASK, we consider two sets of variance functions in our experiments: in the SK case, $V_1(\mathbf{x}) = 0.1|f(\mathbf{x})| + 0.1$ and $V_2(\mathbf{x}) = 0.2|f(\mathbf{x})| + 0.1$; in the SKG case, $V(\mathbf{x}) = 0.2|f(\mathbf{x})| + 0.1$, $V_{\zeta_1}(\mathbf{x}) = 0.5|f(\mathbf{x})| + 0.1$ and $V_{\zeta_2}(\mathbf{x}) = |f(\mathbf{x})| + 0.1$. We have experimented with different numbers of initial space-filling designs and found empirically that the performance of the algorithms is not sensitive to this choice, provided that these numbers are not chosen too small so that sensible estimates of model parameters can be obtained.

Table 3. AISEs (Mean ± SE) Obtained and the Respective Numbers of Simulation Replications and Design Points (Mean ± SE) Required by ASK (Assuming True Model Parameters Are Known) When the AISE Reaches 0.01

AISE	# replications	# design points
8.3e−3 ± 6.3e−4	731.0 ± 62.4	13 ± 0.6

Note. All results are based on 100 independent runs.

Table 4. Mean AISEs ± SE (Mean #design Points ± SE) Obtained When the Estimated AIMSE Reaches 0.01 Under the SK Framework with Different Noise Variance Functions

$V(\mathbf{x})$	Fcn.	ASK	AEES	SMSE
V_1	(1)	1.8e−2 ± 4.4e−3 (13 ± 0.3)	1.9e−2 ± 3.8e−3 (15 ± 0.4)	1.5e−2 ± 4.9e−3 (18 ± 0.4)
	(2)	1.3e−2 ± 3.1e−3 (12 ± 0.6)	1.5e−2 ± 4.3e−3 (12 ± 0.4)	7.3e−3 ± 2.7e−4 (20 ± 0.6)
	(3)	1.0e−2 ± 4.6e−3 (21 ± 0.8)	6.3e−3 ± 1.1e−4 (33 ± 0.4)	1.4e−2 ± 3.7e−3 (22 ± 0.5)
	(4)	1.0e−2 ± 3.5e−3 (55 ± 1.2)	7.8e−3 ± 8.1e−4 (130 ± 2.6)	9.9e−3 ± 2.2e−4 (150 ± 2.8)
	(5)	8.5e−3 ± 3.5e−4 (9 ± 0.3)	2.8e−2 ± 3.5e−3 (10 ± 0.4)	1.2e−2 ± 1.5e−3 (10 ± 0.3)
V_2	(1)	1.6e−2 ± 4.1e−3 (15 ± 0.5)	1.8e−2 ± 6.8e−3 (17 ± 0.3)	1.8e−2 ± 5.2e−3 (19 ± 0.5)
	(2)	1.0e−2 ± 3.8e−3 (14 ± 0.6)	1.4e−2 ± 2.4e−3 (14 ± 0.5)	7.8e−3 ± 3.1e−4 (19 ± 0.6)
	(3)	1.4e−2 ± 5.5e−3 (21 ± 1.4)	7.8e−3 ± 3.0e−4 (36 ± 1.9)	1.3e−2 ± 3.7e−3 (25 ± 0.6)
	(4)	9.7e−3 ± 3.2e−4 (56 ± 2.3)	6.8e−3 ± 1.2e−4 (137 ± 5.4)	8.8e−3 ± 3.8e−4 (154 ± 5.6)
	(5)	9.4e−3 ± 2.6e−4 (9 ± 0.5)	2.5e−2 ± 1.6e−3 (10 ± 0.3)	3.0e−2 ± 2.6e−3 (7 ± 1.1)

Note. All results are based on 100 independent runs.

Table 5. Mean AISEs ± SE (Mean #design Points ± SE) Obtained When the Estimated AIMSE Reaches 0.01 Under the SKG Framework with Different Noise Variance Functions

