Monotonicity Properties of Stochastic Kriging Metamodels and Related Applications

A Dissertation presented

by

Bing Wang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Concentration - Operations Research)

Stony Brook University

January 2016

Stony Brook University

The Graduate School

Bing Wang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation

Jiaqiao Hu - Dissertation Advisor
Associate Professor, Department of Applied Mathematics and Statistics

Song Wu - Chairperson of Defense
Assistant Professor, Department of Applied Mathematics and Statistics

Haipeng Xing - Committee Member
Associate Professor, Department of Applied Mathematics and Statistics

Jin Wang - Committee Member
Professor, Department of Chemistry

Xin Wang - Outside Committee Member
Associate Professor, Department of Electrical and Computer Engineering

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Monotonicity Properties of Stochastic Kriging Metamodels and Related Applications

by

Bing Wang

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Concentration - Operations Research)

Stony Brook University

2016

Stochastic kriging (SK) and stochastic kriging with gradient estimators (SKG) are popular approaches to approximate complex simulation models because of their ability to replace the expensive simulation outputs by metamodel values. Obtaining an accurate SK/SKG metamodel is highly desirable in practice. This dissertation studies the monotonicity properties of the mean squared error (MSE) of optimal SK and SKG predictors. In particular, we show that in both SK and SKG, the MSEs of the corresponding optimal predictors are non-increasing functions of the numbers of design points. Based on these findings, we design an adaptive sequential sampling approach to obtain SK/SKG predictors with a pre-defined level of accuracy. In each step, our approach selects the point that achieves the maximum reduction in the current integrated MSE (IMSE) and adaptively allocates the number of simulation replications. Theoretical analysis is also provided to guarantee that a desired performance can be achieved. We run numerical examples to justify the monotonicity properties of the predictors under both SK and SKG frameworks, and illustrate the effectiveness of the proposed approach by comparing its performance with two other existing methods. The comparison results indicate that our approach can be more efficient both in terms of the number of design points used and the simulation efforts expended.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SK	stochastic kriging
SKG	stochastic kriging with gradient estimator
MSE	mean square error of the SK or SKG predictor
IMSE	integrated mean square error
AISE	average integrated square error
AIMSE	average integrated mean square error
ASK	adaptive sequential kriging algorithm
SMSE	sequential sampling algorithm based on MSE measure
AEES	adaptive exploration-exploitation sampling algorithm

Acknowledgements

I would like to express my sincere gratitude to my advisor Dr. Jiaqiao Hu for his warm guidance, support and encouragement throughout my graduate study.

I would also like to thank my defense committee: Dr. Song Wu, Dr. Haipeng Xing, Dr. Jin Wang and Dr. Xin Wang, as well as my preliminary exam committee chair Dr. Estie Arkin, for their kindness to provide insightful comments on my thesis.

Last but not least, I would like to thank my family and my girlfriend Jiayu for their endless love and support. I am also thankful to all my friends for our invaluable friendship that encourages me all the time.

1 Kriging Metamodeling in Simulation

1.1 Background

In many real life problems, evaluating the performance of an objective function may be computationally expensive and time consuming. For some complicated simulation systems, it may take days or weeks to finish a single simulation run. For example, it is reported that it takes Ford Motor Company about 36-160 hours to run one crash simulation [1]. Suppose on average 50 iterations are needed to solve an optimization problem, and assume each iteration needs one crash simulation, the total computation time would be 75 days to 11 months, which is unacceptable in practice [2]. Similar problems that involve complex and heavy computational tasks are also common in the area of geo-information inference, stock price prediction and computer design experiments. Consequently, obtaining an accurate simulation output in a computationally efficient manner is highly desirable.

Metamodeling, or surrogate model, is a well recognized solution to address the above-mentioned challenge. A metamodel aims at approximating the response of a simulation model based on limited evaluations of the model. It is considered as a model of the simulation model. Metamodels usually focus on studying the the input and output relationships. By fitting the correct metamodel, a computationally expensive simulation model can be replaced by its metamodel, and the common statistical analysis can be applied. Commonly used metamodels include radial basis functions, neural networks and kriging, etc. In this dissertation, we focus on one of the most popular metamodels, kriging. Kriging and its related methods are advantageous for the efficient usage of simulation computational resources. Specifically, they are capable of obtaining accurate estimates of performance measures with relatively small amount of design points, which is particularly important when the cost of obtaining sample data is high. Kriging and its related methods have been widely used in environmental science, mining, natural resources and engineering ([3], [4], [5], [6], [7]).

Kriging was originally proposed in geostatistics studies by the South African mining engineer Krige. The mathematical mechanism was later developed by Matheron[8] in 1963. It is an interpolation-based metamodeling technique to mimic the behavior of an unknown objective function. Originally kriging models were applied in deterministic simulation models; see Sacks et al.'s 1989 article [9] and recent publications [10],[11],[12]. Recently, kriging has been extended to stochastic settings. Stochastic kriging (SK) considers sampling noise which is inherent to stochastic simulation systems [13]. In adddition, it is well known in the community of design and analysis of computer experiments (DACE) that gradient information are valuable in smoothing the response surface [10]. In order to incorporate gradient information into the original stochastic kriging model, a new stochastic kriging with gradient estimators (SKG) method was developed by Chen et al. [14]. SKG was able to achieve better prediction performance as the expense of using additional computational efforts in obtaining gradient estimates.

1.2 Kriging

1.2.1 Problem Setting

Consider the problem of describing the response surface of an unknown function $f(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$, where $\mathbf{x}$ is a vector of design variables and $\mathcal{X}$ is a compact full-dimensional subset of $\mathbb{R}^d$. The goal is to model the unknown response surface $f(\mathbf{x})$. Different assumptions on $f(\mathbf{x})$ will lead to different forms of kriging metamodels. For example, when the response $f(\mathbf{x})$ can be evaluated exactly, universal and simple kriging models can be fitted. On the other hand, when $f(\mathbf{x})$ can only be estimated in a path-wise manner through stochastic simulations, stochastic kriging models can be applied. In this section, three types of kriging models in deterministic setting will be introduced.

1.2.2 Universal Kriging

In deterministic computer experiments, the response $f(\mathbf{x})$ can be observed without noise and a metamodel can be developed after observing $f(\mathbf{x})$ at several design points. A successful approach is to formulate this problem into statistical framework by representing the unknown response surface as

$$Y(\mathbf{x}) = \mathbf{f}(\mathbf{x})^{\intercal}\boldsymbol{\beta} + M(\mathbf{x}) \tag{1.1}$$

where $\mathbf{f}(\mathbf{x}) \subseteq \mathbb{R}^p$ is a known vector of user specified basis functions, $\boldsymbol{\beta} \subseteq \mathbb{R}^p$ is an unknown parameter vector that needs to be estimated. This form of model is called "universal kriging".

In (1.1), M is a realization of a zero mean second-order stationary Gaussian random field. A random field is a generalization of a stochastic process such that the underlying parameter need no longer be a simple real or integer valued "time", but can instead take values that are multidimensional vectors, or points on some manifold, see [15]. In general, the random field can be thought of as a "function valued" random variable. The values in a random field are usually spatially correlated in some way. It is often assumed that values with adjacent design points do not differ as much as values that are further apart. In other words, the values $M(\mathbf{x}_1)$ and $M(\mathbf{x}_2)$ will be similar if $\mathbf{x}_1$ and $\mathbf{x}_2$ are close to each other in $\mathbb{R}^d$. A Gaussian random field (GRF) is a random field involving Gaussian probability density functions of the variables. In particular, a one-dimensional GRF is also called a Gaussian process.

Thus, the response $Y(\mathbf{x})$ is modeled using a trend term $\mathbf{f}(\mathbf{x})^\intercal \boldsymbol{\beta}$ representing the mean response value and a noise term $M(\mathbf{x})$ quantifying our uncertainty about the unknown true response at $\mathbf{x}$. This kind of uncertainty is referred to as extrinsic uncertainty [13]. With this type of metamodel, many statistical concepts like mean square error (MSE) of prediction estimation can be analyzed rigorously.

Suppose that we have selected a set of design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ and run the simulation experiments on these points. A set of deterministic response $\mathbf{y}_d = (y(\mathbf{x}_1), y(\mathbf{x}_2), \ldots, y(\mathbf{x}_k))^\intercal$ is observed. To provided more technical details about the model, 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 $\Sigma_M(\mathbf{x}, \cdot) = (\mathrm{Cov}[M(\mathbf{x}), M(\mathbf{x}_1)], \ldots, \mathrm{Cov}[M(\mathbf{x}), M(\mathbf{x}_k)])^\intercal$ represent the spatial covariances between (un-explored point) $\mathbf{x}$ and all design points. 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. It is assumed that Σ_M and $\Sigma_M(\mathbf{x}, \cdot)$ are completely known in the following analysis. However, it is almost impossible to get such information in practice. Thus, they have to be estimated statistically. This issue will be discussed later.

Consider the linear predictor

$$\begin{aligned}\hat{y}_d(\mathbf{x}) &= \sum_{i=1}^{k} \lambda_i(\mathbf{x}) \cdot y(\mathbf{x}_i) \\ &= \lambda^\intercal \cdot \mathbf{y}_d \end{aligned} \tag{1.2}$$

of $f(\mathbf{x})$ at a unexplored point $\mathbf{x}$, where $\lambda = (\lambda_1(\mathbf{x}), \lambda_2(\mathbf{x}), \ldots, \lambda_k(\mathbf{x}))^\intercal$ is a weight vector. The predictor $\hat{y}_d(\mathbf{x})$ is a weighted linear combination of observed values. We can replace $\mathbf{y}_d$ by the corresponding random quantity $\mathbf{Y}_d = (Y(\mathbf{x}_1), Y(\mathbf{x}_2), \ldots, Y(\mathbf{x}_k))^\intercal$ and then compute the MSE of the estimated predictor. By minimizing

$$MSE(\hat{y}_d(\mathbf{x})) = E[\lambda^\intercal \cdot \mathbf{Y}_d - Y(\mathbf{x})]^2, \tag{1.3}$$

we can get the best linear unbiased predictor (BLUP)

$$\hat{y}_d(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\intercal \hat{\boldsymbol{\beta}} + \Sigma_M(\mathbf{x}, \cdot)^\intercal \Sigma_M^{-1}(\mathbf{y}_d - \mathbf{F}\hat{\boldsymbol{\beta}}), \tag{1.4}$$

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

$$MSE(\hat{y}_d(\mathbf{x})) = \Sigma_M(\mathbf{x}, \mathbf{x}) - \Sigma_M(\mathbf{x}, \cdot)^\intercal \Sigma_M^{-1} \Sigma_M(\mathbf{x}, \cdot) + \eta^\intercal \left(\mathbf{F}^\intercal \Sigma_M^{-1} \mathbf{F}\right)^{-1} \eta, \tag{1.5}$$

where $\eta = \mathbf{f}(\mathbf{x}) - \mathbf{F}^\intercal \Sigma_M^{-1} \Sigma_M(\mathbf{x}, \cdot)$. See [9],[13],[14] and [16] for more details.

1.2.3 Simple Kriging

Simple kriging can be viewed as a special case of universal kriging. In particular, if the trend term $\mathbf{f}(\mathbf{x})^\intercal \boldsymbol{\beta}$ is simplified as a known constant β, then the model (1.1) is called simple kriging. The linear predictor and its corresponding MSE estimator can be derived as (see [17] for more details)

$$\begin{aligned} \hat{y}_d(\mathbf{x}) &= \beta + \Sigma_M(\mathbf{x}, \cdot)^\intercal \Sigma_M^{-1}(\mathbf{y}_d - \beta) \\ MSE(\hat{y}_d(\mathbf{x})) &= \Sigma_M(\mathbf{x}, \mathbf{x}) - \Sigma_M(\mathbf{x}, \cdot)^\intercal \Sigma_M^{-1} \Sigma_M(\mathbf{x}, \cdot). \end{aligned} \tag{1.6}$$

Although the model is greatly simplified, simple kriging is not widely used because obtaining the constant trend information is intractable in most cases. On the other hand, specifying basis functions in advance and applying universal kriging require extensive computation effort and prior knowledge. Ordinary kriging helps to overcome these difficulties and has been widely used in real life problems.

1.2.4 Ordinary Kriging

Similar to simple kriging, ordinary kriging also assumes that the trend term in (1.1) is replaced by a constant β. However, the constant β in ordinary kriging is unknown and thus needs to be estimated. Based on the similar idea as introduced in universal kriging, it is not difficult to derive the linear predictor and its corresponding MSE estimator as follows (see [8], [9], [17] and [18]).

$$\begin{aligned} \hat{y}_d(\mathbf{x}) &= \hat{\beta} + \Sigma_M(\mathbf{x},\cdot)^\intercal \Sigma_M^{-1}(\mathbf{y}_d - \hat{\beta}), \\ MSE(\hat{y}_d(\mathbf{x})) &= \Sigma_M(\mathbf{x},\mathbf{x}) - \Sigma_M(\mathbf{x},\cdot)^\intercal \Sigma_M^{-1}\Sigma_M(\mathbf{x},\cdot) + \eta^\intercal \left(\mathbf{1}_k^\intercal \Sigma_M^{-1}\mathbf{1}_k\right)^{-1}\eta, \end{aligned} \tag{1.7}$$

where $\hat{\beta} = (\mathbf{1}_k^\intercal \Sigma_M^{-1}\mathbf{1}_k)^{-1}\mathbf{1}_k^\intercal \Sigma_M^{-1}\mathbf{y}_d$, and $\eta = 1 - \mathbf{1}_k^\intercal \Sigma_M^{-1}\Sigma_M(\mathbf{x},\cdot)$.

By comparing (1.4), (1.5) with (1.7), it is not difficult to observe that the basis function in the trend term can be simplified to $\mathbf{f}(\mathbf{x})^\intercal\boldsymbol{\beta} = \beta$ (i.e., $p = 1$, $\mathbf{f}(\mathbf{x}) = 1\ \forall\mathbf{x}$ and $\mathbf{F} = \mathbf{1}_k$). Please see [9],[13],[14],[16] and [18] for more details.

In general, ordinary kriging is the most commonly used method among the kriging family. However, it is limited to deterministic settings. A more general stochastic kriging (SK) method was proposed by Ankenman et al. [13] to handle the situation when the observed response is random.

1.3 Stochastic Kriging

In stochastic settings, it is assumed that the true response $f(\mathbf{x})$ at each point $\mathbf{x}$ cannot be observed exactly but can be estimated in a path-wise manner through stochastic simulations.

Suppose that there is 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 $\bar{y}(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} y_j(\mathbf{x}_i)$ and $y_j(\mathbf{x}_i)$ is the simulation output at $\mathbf{x}_i$ obtained on the jth replication run. Stochastic kriging assumes $y_j(\mathbf{x}_i)$ to be of the following form:

$$
\begin{aligned}
y_j(\mathbf{x}_i) &= \mathbf{f}(\mathbf{x}_i)^\intercal \boldsymbol{\beta} + M(\mathbf{x}_i) + \epsilon_j(\mathbf{x}_i) \\
&= Y(\mathbf{x}_i) + \epsilon_j(\mathbf{x}_i),
\end{aligned} \tag{1.8}
$$

where $\mathbf{f}(\mathbf{x}_i) \subseteq \mathbb{R}^p$, $\boldsymbol{\beta} \subseteq \mathbb{R}^p$ and M are defined exactly the same as in universal kriging. Different from universal kriging, an additional term $\epsilon_j(\mathbf{x}_i)$ is used in Equation (1.8). It is interpreted as the intrinsic noise and is primarily used in stochastic kriging to model the simulation noise in the jth replication run at $\mathbf{x}_i$.

To provide the predictor and MSE estimator in stochastic settings, we assume that Σ_M, $\Sigma_M(\mathbf{x}, \cdot)$ and $\mathbf{F}$ are defined the same as that in universal kriging. Let Σ_ϵ be the $k \times k$ covariance matrix associated with the intrinsic simulation noise with $(i,j)th$ element $\text{Cov}[\bar{\epsilon}(\mathbf{x}_i), \bar{\epsilon}(\mathbf{x}_j)]$, where $\bar{\epsilon}(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} \epsilon_j(\mathbf{x}_i)$ for all $i = 1, \ldots, k$.

Under the above notation, it has been shown in [13] that when $\boldsymbol{\beta}, \Sigma_M(\mathbf{x}, \cdot)$ and Σ_M are known, the MSE-optimal predictor is of the form

$$
\hat{y}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\intercal \boldsymbol{\beta} + \Sigma_M(\mathbf{x}, \cdot)^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} (\bar{\mathbf{y}} - \mathbf{F}\boldsymbol{\beta}) \tag{1.9}
$$

and the corresponding optimal MSE is given by

$$MSE(\hat{y}(\mathbf{x})) = \Sigma_M(\mathbf{x},\mathbf{x}) - \Sigma_M(\mathbf{x},\cdot)^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_M(\mathbf{x},\cdot). \tag{1.10}$$

On the other hand, when $\Sigma_M(\mathbf{x},\cdot)$ and Σ_M are known, but $\boldsymbol{\beta}$ is estimated via the generalized least squares estimator, the MSE-optimal predictor becomes (see, e.g., [19])

$$\hat{y}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\intercal\hat{\boldsymbol{\beta}} + \Sigma_M(\mathbf{x},\cdot)^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}(\bar{\mathbf{y}} - \mathbf{F}\hat{\boldsymbol{\beta}}), \tag{1.11}$$

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

$$MSE(\hat{y}(\mathbf{x})) = \Sigma_M(\mathbf{x},\mathbf{x}) - \Sigma_M(\mathbf{x},\cdot)^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_M(\mathbf{x},\cdot) + \eta^\intercal\left(\mathbf{F}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{F}\right)^{-1}\eta, \tag{1.12}$$

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

Throughout the thesis, we make the following assumption on the intrinsic and extrinsic noises in stochastic kriging metamodel.