$V_\zeta(\mathbf{x})$	Fcn.	ASK	AEES	SMSE
V_{ζ_1}	(1)	9.8e−3 ± 5.6e−4 (5 ± 0.6)	9.4e−3 ± 7.2e−4 (11 ± 0.6)	1.5e−2 ± 1.5e−3 (7 ± 0.8)
	(2)	1.3e−2 ± 6.7e−4 (5 ± 0.3)	1.5e−2 ± 2.0e−3 (12 ± 0.3)	6.0e−3 ± 4.7e−4 (7 ± 0.6)
	(3)	7.0e−3 ± 6.2e−4 (7 ± 0.3)	5.6e−3 ± 5.2e−4 (12 ± 0.3)	9.8e−3 ± 8.3e−4 (9 ± 0.4)
	(4)	8.6e−3 ± 1.6e−4 (47 ± 1.1)	9.5e−3 ± 2.3e−4 (78 ± 2.2)	9.6e−3 ± 8e−4 (65 ± 1.3)
	(5)	9.5e−3 ± 5.8e−4 (7 ± 0.4)	1.0e−2 ± 5.2e−4 (8 ± 0.6)	9.5e−3 ± 6.9e−4 (7 ± 0.5)
V_{ζ_2}	(1)	8.5e−3 ± 5.5e−3 (7 ± 0.5)	1.1e−2 ± 8.3e−4 (11 ± 0.6)	7.1e−3 ± 6.4e−4 (8 ± 0.7)
	(2)	1.0e−2 ± 6.3e−4 (5 ± 0.4)	1.8e−2 ± 1.5e−3 (12 ± 0.3)	7.6e−3 ± 9.0e−4 (7 ± 0.8)
	(3)	9.1e−3 ± 9.0e−4 (7 ± 0.5)	6.7e−3 ± 5.0e−4 (17 ± 1.1)	9.1e−3 ± 8.1e−4 (9 ± 0.6)
	(4)	9.6e−3 ± 2.0e−4 (49 ± 0.3)	1.1e−2 ± 5.2e−4 (80 ± 0.9)	9.0e−3 ± 1.5e−4 (75 ± 1.8)
	(5)	8.9e−3 ± 3.0e−4 (11 ± 0.5)	9.3e−3 ± 3.8e−4 (13 ± 0.4)	9.1e−3 ± 4.2e−4 (11 ± 0.4)

Note. All results are based on 100 independent runs.

In the experiments considered in this section, we set the number of initial space-filling designs to 10 under SK and 5 under SKG, which work reasonably well in all test cases. For a general rule of thumb on selecting the number of design points in an initial experiment, we refer the reader to Loeppky et al. (2009). At each initial point, 30 replications are allocated to evaluate the function performance. We stop all three algorithms when the estimate $\widehat{\text{IMSE}}$ falls below $\varepsilon = 0.01|\mathscr{X}|$, which corresponds to an estimated AIMSE target of 0.01.

In Tables 4 and 5, we record the mean AISEs obtained and the numbers of additional design points required by the three algorithms on each of the respective test cases. Test results indicate that ASK delivers superior performance over the other two competing algorithms in all cases, in the sense that it reaches the desired accuracy by using the least number of design points. In addition, note that the AISEs obtained by the three algorithms are reasonably close to the prescribed AIMSE threshold. In particular, we see that in ASK, the actual mean AISEs are smaller than 0.01 in 15 out of the 20 cases, and those AISE values that are larger than the AIMSE threshold mostly occur in the SK case when the number of design points is relatively small. This suggests that, as compared to AISE, the influence of using the estimated AIMSE as a stopping criterion to measure the predictor performance is typically not significant provided that the number of design points is not very small.

To further compare the computational efficiency of ASK, AEES, and SMSE, we use the AISE as a stopping criterion and continue running each algorithm until the AISE drops below 0.01. Figures 3–6 show the performance of these algorithms by plotting the mean AISE curves (in logarithmic scale) as functions of the numbers of added design points in each of the respective test cases. Tables 6 and 7 give the numbers of simulation replications consumed by different algorithms to attain the desired AISE level. It is easy to observe that ASK outperforms, or at least shows comparable performance to, AEES and SMSE in terms of both the number of simulation replications and the number of design points used. In addition, we see that under the SKG framework, the numbers of design points required by ASK to achieve the AISE target are generally between 10 and 54 (including the initial points), which are much

Figure 3. Performance of ASK, AEES, and SMSE Under the SK Framework on Test Functions (1)–(3)

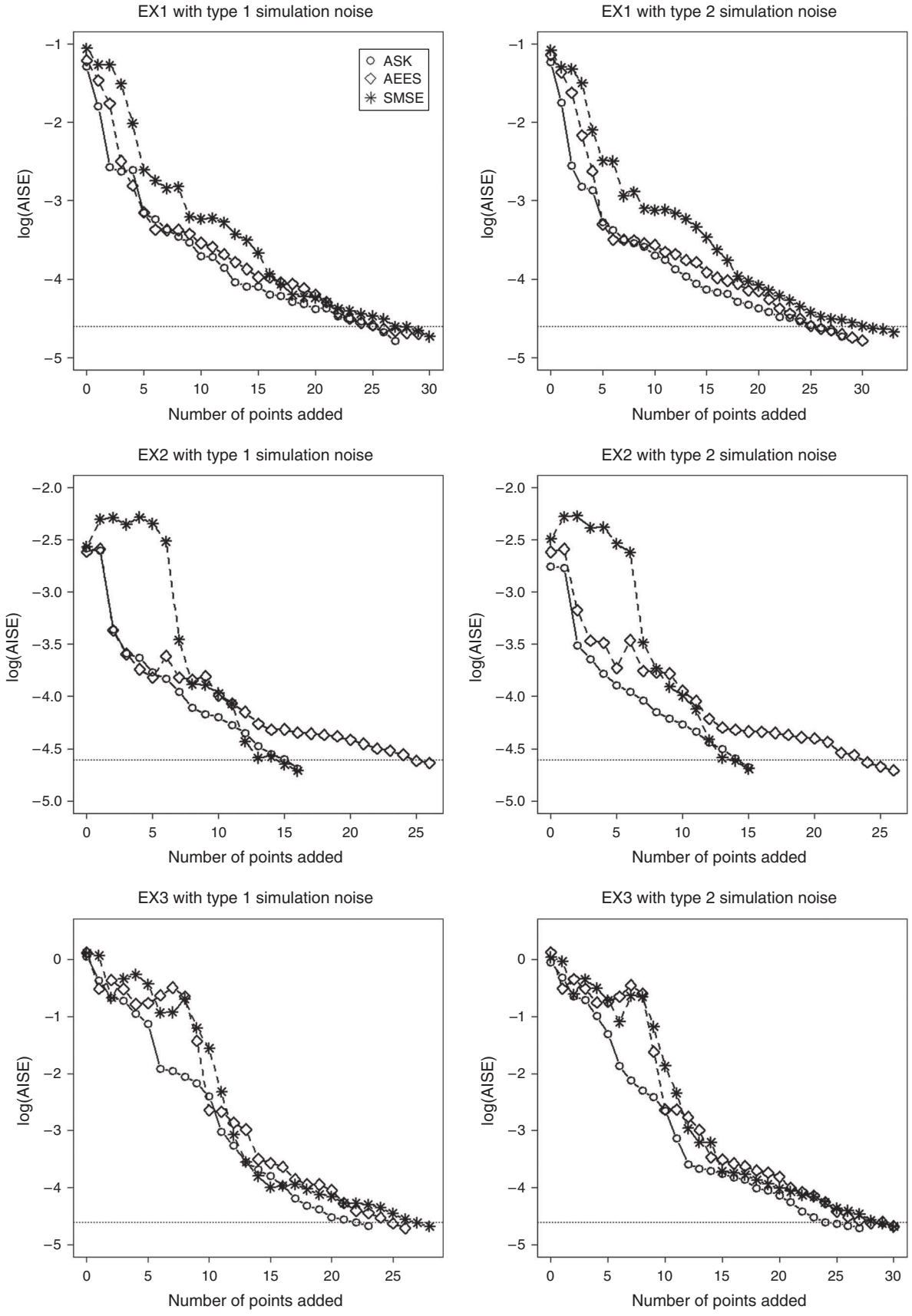
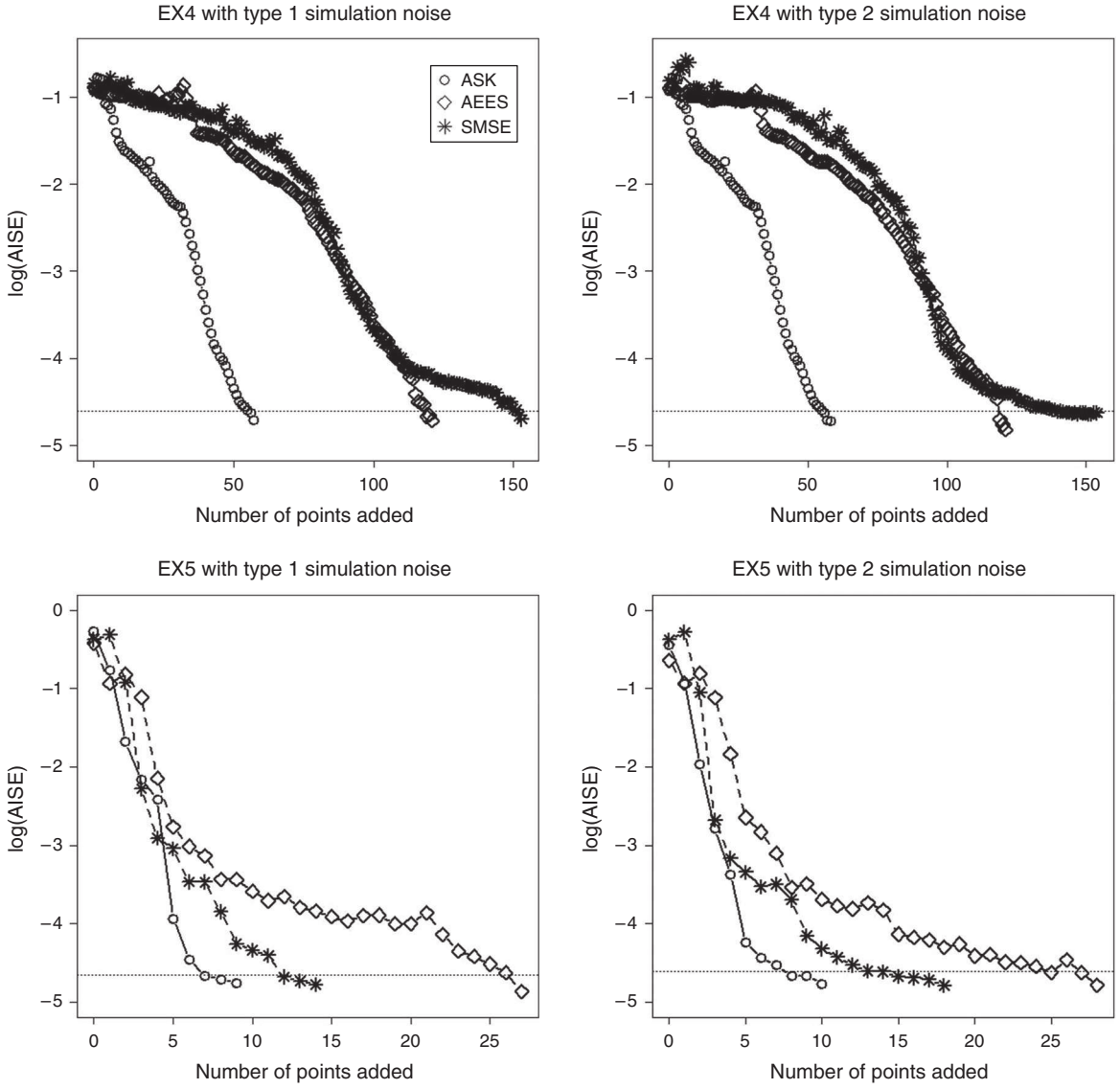