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. $N(0, V(\mathbf{x}_i))$, independent of $\epsilon_j(\mathbf{x}_h)$ for all j and $h \neq i$, and inde-*

pendent of M.

Assumption 1 indicates 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 θ 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 $\mathbf{x} \in \mathcal{X}$.

1.4 Stochastic Kriging with Gradient Information

Based on the SK model, suppose in addition to simulation outputs $y_j(\mathbf{x}_i)$, we can also obtain their unbiased gradient estimates $\mathcal{D}_j(\mathbf{x}_i) = \big(\mathcal{D}_j^1(\mathbf{x}_i), \ldots, \mathcal{D}_j^d(\mathbf{x}_i)\big)^\intercal$ at $\mathbf{x}_i$ on the jth replication run. For such a setting, [14] has introduced an augmented kriging model called stochastic kriging with gradient estimators that explicitly incorporates gradient estimators in constructing an SK predictor. Specifically, each (partial) gradient estimator $\mathcal{D}_j^r(\mathbf{x}_i)$, $r = 1, \ldots, d$, is assumed to take the form

$$\mathcal{D}_j^r(\mathbf{x}_i) = \frac{\partial Y(\mathbf{x}_i)}{\partial x_r} + \zeta_j^r(\mathbf{x}_i) = \left(\frac{\partial \mathbf{f}(\mathbf{x}_i)}{\partial x_r}\right)^{\intercal} \boldsymbol{\beta} + \frac{\partial M(\mathbf{x}_i)}{\partial x_r} + \zeta_j^r(\mathbf{x}_i) = D^r(\mathbf{x}_i) + \zeta_j^r(\mathbf{x}_i), \tag{1.13}$$

where $\zeta_j^r(\mathbf{x}_i)$, $r = 1, \ldots, d$ are the simulation noises associated with gradient estimators at $\mathbf{x}_i$. The following assumption of ζ_j^r is made through the thesis.

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

Under the 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 $\zeta_j^r(\mathbf{x}_i)$ at the same design point $\mathbf{x}_i$ within the same replication j.

When gradient information is available, the notation used in SK should be augmented to the SKG version by incorporating the derivative information. We use the "+" symbol to distinguish any quantity that is related to

SKG. In particular, the augmented response vector $\bar{\mathbf{y}}^+$ is written as

$$\begin{aligned}\bar{\mathbf{y}}^+ &= (\bar{y}(\mathbf{x}_1), \ldots, \bar{y}(\mathbf{x}_k))^\intercal + (\bar{\mathcal{D}}^1(\mathbf{x}_1), \ldots, \bar{\mathcal{D}}^1(\mathbf{x}_k), \ldots, \bar{\mathcal{D}}^d(\mathbf{x}_1), \ldots, \bar{\mathcal{D}}^d(\mathbf{x}_k)\big)^\intercal \\ &= (\bar{\mathbf{y}}^\intercal, \bar{\mathcal{D}}^\intercal)^\intercal \end{aligned} \tag{1.14}$$

where $\bar{\mathcal{D}}^r(\mathbf{x}_i) = \sum_{j=1}^{n_i} \mathcal{D}_j^r(\mathbf{x}_i), r = 1, \ldots, d$ and $j = 1, \ldots, n_i$. In addition, the augmented spatial covariance matrix Σ_M^+ can be written explicitly as

$$\Sigma_M^+ = \begin{pmatrix} C_{0,0}^M(1,1) & \cdots & C_{0,0}^M(1,k) & \cdots & C_{0,d}^M(1,1) & \cdots & C_{0,d}^M(1,k) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ C_{0,0}^M(k,1) & \cdots & C_{0,0}^M(k,k) & \cdots & C_{0,d}^M(k,1) & \cdots & C_{0,d}^M(k,k) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ C_{d,0}^M(1,1) & \cdots & C_{d,0}^M(1,k) & \cdots & C_{d,d}^M(1,1) & \cdots & C_{d,d}^M(1,k) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ C_{d,0}^M(k,1) & \cdots & C_{d,0}^M(k,k) & \cdots & C_{d,d}^M(k,1) & \cdots & C_{d,d}^M(k,k) \end{pmatrix}$$

and

$$\begin{aligned} &C_{0,0}^M(i,h) = Cov(Y(\mathbf{x}_i), Y(\mathbf{x}_h)), C_{0,r}^M(i,h) = Cov(Y(\mathbf{x}_i), D^r(\mathbf{x}_h)) \\ &C_{r,0}^M(i,h) = Cov(D^r(\mathbf{x}_i), Y(\mathbf{x}_h)), C_{r,g}^M(i,h) = Cov(D^r(\mathbf{x}_i), D^g(\mathbf{x}_h)) \\ &C_{r,r}^M(i,h) = Cov(D^r(\mathbf{x}_i), D^r(\mathbf{x}_h)) \end{aligned} \tag{1.15}$$

where $r, g = 1, \ldots, d$ and $i, h = 1, \ldots, k$. Moreover, let

$$\begin{aligned} Var[\epsilon_j(\mathbf{x}_i)] = \sigma_{i0}^2, Var[\zeta_j^r(\mathbf{x}_i)] = \sigma_{ir}^2 \\ Corr[\epsilon_j(\mathbf{x}_i), \zeta_j^r(\mathbf{x}_i)] = \varrho_i^{(0,r)}, Corr[\zeta_j^r(\mathbf{x}_i), \zeta_j^s(\mathbf{x}_i)] = \varrho_i^{(r,s)} \end{aligned} \tag{1.16}$$

where $r, s = 1, \ldots, d$ and $r \neq s$. So the augmented intrinsic noise covariance matrix Σ_ϵ^+ is changed to a $k(d+1) \times k(d+1)$ matrix with elements specified as

$$\begin{aligned} &\Sigma_\epsilon^+[rk + i, sk + i] = \frac{\varrho_i^{(r,s)} \sigma_{ir} \sigma_{is}}{n_i} \quad r, s = 0, \ldots, d, r \neq s, i = 1, \ldots, k \\ &\Sigma_\epsilon^+[rk + l, sk + h] = 0 \quad l \neq h \end{aligned} \tag{1.17}$$

Finally, the augmented covariance vector between a given point $\mathbf{x}$ and all design points is given as

$$\begin{aligned} \Sigma_M^+(\mathbf{x}, \cdot) &= (\Sigma_M(\mathbf{x}, \cdot)^\intercal, \Sigma_{M,d}(\mathbf{x}, \cdot)^\intercal)^\intercal \\ &= (\mathrm{Cov}[Y(\mathbf{x}), Y(\mathbf{x}_1)], \ldots, \mathrm{Cov}[Y(\mathbf{x}), Y(\mathbf{x}_k)])^\intercal + (\mathrm{Cov}[Y(\mathbf{x}), D^1(\mathbf{x}_1)], \ldots, \\ &\quad \mathrm{Cov}[Y(\mathbf{x}), D^1(\mathbf{x}_k)], \ldots, \mathrm{Cov}[Y(\mathbf{x}), D^d(\mathbf{x}_1)], \ldots, \mathrm{Cov}[Y(\mathbf{x}), D^d(\mathbf{x}_k)])^\intercal \end{aligned} \tag{1.18}$$

and let the augmented vector of basis functions be $\mathbf{F}^+ = (\mathbf{F}^\intercal, \mathbf{F}_d^\intercal)^\intercal$, where $\mathbf{F}_d = \left(\frac{\partial \mathbf{f}(\mathbf{x}_1)}{\partial x_{11}}, \ldots, \frac{\partial \mathbf{f}(\mathbf{x}_k)}{\partial x_{k1}}, \ldots, \frac{\partial \mathbf{f}(\mathbf{x}_1)}{\partial x_{1d}}, \ldots, \frac{\partial \mathbf{f}(\mathbf{x}_k)}{\partial x_{kd}}\right)^\intercal$.

It has been shown in [14] that when $\boldsymbol{\beta}$ is known, the MSE-optimal predictor is given as

$$\hat{y}^+(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\intercal \boldsymbol{\beta} + \Sigma_M^+(\mathbf{x}, \cdot)^\intercal (\Sigma_M^+ + \Sigma_\epsilon^+)^{-1} (\bar{\mathbf{y}} - \mathbf{F}^+ \boldsymbol{\beta}) \tag{1.19}$$

and the corresponding optimal MSE is given by

$$MSE(\hat{y}^+(\mathbf{x})) = \Sigma_M(\mathbf{x}, \mathbf{x}) - \Sigma_M^+(\mathbf{x}, \cdot)^\intercal (\Sigma_M^+ + \Sigma_\epsilon^+)^{-1} \Sigma_M^+(\mathbf{x}, \cdot). \tag{1.20}$$

On the other hand, when $\Sigma_M^+(\mathbf{x}, \cdot)$ and Σ_M^+ are known, but $\boldsymbol{\beta}$ is estimated via the generalized least squares estimator, the MSE-optimal predictor becomes (see, e.g., [14] and [19])

$$\hat{y}^+(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\intercal \hat{\boldsymbol{\beta}} + \Sigma_M^+(\mathbf{x}, \cdot)^\intercal (\Sigma_M^+ + \Sigma_\epsilon^+)^{-1} (\bar{\mathbf{y}}^+ - \mathbf{F}^+ \hat{\boldsymbol{\beta}}), \tag{1.21}$$

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

$$MSE(\hat{y}^+(\mathbf{x})) = \Sigma_M(\mathbf{x},\mathbf{x}) - \Sigma_M^+(\mathbf{x},\cdot)^\intercal(\Sigma_M^+ + \Sigma_\epsilon^+)^{-1}\Sigma_M^+(\mathbf{x},\cdot) + \eta^{+\intercal}\left(\mathbf{F}^{+\intercal}(\Sigma_M^+ + \Sigma_\epsilon^+)^{-1}\mathbf{F}^+\right)^{-1}\eta^+, \tag{1.22}$$

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

To estimate the gradient estimator of simulation outputs, many methods such as infinitesimal perturbation analysis (IPA), the likelihood ratio/score function, finite difference gradient estimation and simultaneous perturbation gradient estimation ([20], [21], [22], [23] and [24]) can be applied; see, e.g., Chen et al. [14].

1.5 Parameter Estimation in SK and SKG

To apply SK/SKG metamodels in practice, it is crucial to obtain accurate estimates of the model parameters. In SK/SKG metamodels, two types of parameters need to be estimated: the parameters in the trend term, like $\boldsymbol{\beta}$; and the parameters in the kriging correlation model, like $\tau^2, \boldsymbol{\theta}$. There are multiple choices of the correlation model as long as they satisfy Assumption 1. In practice, a common choice of the correlation model is $R_M(\mathbf{x}_i, \mathbf{x}_j) = exp\{-\sum_{r=1}^d \theta_r(x_{ir} - x_{jr})^2\}$. A constant trend model $\mathbf{f}(\mathbf{x})^\intercal\boldsymbol{\beta} = \beta$ (i.e., $p = 1$,

$\mathbf{f}(\mathbf{x}) = 1\ \forall \mathbf{x}$ and $\mathbf{F} = \mathbf{1}_k$) is also suggested [13]. These selections of correlation model and trend term will be used in Sections 3 and 4.

To estimate the parameters in SK/SKG metamodels, a common approach is to search for the maximum likelihood estimator. In particular, the log-likelihood function for the SK metamodel is derived in ([13]) as follows.

$$l(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}) = -ln((2\pi)^{\frac{k}{2}}) - \frac{1}{2}ln[\|\Sigma_M + \Sigma_\epsilon\|] - \frac{1}{2}(\bar{\mathbf{y}} - \beta \mathbf{1}_k)^\intercal [\Sigma_M + \Sigma_\epsilon]^{-1}(\bar{\mathbf{y}} - \beta \mathbf{1}_k) \tag{1.23}$$

The log-likelihood function for the SKG metamodel has almost the same form of (1.24), except that $\Sigma_M, \Sigma_\epsilon, \bar{\mathbf{y}}$ are needed to be replaced by $\Sigma_M^+, \Sigma_\epsilon^+, \bar{\mathbf{y}}^+$ ([14]).

$$\begin{aligned} l(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}) = &- ln((2\pi)^{\frac{k}{2}}) - \frac{1}{2}ln[\|\Sigma_M^+ + \Sigma_\epsilon^+\|] - \\ &\frac{1}{2}(\bar{\mathbf{y}}^+ - \beta \mathbf{1}_{k(d+1)})^\intercal [\Sigma_M^+ + \Sigma_\epsilon^+]^{-1}(\bar{\mathbf{y}}^+ - \beta \mathbf{1}_{k(d+1)}) \end{aligned} \tag{1.24}$$

A number of numerical methods are recommended in [25] to search for the MLEs $(\hat{\beta}, \hat{\tau}^2, \hat{\boldsymbol{\theta}})$. In general, the procedure of fitting a SK/SKG metamodel is as follows:

1. Decide a set of design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ and obtain the simulation

outputs on each design point.

2. Estimate the variance of intrinsic noise $\hat{V}(\mathbf{x})$ and then obtain $\hat{\Sigma}_\epsilon$.

3. Use the $\hat{\Sigma}_\epsilon$ instead of Σ_ϵ, and maximize the log-likelihood function (1.24) or (1.23).

4. Plug $\hat{\beta}, \hat{\tau}^2, \hat{\boldsymbol{\theta}}$ into the corresponding predictors.

More details about the estimation of intrinsic noise are available in [13] and [14]. When choosing the metamodel parameters, several issues, as discussed below, should be addressed ([13], [26] and [27]).

First, it may be helpful to normalize the design points in some special situations, especially when different components have different variabilities. The design points should be normalized by transforming each component separately, like mapping x_{ih} to $\frac{x_{ih} - min_{j=1,\ldots,k} x_{jh}}{max_{j=1,\ldots,k} x_{jh} - min_{j=1,\ldots,k} x_{jh}}$. This kind of affine transformation merely changes the scale of each component of $\boldsymbol{\theta}$. The normalization trick can be neglected if $\boldsymbol{\theta}$ is set as a scalar.

Second, when searching for the MLEs $(\hat{\beta}, \hat{\tau}^2, \hat{\boldsymbol{\theta}})$, many nonlinear solvers require a starting point. Although there is not a strict rule about how to choose a starting point, it is suggested to start with a moderate value for $(\hat{\beta}, \hat{\tau}^2, \hat{\boldsymbol{\theta}})$. For example, initialize β, τ^2 to be the sample mean and sample variance of the output measures $\{\bar{y}(\mathbf{x}_1), \ldots, \bar{y}(\mathbf{x}_k)\}$ and set the initial r_{th} component of $\boldsymbol{\theta}$, θ_r as the value that solves the equation $0.5 = (R_M(\bar{d}, \theta_r))^p$, where $\bar{d}$ denotes the average distance among all design points, such as $\bar{d} = \frac{2\sum_{i=1}^{k}\sum_{j=i+1}^{k}\|x_ir - x_jr\|}{k(k-1)}$.

Another issue is how to avoid numerical instability when optimizing the log-likelihood function. Numerical instabilities may happen when some rows of $\Sigma_M + \Sigma_\epsilon$ are nearly zero or collinear. These situations can be avoided by special experimental designs. One solution is to avoid selecting repeated design points. Another solution is to add constraints such as $\sum_{r=1}^{p}\theta_r > v_0$ for a small v_0. Besides, it is also important to include the constraint $\tau^2 > 0$ in the optimization. Many spatial correlation models also require the constraint $\boldsymbol{\theta} > 0$ as well.

Under the stochastic kriging framework, it has been shown in [27] that the estimator $\hat{\boldsymbol{\theta}}$ can be affected by the intrinsic noises. When the variance of both intrinsic and extrinsic noise are known, the additional prediction error caused by the uncertainty of parameter estimation increases as the variance of intrinsic noise increases. Therefore, it is important to locate sufficient simulation efforts on each design point in order to obtain a well-built metamodel.

1.6 Applications of Kriging Related Methods

As we have introduced in previous sections, kriging-related methods present several interesting properties that make them popular in various disciplines. First of all, the simple/ordinary/universal kriging is an exact interpolation method. That is, the points of the design of experiments are interpolated. This makes kriging very popular in mechanical design of experiments. Besides, the uncertainty on the prediction (the MSE of the predictor) provided by kriging can be used to identify the least reliable prediction among a population. As a result, developing sequential learning methods based on this characteristic is possible because the metamodel can be refined step by step. Stochastic kriging extends the area of application by accounting for both intrinsic and extrinsic noises. Many non-deterministic models can be approximated and replaced by SK metamodels. Recently, the theory of SK metamodel is further developed by incorporating the gradient information. The work of Chen et.al in [14] indicates that adding gradient information may lead to more accurate prediction performance under some ideal assumptions. With well-developed theories on gradient estimation, SKG may become popular in certain situations.