Figure 4. Performance of ASK, AEES, and SMSE Under the SK Framework on Test Functions (4)–(5)



smaller than those required in the SK case (generally between 19 and 66). This empirically illustrates the benefit of using gradient information for improving the prediction performance of kriging metamodels. However, to rigorously quantify the potential reductions in the numbers of design points and simulation replications as a result of using SKG as opposed to SK can be very difficult, and is clearly an open issue that merits further investigation.

Another observation is that, from Figures 3, 4 and Table 6, the mean numbers of simulation replications in the SK case become larger when the noise variance increases, whereas the numbers of added design points required by each algorithm are very close under different noise variance functions. An intuitive explanation is that in all three algorithms, the number of simulation replications is chosen proportional to the intrinsic noise variance, which leads to an almost constant simulation estimation precision at all design points. Thus, a change in the noise variance will mainly impact the overall number of simulation replications consumed, but may have little influence on the selection of design points. We also find from Table 6 that when the variability in simulation responses is increased by a factor of 2 (i.e., when the variance function is changed from V_1 to V_2), the number of simulation replications required by SK increases by a factor of between 1.1 and 2.9. In contrast, a two-fold increase in the variance of gradient estimates in the SKG case (i.e., from V_{ζ_1} to V_{ζ_2} in Table 7) merely results in a 0.5% to 30% increase in the number of simulation runs. This seems to suggest that the variability in gradient estimates tends to have a smaller influence on the algorithm performance than the variability in simulation responses. Of course, a

Figure 5. Performance of ASK, AEES, and SMSE Under the SKG Framework on Test Functions (1)–(3)