In general, the kriging family are mainly applied for two different purposes: prediction and optimization. A brief review of applications on each purpose is provided as follows.

1.6.1 Application in Prediction

The most intuitive and important application of kriging-related models is the approximation of a specific model. Since kriging is originally proposed in geostatistics, it has been widely used in geostatistical spatial analysis and environmental research. Its popularity in geostatistics depends on its ability of successfully expressing statistical relationships among spatially distributed data. So kriging is a good choice to solve many practical problems, for example, the mapping and control of soil variation [4], [5] and [28] the air pollution exposure assessment [7] and the regional ground water investigation [6].

Besides the popularity in geostatistics and environmental science, kriging is also a productive tool for performance prediction in finance and computer science. For example, SK metamodel can be employed to estimate the risks and their sensitivities in asset portfolios because it can avoid time consuming calculations of risk measures, see Chen et.al in [29] and Liu et.al in [30]. In addition, Baysal et.al designed a method for simulating the hedging and trading strategy based on SK metamodel [31]. Another example is the pricing of securities with SK metamodels [32].

1.6.2 Application in Simulation Optimization

Optimizing a simulated system via kriging related methods is a promising and fast-developing research topic in the recent decade, because of its ability to accurately approximate a response surface [33], [34]. One of the most popular algorithms is the efficient global optimization (EGO) algorithm proposed by Jones et.al in [11]. In this algorithm, kriging is used to model the nonlinear and multimodal response surface. The key of using kriging in global optimization lies in balancing the need of exploiting the approximated surface with the need of improving the quality of approximation. Huang et.al [35] extended EGO to address stochastic black-box systems by proposing an augmented expected improvement function as the new sample selection criterion. A sequential kriging optimization method based on this criterion is also developed [12].

2 Monotonicity Property of SK and SKG Predictors

2.1 Motivation

Obtaining accurate SK or SKG predictors requires careful selection of design points. A commonly used method in practice is to generate design points all at once according to predefined space-filling design. However, it has been shown in [9] and [36] that fitting the metamodel in a sequential manner, which chooses one design point each time, has the advantage of adaptively and sequentially updating the metamodel. As a result, a new design point can be selected based on the updated metamodel and the information carried by previous design points. A lot of sequential sampling approaches have been introduced and found demonstrating better performance than space-filling design in deterministic kriging models. However, it is suspectable that these approaches can be applied with SK/SKG metamodel because of the intrinsic noise embedded in SK/SKG models. More fundamentally, it is not clear whether the information carried by new design points will result in improvement of the predictive performance.

In this section, we focus on investigating the performance of SK/SKG predictors in a sequential way when only one design point is added at a time. Our main results show that when the SK/SKG parameters are known or estimated, the MSE of SK/SKG predictor is monotonically non-increasing as the number of design points increases. In particular, we consider both cases when the trend terms are known or estimated and provide the reduction of the MSE estimators. It is interesting to see whether the prediction performance of SK/SKG metamodel can always be improved by adding new design points. Besides, we would like to compare the SK and SKG metamodels. We want to see if it will be helpful to incorporate the gradient information in SK. We also want to check how to quantify the benefaction of doing so. All of the problems mentioned above will be investigated in this section.

2.2 Monotonic Performance of SK Predictors

In this section, we use the subscript k to denote 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 $\{\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{x}_{k+1}\}$.

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

Lemma 2.1. *If Assumption 1 holds, then the matrix* $(\Sigma_{M_k}+\Sigma_{\epsilon_k})^{-1}$ *is positive definite for all* k*.*

Proof of Theorem 2.1. Since Σ_{M_k} is the covariance matrix of the unknown responses $Y(\mathbf{x}_1),\ldots,Y(\mathbf{x}_k)$, Assumption 1 implies that it is positive definite. On the other hand, the covariance matrix Σ_{ϵ_k} associated with the intrinsic noise is positive semi-definite. Thus, $\Sigma_{M_k}+\Sigma_{\epsilon_k}$ is positive definite. This shows that $(\Sigma_{M_k}+\Sigma_{\epsilon_k})^{-1}$ is also positive definite. □

Let $\mathbf{x}_0$ be a prediction point, $\hat{y}_k(\mathbf{x}_0)$ be the SK predictor constructed using Equation (1.9) based on a set of k design points $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_k\}$, 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. 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 2.1. *Suppose that* $\mathbf{x}_{k+1}\notin\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}$*. For any prediction point* $\mathbf{x}_0\in\mathcal{X}$*, let* $MSE(\hat{y}_k(\mathbf{x}_0))$ *and* $MSE(\hat{y}_{k+1}(\mathbf{x}_0))$ *denote the MSEs of the predictors* $\hat{y}_k(\mathbf{x}_0)$ *and* $\hat{y}_{k+1}(\mathbf{x}_0)$ *constructed using Equation (1.9). If Assumption 1 holds, then* $MSE(\hat{y}_k(\mathbf{x}_0))\geq MSE(\hat{y}_{k+1}(\mathbf{x}_0))$*.*

Proof of Theorem 2.1. We shall prove the result when the optimal MSE is given by (1.10). In particular, the MSE of $\hat{y}_k$ can be written as $MSE(\hat{y}_k(\mathbf{x}_0))=$

$\tau^2 - \Sigma_{M_k}(\mathbf{x}_0,\cdot)^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0,\cdot)$. After a new design point $\mathbf{x}_{k+1}$ is included in the model, the covariance vector between the prediction point $\mathbf{x}_0$ and all $k+1$ design points can be expressed in terms of $\Sigma_{M_k}(\mathbf{x}_0,\cdot)$ as $\Sigma_{M_{k+1}}(\mathbf{x}_0,\cdot) = (\Sigma_{M_k}(\mathbf{x}_0,\cdot)^\intercal, \Sigma_M(\mathbf{x}_0,\mathbf{x}_{k+1}))^\intercal$, where $\Sigma_M(\mathbf{x}_0,\mathbf{x}_{k+1}) =$ $\mathrm{Cov}(Y(\mathbf{x}_0), Y(\mathbf{x}_{k+1}))$. Similarly, it is not difficult to verify that the sum of the covariance matrices $\Sigma_{M_{k+1}} + \Sigma_{\epsilon_{k+1}}$ can be written in the form

$$\Sigma_{M_{k+1}} + \Sigma_{\epsilon_{k+1}} = \begin{pmatrix} \Sigma_{M_k} + \Sigma_{\epsilon_k} & \Sigma_{k\times 1} \\ \Sigma_{k\times 1}^\intercal & \tau^2 + \Sigma_\epsilon(\mathbf{x}_{k+1},\mathbf{x}_{k+1}) \end{pmatrix},$$

where $\Sigma_{k\times 1}$ is a $k \times 1$ matrix with its ith element given by $\Sigma_M(\mathbf{x}_i,\mathbf{x}_{k+1})$ and $\Sigma_\epsilon(\mathbf{x}_{k+1},\mathbf{x}_{k+1}) = \mathrm{Cov}[\bar{\epsilon}(\mathbf{x}_{k+1}), \bar{\epsilon}(\mathbf{x}_{k+1})]$. By Lemma 2.1, $\Sigma_{M_{k+1}} + \Sigma_{\epsilon_{k+1}}$ is positive definite and thus invertible. Its inverse, denoted by A, can be calculated using the block matrix inversion formula as follows:

$$A = \begin{pmatrix} A_{11} & -(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{k\times 1}\Phi \\ -\Phi\Sigma_{k\times 1}^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1} & \Phi \end{pmatrix},$$

where

$$\begin{aligned}
A_{11} &= (\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1} + (\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{k\times 1}\Phi\Sigma_{k\times 1}^{\intercal}(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1} \\
\Phi &= \Big(\tau^2 - \Sigma_{k\times 1}^{\intercal}(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{k\times 1} + \Sigma_{\epsilon}(\mathbf{x}_{k+1}, \mathbf{x}_{k+1})\Big)^{-1} \\
&= \Big(MSE(\hat{y}_k(\mathbf{x}_{k+1})) + \Sigma_{\epsilon}(\mathbf{x}_{k+1}, \mathbf{x}_{k+1})\Big)^{-1}.
\end{aligned}$$

Thus, it follows that

$$\begin{aligned}
&MSE(\hat{y}_{k+1}(\mathbf{x}_0)) = \tau^2 - \Sigma_{M_{k+1}}(\mathbf{x}_0, \cdot)^{\intercal}(\Sigma_{M_{k+1}} + \Sigma_{\epsilon_{k+1}})^{-1}\Sigma_{M_{k+1}}(\mathbf{x}_0, \cdot) \\
&= \tau^2 - (\Sigma_{M_k}(\mathbf{x}_0, \cdot)^{\intercal}, \Sigma_M(\mathbf{x}_0, \mathbf{x}_{k+1})) \\
&\qquad\qquad \begin{pmatrix} \Sigma_{M_k} + \Sigma_{\epsilon_k} & \Sigma_{k\times 1} \\ \Sigma_{k\times 1}^{\intercal} & \tau^2 + \Sigma_{\epsilon}(\mathbf{x}_{k+1}, \mathbf{x}_{k+1}) \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{M_k}(\mathbf{x}_0, \cdot) \\ \Sigma_M(\mathbf{x}_0, \mathbf{x}_{k+1}) \end{pmatrix} \\
&= \tau^2 - (\Sigma_{M_k}(\mathbf{x}_0, \cdot)^{\intercal}, \Sigma_M(\mathbf{x}_0, \mathbf{x}_{k+1}))A \begin{pmatrix} \Sigma_{M_k}(\mathbf{x}_0, \cdot) \\ \Sigma_M(\mathbf{x}_0, \mathbf{x}_{k+1}) \end{pmatrix} \\
&= \tau^2 - \Big[\Sigma_{M_k}(\mathbf{x}_0, \cdot)^{\intercal}(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0, \cdot) \\
&\quad + \Sigma_{M_k}(\mathbf{x}_0, \cdot)^{\intercal}(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{k\times 1}\Phi\Sigma_{k\times 1}^{\intercal}(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0, \cdot) \\
&\quad - \Sigma_M(\mathbf{x}_0, \mathbf{x}_{k+1})\Phi\Sigma_{k\times 1}^{\intercal}(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0, \cdot) \\
&\quad - \Sigma_{M_k}(\mathbf{x}_0, \cdot)^{\intercal}(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{k\times 1}\Phi\Sigma_M(\mathbf{x}_0, \mathbf{x}_{k+1}) + \Sigma_M(\mathbf{x}_0, \mathbf{x}_{k+1})\Phi\Sigma_M(\mathbf{x}_0, \mathbf{x}_{k+1})\Big]
\end{aligned}$$

$$= \tau^2 - \Sigma_{M_k}(\mathbf{x}_0, \cdot)^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0, \cdot) - \Big(\Sigma_{M_k}(\mathbf{x}_0, \cdot)^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{k\times 1}$$

$$- \Sigma_M(\mathbf{x}_0, \mathbf{x}_{k+1})^\intercal\Big)\Phi\Big(\Sigma_{k\times 1}^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0, \cdot) - \Sigma_M(\mathbf{x}_0, \mathbf{x}_{k+1})\Big)$$

$$= MSE(\hat{y}_k(\mathbf{x}_0)) - \phi(\mathbf{x}_0)^2\Phi,$$

where we have defined $\phi(\mathbf{x}_0) = \Sigma_{k\times 1}^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0, \cdot) - \Sigma_M(\mathbf{x}_0, \mathbf{x}_{k+1})$. Finally, since $\phi(\mathbf{x}_0)^2 \geq 0$ and Φ is a positive scalar, we have $MSE(\hat{y}_k(\mathbf{x}_0)) \geq MSE(\hat{y}_{k+1}(\mathbf{x}_0))$. □

The next result shows that the conclusion of Theorem 2.1 still holds true when the predictors are constructed using Equation (1.11).

Theorem 2.2. *Suppose that $\mathbf{x}_{k+1} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. For any prediction point $\mathbf{x}_0 \in \mathcal{X}$, let $MSE(\hat{y}_k(\mathbf{x}_0))$ and $MSE(\hat{y}_{k+1}(\mathbf{x}_0))$ denote the MSEs of the predictors $\hat{y}_k(\mathbf{x}_0)$ and $\hat{y}_{k+1}(\mathbf{x}_0)$ constructed using Equation (1.11). If Assumption 1 holds and $\mathbf{F}_k$ has full column rank, then $MSE(\hat{y}_k(\mathbf{x}_0)) \geq MSE(\hat{y}_{k+1}(\mathbf{x}_0))$.*

Proof of Theorem 2.2. We use the same shorthand notation A, Φ, and $\phi(\mathbf{x}_0)$ as in the proof of Theorem 2.1. When $\hat{y}_{k+1}(\mathbf{x}_0)$ is constructed using (1.11), its associated MSE becomes

$$
\begin{aligned}
MSE(\hat{y}_{k+1}(\mathbf{x}_0)) = \tau^2 - \Sigma_{M_{k+1}}(\mathbf{x}_0, \cdot)^\intercal(\Sigma_{M_{k+1}} + \Sigma_{\epsilon_{k+1}})^{-1}\Sigma_{M_{k+1}}(\mathbf{x}_0, \cdot) \\
+ \eta_{k+1}(\mathbf{x}_0)^\intercal(\mathbf{F}_{k+1}^\intercal(\Sigma_{M_{k+1}} + \Sigma_{\epsilon_{k+1}})^{-1}\mathbf{F}_{k+1})^{-1}\eta_{k+1}(\mathbf{x}_0),
\end{aligned}
\tag{2.1}
$$

where $\eta_{k+1}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0) - \mathbf{F}_{k+1}^\intercal(\Sigma_{M_{k+1}} + \Sigma_{\epsilon_{k+1}})^{-1}\Sigma_{M_{k+1}}(\mathbf{x}_0, \cdot)$. From the proof of Theorem 2.1, it is easy to see that $\Sigma_{M_{k+1}}(\mathbf{x}_0, \cdot)^\intercal(\Sigma_{M_{k+1}} + \Sigma_{\epsilon_{k+1}})^{-1}\Sigma_{M_{k+1}}(\mathbf{x}_0, \cdot) = \Sigma_{M_k}(\mathbf{x}_0, \cdot)^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0, \cdot) + \phi(\mathbf{x}_0)^2\Phi$. Thus, the right-hand-side of (2.1) can be written in terms of $MSE(\hat{y}_k(\mathbf{x}_0))$ as

$$
\begin{aligned}
&MSE(\hat{y}_{k+1}(\mathbf{x}_0)) = \\
&\qquad MSE(\hat{y}_k(\mathbf{x}_0)) - \phi(\mathbf{x}_0)^2\Phi - \eta_k(\mathbf{x}_0)^\intercal(\mathbf{F}_k^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\mathbf{F}_k)^{-1}\eta_k(\mathbf{x}_0) \\
&\qquad + \eta_{k+1}(\mathbf{x}_0)^\intercal(\mathbf{F}_{k+1}^\intercal(\Sigma_{M_{k+1}} + \Sigma_{\epsilon_{k+1}})^{-1}\mathbf{F}_{k+1})^{-1}\eta_{k+1}(\mathbf{x}_0).
\end{aligned}
\tag{2.2}
$$

Regarding the last term in (2.2), we have

$$
\begin{aligned}
&\mathbf{F}_{k+1}^\intercal(\Sigma_{M_{k+1}} + \Sigma_{\epsilon_{k+1}})^{-1}\mathbf{F}_{k+1} \\
&= (\mathbf{F}_k^\intercal, \mathbf{f}(\mathbf{x}_{k+1}))A \begin{pmatrix} \mathbf{F}_k \\ \mathbf{f}(\mathbf{x}_{k+1})^\intercal \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbf{F}_k^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\mathbf{F}_k + \\
&\quad \big(\mathbf{F}_k^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{k\times 1} - \mathbf{f}(\mathbf{x}_{k+1})\big)\Phi\big(\Sigma_{k\times 1}^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\mathbf{F}_k - \mathbf{f}(\mathbf{x}_{k+1})^\intercal\big) \\
&= \mathbf{F}_k^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\mathbf{F}_k + \psi^\intercal\Phi\psi, \qquad (2.3)
\end{aligned}
$$

where $\psi = \Sigma_{k\times 1}^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\mathbf{F}_k - \mathbf{f}(\mathbf{x}_{k+1})^\intercal$. On the other hand,

$$
\begin{aligned}
\eta_{k+1}(\mathbf{x}_0) &= \mathbf{f}(\mathbf{x}_0) - \mathbf{F}_{k+1}^\intercal(\Sigma_{M_{k+1}} + \Sigma_{\epsilon_{k+1}})^{-1}\Sigma_{M_{k+1}}(\mathbf{x}_0, \cdot) \\
&= \mathbf{f}(\mathbf{x}_0) - (\mathbf{F}_k^\intercal, \mathbf{f}(\mathbf{x}_{k+1}))A\begin{pmatrix} \Sigma_{M_k}(\mathbf{x}_0, \cdot) \\ \\ \Sigma_M(\mathbf{x}_0, \mathbf{x}_{k+1}) \end{pmatrix} \\
&= \mathbf{f}(\mathbf{x}_0) - \mathbf{F}_k^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0, \cdot) \\
&\quad - \big(\mathbf{F}_k^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{k\times 1} - \mathbf{f}(\mathbf{x}_{k+1})\big)\Phi \\
&\quad \big(\Sigma_{k\times 1}^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0, \cdot) - \Sigma_M(\mathbf{x}_0, \mathbf{x}_{k+1})\big) \\
&= \eta_k(\mathbf{x}_0) - \psi^\intercal\Phi\phi(\mathbf{x}_0) \qquad (2.4)
\end{aligned}
$$