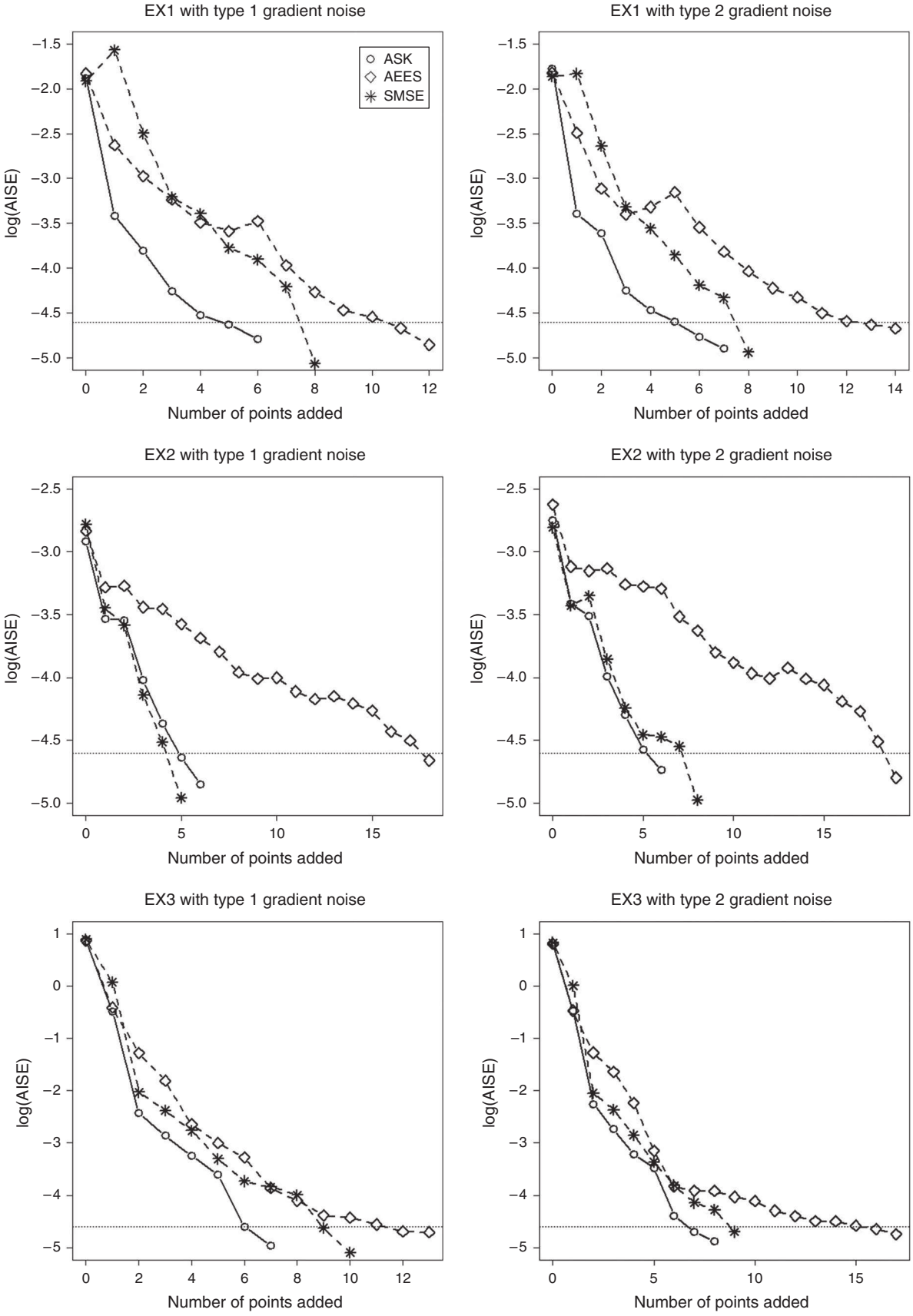


Figure 6. Performance of ASK, AEES, and SMSE Under the SKG Framework on Test Functions (4)–(5)