Substituting (2.3) and (2.4) into (2.2), we obtain

$$
\begin{aligned}
&MSE(\hat{y}_{k+1}(\mathbf{x}_0)) = \\
&\quad MSE(\hat{y}_k(\mathbf{x}_0)) - \phi(\mathbf{x}_0)^2\Phi - \eta_k(\mathbf{x}_0)^\intercal(\mathbf{F}_k^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\mathbf{F}_k)^{-1}\eta_k(\mathbf{x}_0) + \\
&\quad (\eta_k(\mathbf{x}_0) - \psi^\intercal\Phi\phi(\mathbf{x}_0))^\intercal(\mathbf{F}_k^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\mathbf{F}_k + \psi^\intercal\Phi\psi)^{-1}(\eta_k(\mathbf{x}_0) - \psi^\intercal\Phi\phi(\mathbf{x}_0)).
\end{aligned}
\tag{2.5}
$$

Next, define $W = \mathbf{F}_k^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\mathbf{F}_k$. Since $\mathbf{F}_k$ is assumed to have full column rank and $(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}$ is positive definite, it follows that W is also positive definite and its inverse W^{-1} exists. Therefore, by the Sherman-Morrison-Woodbury formula, we have

$$
(\mathbf{F}_k^\intercal(\Sigma_{M_k} + \Sigma_{\epsilon_k})^{-1}\mathbf{F}_k + \psi^\intercal\Phi\psi)^{-1} = (W + \psi^\intercal\Phi\psi)^{-1} = W^{-1} - \frac{W^{-1}\psi^\intercal\psi W^{-1}}{\Phi^{-1} + \psi W\psi^\intercal}.
\tag{2.6}
$$

Finally, substituting (2.6) into (2.5), we get

$$
\begin{aligned}
MSE(\hat{y}_{k+1}(\mathbf{x}_0)) &= MSE(\hat{y}_k(\mathbf{x}_0)) - \phi(\mathbf{x}_0)^2\Phi - \eta_k(\mathbf{x}_0)^\intercal W^{-1}\eta_k(\mathbf{x}_0) \\
&\quad + (\eta_k(\mathbf{x}_0) - \psi^\intercal\Phi\phi(\mathbf{x}_0))^\intercal\Big(W^{-1} - \frac{W^{-1}\psi^\intercal\psi W^{-1}}{\Phi^{-1} + \psi W^{-1}\psi^\intercal}\Big)(\eta_k(\mathbf{x}_0) - \psi^\intercal\Phi\phi(\mathbf{x}_0)) \\
= MSE(\hat{y}_k(\mathbf{x}_0)) &- \phi(\mathbf{x}_0)^2\Phi - \eta_k(\mathbf{x}_0)^\intercal\frac{W^{-1}\psi^\intercal\psi W^{-1}}{\Phi^{-1} + \psi W^{-1}\psi^\intercal}\eta_k(\mathbf{x}_0)
\end{aligned}
$$

$$
\begin{aligned}
&- \eta_k(\mathbf{x}_0)^\intercal W^{-1}\psi^\intercal \Phi\phi(\mathbf{x}_0) + \eta_k(\mathbf{x}_0)^\intercal \frac{W^{-1}\psi^\intercal\psi W^{-1}}{\Phi^{-1}+\psi W^{-1}\psi^\intercal}\psi^\intercal\Phi\phi(\mathbf{x}_0)\\
&- \phi(\mathbf{x}_0)\Phi\psi W^{-1}\eta_k(\mathbf{x}_0) + \phi(\mathbf{x}_0)\Phi\psi\frac{W^{-1}\psi^\intercal\psi W^{-1}}{\Phi^{-1}+\psi W^{-1}\psi^\intercal}\eta_k(\mathbf{x}_0)\\
&+ \phi(\mathbf{x}_0)\Phi\psi W^{-1}\psi^\intercal\Phi\phi(\mathbf{x}_0) - \phi(\mathbf{x}_0)\Phi\psi\frac{W^{-1}\psi^\intercal\psi W^{-1}}{\Phi^{-1}+\psi W^{-1}\psi^\intercal}\psi^\intercal\Phi\phi(\mathbf{x}_0)\\
&= MSE(\hat{y}_k(\mathbf{x}_0)) - \phi(\mathbf{x}_0)^2\Phi - \frac{(\eta_k(\mathbf{x}_0)^\intercal W^{-1}\psi^\intercal)^2}{\Phi^{-1}+\psi W^{-1}\psi^\intercal} - \frac{2\phi(\mathbf{x}_0)(\eta_k(\mathbf{x}_0)^\intercal W^{-1}\psi^\intercal)}{\Phi^{-1}+\psi W^{-1}\psi^\intercal}\\
&\quad + \frac{\phi(\mathbf{x}_0)^2\Phi\psi W^{-1}\psi^\intercal}{\Phi^{-1}+\psi W^{-1}\psi^\intercal}\\
&= MSE(\hat{y}_k(\mathbf{x}_0)) - \frac{(\phi(\mathbf{x}_0)+\eta_k(\mathbf{x}_0)^\intercal W^{-1}\psi^\intercal)^2}{\Phi^{-1}+\psi W^{-1}\psi^\intercal}\\
&\leq MSE(\hat{y}_k(\mathbf{x}_0)).
\end{aligned}
$$

This completes the proof of the Theorem 2. □

To summarize, we have shown that under the SK framework, no matter the parameter vector β is known or estimated, the MSE of the corresponding predictor is monotonically non-increasing as the number of design points increases.

2.3 Monotonic Performance of SKG Predictors

In this section, we investigate the monotonicity property for the MSE of predictors under the SKG setting. In addition, we compare the monotonicity properties between SK framework and SKG framework. We validate the hypothesis that additional gradient information is helpful to improve the prediction accuracy in general cases. In addition, we will quantify the improvement and amount of reduction in the prediction under this setting.

Before going into detailed proofs, the following notation is needed. Let $\Sigma_{y,d} = E[(\bar{\mathbf{y}} - E[\bar{\mathbf{y}}])(\bar{\mathcal{D}} - E[\bar{\mathcal{D}}])^\intercal]$ be the $k \times kd$ cross-covariance matrix between the averaged simulation responses and gradient estimators. Let $\Sigma_{d,d} = E[(\bar{\mathcal{D}} - E[\bar{\mathcal{D}}])(\bar{\mathcal{D}} - E[\bar{\mathcal{D}}])^\intercal]$ be the $kd \times kd$ covariance matrix of the averaged gradient estimators at all design points. Based on the introduction of Section 1.4 and the following assumptions, we will show the monotonicity properties for SKG model.

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

Assumption 3 guarantees that the Gaussian process M has differentiable sample paths and ensures the validity of (1.13). By the linearity of the differential operator, the first order partial derivative process $D^r(\mathbf{x}_i), r = 1, \ldots, d$ of $Y(\mathbf{x}_i)$ is also Gaussian; see e.g., [14]. Thus, it is natural to also assume the following condition:

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

The main results of this section are given in Theorems 2.3 and 2.4, 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, for simplicity we omit the subscript k in these theorems.

Theorem 2.3. *Let* $\hat{y}(\mathbf{x})$ *be the SK predictor constructed using (1.9) 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* (1.9). *If Assumptions 1, 2, 3 and 4 hold, then* $MSE(\hat{y}(\mathbf{x})) \geq MSE(\hat{y}^+(\mathbf{x}))$ *for any* $\mathbf{x} \in \mathcal{X}$.