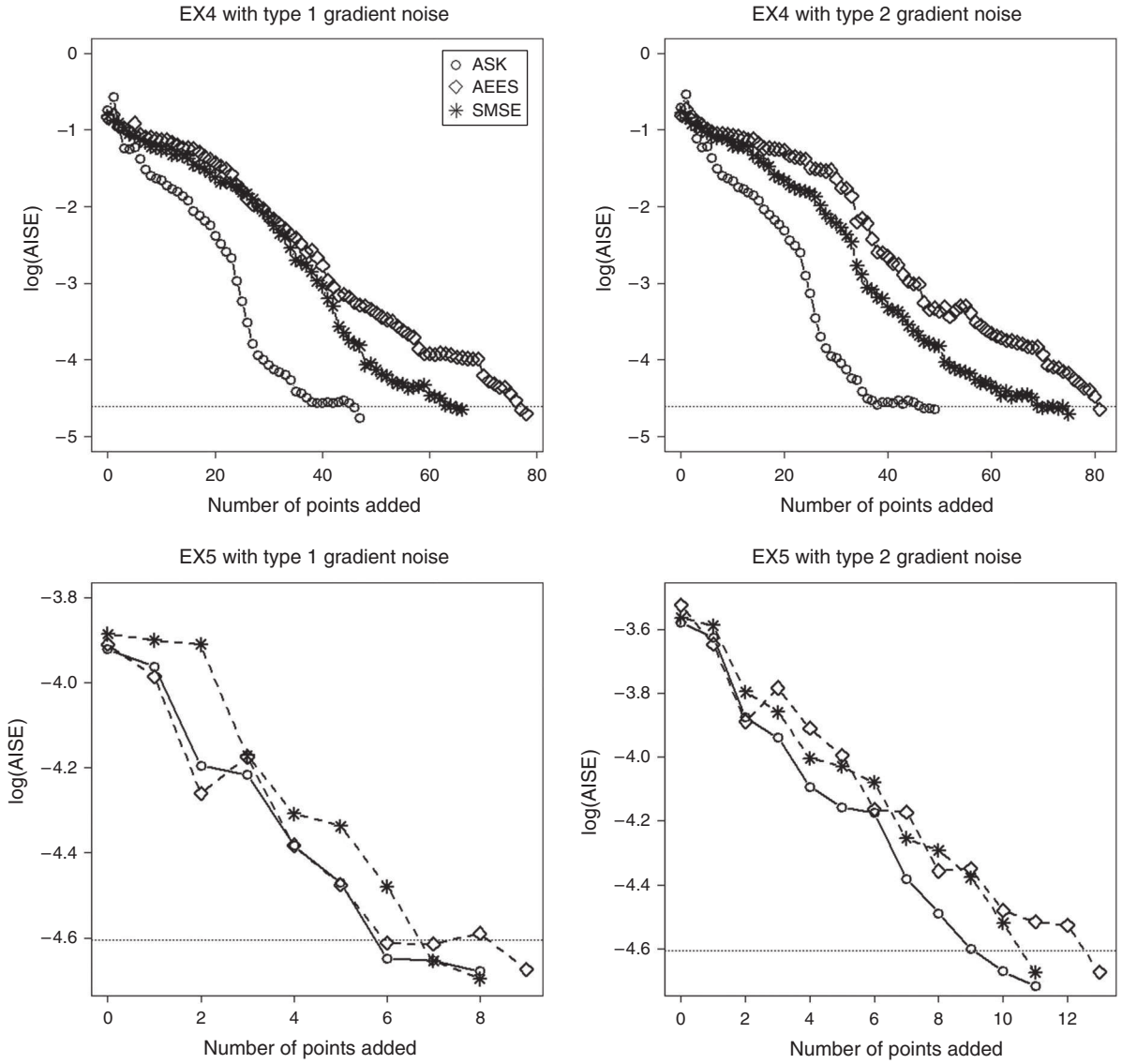


Table 6. Mean #simulation Replications ± SE (Mean #design Point ± SE) Required to Reach an AISE of at Least 0.01 for Different Algorithms Under the SK Framework with Different Noise Variance Functions

$V(\mathbf{x})$	Fcn.	ASK	AEES	SMSE
V_1	(1)	839.7 ± 8.1 (27 ± 1.1)	1,504.3 ± 16.8 (29 ± 0.9)	1,009.5 ± 13.9 (30 ± 1.2)
	(2)	561.6 ± 5.4 (16 ± 0.6)	1,279.9 ± 11.9 (26 ± 0.4)	553.0 ± 4.0 (16 ± 0.5)
	(3)	898.9 ± 13.0 (23 ± 0.6)	1,471.7 ± 31.3 (26 ± 0.3)	1,079.9 ± 16.9 (28 ± 0.6)
	(4)	1,467.4 ± 15.1 (57 ± 2.4)	2,513.4 ± 45.2 (121 ± 3.3)	3,410.7 ± 6.1 (153 ± 3.2)
	(5)	470.8 ± 5.7 (9 ± 0.4)	1,229.4 ± 4.0 (27 ± 0.6)	589.9 ± 9.0 (14 ± 0.7)
V_2	(1)	1,132.2 ± 23.6 (28 ± 0.5)	2,048.0 ± 31.4 (30 ± 0.6)	1,514 ± 31.1 (33 ± 0.5)
	(2)	636.4 ± 7.9 (15 ± 0.4)	1,648.5 ± 17.7 (26 ± 0.5)	614.3 ± 7.3 (15 ± 0.5)
	(3)	1,354.2 ± 20.9 (27 ± 0.6)	2,264.0 ± 35.9 (30 ± 0.4)	1,663.0 ± 23.7 (30 ± 0.5)
	(4)	4,213.2 ± 15.2 (58 ± 3.2)	4,505.4 ± 58.7 (121 ± 2.7)	5,035.0 ± 13.2 (154 ± 2.8)
	(5)	563.2 ± 14.6 (10 ± 0.7)	1,592.1 ± 11.9 (28 ± 0.5)	836.3 ± 21.7 (18 ± 0.4)

Note. All results are based on 100 independent runs.

Table 7. Mean #simulation Replications ± SE (Mean #design Points ± SE) Required to Reach an AISE of at Least 0.01 for Different Algorithms Under the SKG Framework with Different Gradient Noise Variance Functions

$V_\zeta(\mathbf{x})$	Fcn.	ASK	AEES	SMSE
V_{ζ_1}	(1)	333.4 ± 0.9 (6 ± 0.7)	369.9 ± 0.8 (12 ± 0.4)	554.7 ± 0.9 (8 ± 0.5)
	(2)	373.9 ± 1.4 (6 ± 0.4)	575.3 ± 1.4 (18 ± 0.7)	276.4 ± 0.5 (5 ± 0.7)
	(3)	437.5 ± 2.9 (7 ± 0.7)	428.1 ± 1.5 (13 ± 0.6)	775.6 ± 4.1 (10 ± 0.6)
	(4)	2,111.7 ± 27.9 (47 ± 1.2)	2,187.7 ± 2.5 (78 ± 2.1)	2,758.7 ± 16.5 (66 ± 1.4)
	(5)	375.8 ± 4.1 (8 ± 0.3)	405.0 ± 0.8 (9 ± 0.5)	520.0 ± 5.2 (8 ± 0.5)
V_{ζ_2}	(1)	359.4 ± 1.1 (7 ± 0.5)	408.1 ± 1.3 (14 ± 0.3)	554.4 ± 1.0 (8 ± 0.6)
	(2)	375.5 ± 1.4 (6 ± 0.6)	601.6 ± 1.4 (19 ± 0.5)	389.6 ± 0.5 (8 ± 0.6)
	(3)	456.3 ± 4.2 (8 ± 0.4)	518.3 ± 1.2 (17 ± 0.7)	891.1 ± 2.2 (9 ± 0.7)
	(4)	2,192.3 ± 18.8 (49 ± 0.5)	2,276.0 ± 9.3 (81 ± 1.1)	3,194.0 ± 3 (75 ± 2.2)
	(5)	493.6 ± 13.5 (11 ± 0.4)	551.6 ± 4.0 (13 ± 0.6)	736.1 ± 11.0 (11 ± 0.7)

Note. All results are based on 100 independent runs.

more comprehensive study needs to be carried out to confirm this finding.

6. Conclusion

In this paper, we have investigated the performance of SK and SKG metamodels in a fully sequential setting. Our main contributions are the theoretical findings that the MSE of an SK (SKG) predictor is monotonically decreasing in the number of design points and that the use of available gradient information in SKG can in general lead to improved prediction performance. These findings not only complement existing results in the (stochastic) kriging literature, but may also have utility in the design and analysis of sequential sampling procedures under both SK and SKG frameworks. In particular, we have proposed a new design procedure called adaptive sequential kriging for adaptively selecting design points and simulation allocations in obtaining an SK (SKG) predictor with a prescribed level of accuracy. By exploiting the monotonicity of SK and SKG, we have theoretically justified the validity of the proposed procedure. Our preliminary numerical results indicate that ASK may yield high-quality SK (SKG) predictors within a small number of simulation replications and provide superior performance over some of the existing procedures.

It should be noted that the monotonicity results established in this paper rely critically on the specific forms of the optimal predictor MSEs, which are derived under the assumption that the model parameters are either perfectly known or predetermined. Unfortunately, this assumption is rarely satisfied in practice. When model parameters are estimated using simulation samples, these parameters themselves become random variables, in which case the monotonicity results hold with respect to the plug-in estimators of the MSEs (conditional on the estimates based on previous design points), but may not hold for the true MSEs. Thus, an important future research topic is to investigate whether this idealistic assumption of known model parameters can be removed and to what extent the monotonicity results can be generalized when the true parameters are replaced with their estimates.

Acknowledgments

The authors thank the department editor, associate editor, and two anonymous reviewers for their many helpful comments and suggestions, which have led to a substantially improved paper.

References

Abramowitz M, Stegun IA (1972) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (National Bureau of Standards, Gaithersburg, MD).

Ajdari A, Mahlooji H (2014) An adaptive exploration-exploitation algorithm for constructing metamodels in random simulation using a novel sequential experimental design. *Comm. Statist.—Simulation Comput.* 43(5):947–968.

Ankenman B, Nelson BL, Staum J (2010) Stochastic kriging for simulation metamodeling. *Oper. Res.* 58(2):371–382.

Barton RR (2009) Simulation optimization using metamodels. *Proc. 2009 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 230–238.

Barton RR, Meckesheimer M (2006) Metamodel-based simulation optimization. Henderson SG, Nelson BL, eds. *Handbooks Oper. Res. Management Sci.* Vol. 13 (North-Holland, Amsterdam), 535–574.

Chang KH, Hong LJ, Wan H (2013) Stochastic trust-region response-surface method (STRONG)—A new response-surface framework for simulation optimization. *INFORMS J. Comput.* 25(2):230–243.

Chen X, Ankenman BE, Nelson BL (2013a) Enhancing stochastic kriging metamodels with gradient estimators. *Oper. Res.* 61(2): 512–528.

Chen X, Wang K, Yang F (2013b) Stochastic kriging with qualitative factors. *Proc. 2013 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 790–801.

den Hertog D, Kleijnen J, Siem A (2006) The correct kriging variance estimated by bootstrapping. *J. Oper. Res. Soc.* 57(4):400–409.

Hu J, Hu P (2011) Annealing adaptive search, cross-entropy, and stochastic approximation in global optimization. *Naval Res. Logist. (NRL)* 58(5):457–477.

Jin R, Chen W, Sudjianto A (2002) On sequential sampling for global metamodeling in engineering design. *ASME 2002 Internat. Design Engrg. Tech. Conf. Comput. Inform. Engrg. Conf.* (ASME, New York), 539–548.

Kleijnen JP (2007) *Design and Analysis of Simulation Experiments*, Vol. 111 (Springer International, Cham, Switzerland).

Kleijnen JP (2009) Kriging metamodeling in simulation: A review. *Eur. J. Oper. Res.* 192(3):707–716.

Loeppky JL, Sacks J, Welch WJ (2009) Choosing the sample size of a computer experiment: A practical guide. *Technometrics* 51(4):366–376.

Qu H, Fu MC (2014) Gradient extrapolated stochastic kriging. *ACM Trans. Modeling Comput. Simulation* 24(4):Article No. 23.

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

Sacks J, Welch WJ, Mitchell TJ, Wynn HP (1989) Design and analysis of computer experiments. *Statist. Sci.* 4(4):409–423.

Santner TJ, Williams BJ, Notz W (2003) *The Design and Analysis of Computer Experiments* (Springer, New York).

Ulaganathan S, Couckuyt I, Dhaene T, Laermans E, Degroote J (2014) On the use of gradients in kriging surrogate models. *Proc.* 2014 *Winter Simulation Conf.* (IEEE, Piscataway, NJ), 2692–2701.

Vazquez E, Bect J (2010) Pointwise consistency of the kriging predictor with known mean and covariance functions. *mODa* 9—*Advances in Model-Oriented Design and Analysis. Contributions to Statistics* (Physica-Verlag, Heidelberg, Germany), 221–228.

Viana FA, Simpson TW, Balabanov V, Toropov V (2014) Metamodeling in multidisciplinary design optimization: How far have we really come? *AIAA J.* 52(4):670–690.

Wang GG, Shan S (2006) Review of metamodeling techniques in support of engineering design optimization. *J. Mech. Design* 129(4):370–380.

Wang L (2005) A hybrid genetic algorithm-neural network strategy for simulation optimization. *Appl. Math. Comput.* 170(2): 1329–1243.

Xie W, Nelson B, Staum J (2010) The influence of correlation functions on stochastic kriging metamodels. *Proc.* 2010 *Winter Simulation Conf.* (IEEE, Piscataway, NJ), 1067–1078.

Xie W, Nelson BL, Barton RR (2014) A bayesian framework for quantifying uncertainty in stochastic simulation. *Oper. Res.* 62(6): 1439–1452.

Yang F, Liu J, Nelson BL, Ankenman BE, Tongarlak M (2011) Metamodelling for cycle time-throughput-product mix surfaces using progressive model fitting. *Production Planning and Control* 22(1):50–68.

Author Queries

A1 Au: Please confirm names, affiliations, and email addresses are okay as set. Provide author(s) ORCID number(s) if applicable.

A2 Au: Please confirm keywords.

A3 Au: Please confirm heading levels are correct.

A4 Au: Please verify that all equations are set correctly.

A5 Au: We have changed "to choose" to "choosing" here. Please check.

A6 Au: We have changed "proportional" to "proportionally" here. Please check.

A7 Au: We have changed "search" to "searches" here. Please check.

A8 Au: Please check that all figures and tables are set correctly.

A9 Au: In Tables 1–7, we have maintained first sentence alone as caption and rest as notes. Kindly check.

A10 Au: All figures were relabeled. Kindly check.