Proof. Proof of Theorem 3. Under Assumptions 1 and 2, it is not difficult to verify that $\Sigma^+_M + \Sigma^+_\epsilon$ can be written in the block form $\Sigma^+_M + \Sigma^+_\epsilon = \begin{pmatrix} \Sigma_M + \Sigma_\epsilon & \Sigma_{y,d} \\ \Sigma^\intercal_{y,d} & \Sigma_{d,d} \end{pmatrix}$. Moreover, Assumptions 1, 2, 3 and 4 imply that both $\Sigma_M + \Sigma_\epsilon$ and $\Sigma_{d,d}$ are invertible. As a result, the symmetric matrix $\Sigma^+_M + \Sigma^+_\epsilon$ is invertible and its inverse J can be calculated using the block matrix inversion formula as follows:

$$J = \begin{pmatrix} (\Sigma_M + \Sigma_\epsilon)^{-1} + (\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_{y,d}Q\Sigma_{y,d}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1} & -(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_{y,d}Q \\ -Q\Sigma_{y,d}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1} & Q \end{pmatrix},$$

where $Q = (\Sigma_{d,d} - \Sigma_{y,d}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_{y,d})^{-1}$.

Next we show that Q is positive definite. To proceed, note that Assumption 3 and 4 imply that Σ_M^+ is positive definite. This, together with the positive semi-definiteness of the covariance matrix Σ_ϵ^+, indicates that the sum $\Sigma_M^+ + \Sigma_\epsilon^+$ is also positive definite. In addition, we note that $\begin{pmatrix} I & 0 \\ -((\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_{y,d})^\intercal & I \end{pmatrix} \begin{pmatrix} \Sigma_M + \Sigma_\epsilon & \Sigma_{y,d} \\ \Sigma_{y,d}^\intercal & \Sigma_{d,d} \end{pmatrix} \begin{pmatrix} I & -(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_{y,d} \\ 0 & I \end{pmatrix}$ $= \begin{pmatrix} \Sigma_M + \Sigma_\epsilon & 0 \\ 0 & Q^{-1} \end{pmatrix}$. Since the matrix $\begin{pmatrix} I & -(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_{y,d} \\ 0 & I \end{pmatrix}$ has full column rank and $\Sigma_M^+ + \Sigma_\epsilon^+$ is positive definite, the matrix $\begin{pmatrix} \Sigma_M + \Sigma_\epsilon & 0 \\ 0 & Q^{-1} \end{pmatrix}$ must be positive definite. This shows that the principal submatrix Q^{-1} (and hence Q) is positive definite.

The MSE of the SKG predictor $\hat{y}^+(\mathbf{x})$ can be obtained via (1.10) by replacing the corresponding quantities in the equation with their augmented counterparts. Specifically, we have

$$
\begin{aligned}
&MSE(\hat{y}^{+}(\mathbf{x})) = \tau^2 - \Sigma_M^{+}(\mathbf{x},\cdot)^{\intercal}(\Sigma_M^{+} + \Sigma_\epsilon^{+})^{-1}\Sigma_M^{+}(\mathbf{x},\cdot) \\
&= \tau^2 - (\Sigma_M(\mathbf{x},\cdot)^{\intercal}, \Sigma_{M,d}(\mathbf{x},\cdot)^{\intercal}) \begin{pmatrix} \Sigma_M + \Sigma_\epsilon & \Sigma_{y,d} \\ \Sigma_{y,d}^{\intercal} & \Sigma_{d,d} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_M(\mathbf{x},\cdot) \\ \Sigma_{M,d}(\mathbf{x},\cdot) \end{pmatrix} \\
&= \tau^2 - (\Sigma_M(\mathbf{x},\cdot)^{\intercal}, \Sigma_{M,d}(\mathbf{x},\cdot)^{\intercal}) J \begin{pmatrix} \Sigma_M(\mathbf{x},\cdot) \\ \Sigma_{M,d}(\mathbf{x},\cdot) \end{pmatrix} \\
&= \tau^2 - \Sigma_M(\mathbf{x},\cdot)^{\intercal}(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_M(\mathbf{x},\cdot) \\
&\quad - \Sigma_M(\mathbf{x},\cdot)^{\intercal}(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_{y,d} Q \Sigma_{y,d}^{\intercal}(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_M(\mathbf{x},\cdot) \\
&\quad + \Sigma_{M,d}(\mathbf{x},\cdot) Q \Sigma_{y,d}^{\intercal}(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_M(\mathbf{x},\cdot) + \Sigma_M(\mathbf{x},\cdot)^{\intercal}(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_{y,d} Q \Sigma_{M,d}(\mathbf{x},\cdot) \\
&\quad - \Sigma_{M,d}(\mathbf{x},\cdot)^{\intercal} Q \Sigma_{M,d}(\mathbf{x},\cdot) \\
&= MSE(\hat{y}(\mathbf{x})) - \kappa^{\intercal} Q \kappa,
\end{aligned}
$$

where we have defined $\kappa = \Sigma_{y,d}^{\intercal}(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_M(\mathbf{x},\cdot) - \Sigma_{M,d}(\mathbf{x},\cdot)$. Consequently, the desired claim follows from the positive definiteness of Q. □

The following result shows that the conclusion of Theorem 2.3 still holds true when SKG predictors are constructed using Equation (1.11).

Theorem 2.4. *Let $\hat{y}(\mathbf{x})$ be the SK predictor constructed using (1.11) 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 (1.11). If $\mathbf{F}$ has full column rank and Assumptions 1, 2, 3 and 4 hold, then $MSE(\hat{y}(\mathbf{x})) \geq MSE(\hat{y}^+(\mathbf{x}))$ for any $\mathbf{x} \in \mathcal{X}$.

Proof. Proof of Theorem 4. We follow the same notation J, Q, and κ used in the proof of Theorem 2.3. When $\hat{y}^+(\mathbf{x})$ is constructed using (1.11), its associated MSE becomes

$$\begin{aligned} MSE(\hat{y}^+(\mathbf{x})) = \tau^2 &- \Sigma_M^+(\mathbf{x}, \cdot)^\intercal(\Sigma_M^+ + \Sigma_\epsilon^+)^{-1}\Sigma_M^+(\mathbf{x}, \cdot) \\ &+ \eta^+(\mathbf{x})^\intercal(\mathbf{F}^{+\intercal}(\Sigma_M^+ + \Sigma_\epsilon^+)^{-1}\mathbf{F}^+)^{-1}\eta^+(\mathbf{x}), \end{aligned} \tag{2.7}$$

where $\eta^+(\mathbf{x}) = \mathbf{f}(\mathbf{x}) - \mathbf{F}^{+\intercal}(\Sigma_M^+ + \Sigma_\epsilon^+)^{-1}\Sigma_M^+(\mathbf{x}, \cdot)$. From the proof of Theorem 2.3, it is easy to see that $\tau^2 - \Sigma_M^+(\mathbf{x}, \cdot)^\intercal(\Sigma_M^+ + \Sigma_\epsilon^+)^{-1}\Sigma_M^+(\mathbf{x}, \cdot) = \tau^2 - \Sigma_M(\mathbf{x}, \cdot)^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_M(\mathbf{x}, \cdot) - \kappa^\intercal Q\kappa$. Thus, the right-hand-side of (2.7) can be written in terms of $MSE(\hat{y}(\mathbf{x}))$ as

$$\begin{aligned} MSE(\hat{y}^+(\mathbf{x})) = MSE(\hat{y}(\mathbf{x})) &- \kappa^\intercal Q\kappa - \eta(\mathbf{x})^\intercal(\mathbf{F}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{F})^{-1}\eta(\mathbf{x}) \\ &+ \eta^+(\mathbf{x})^\intercal(\mathbf{F}^{+\intercal}(\Sigma_M^+ + \Sigma_\epsilon^+)^{-1}\mathbf{F}^+)^{-1}\eta^+(\mathbf{x}). \end{aligned} \tag{2.8}$$

Regarding the last term in (2.8), we have

$$
\begin{aligned}
&\mathbf{F}^{+\intercal}(\Sigma_M^+ + \Sigma_\epsilon^+)^{-1}\mathbf{F}^+ \\
&= (\mathbf{F}^\intercal, \mathbf{F}_d^\intercal) J \begin{pmatrix} \mathbf{F} \\ \\ \mathbf{F}_d \end{pmatrix} \\
&= \mathbf{F}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{F} + \mathbf{F}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_{y,d} Q \Sigma_{y,d}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{F} \\
&\quad - \mathbf{F}_d^\intercal Q \Sigma_{y,d}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{F} - \mathbf{F}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_{y,d} Q \mathbf{F}_d + \mathbf{F}_d^\intercal Q \mathbf{F}_d \\
&= \mathbf{F}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{F} + \gamma^\intercal Q \gamma, \qquad (2.9)
\end{aligned}
$$

where $\gamma = \Sigma_{y,d}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{F} - \mathbf{F}_d$. On the other hand,

$$
\begin{aligned}
\eta^+(\mathbf{x}) &= \mathbf{f}(\mathbf{x}) - \mathbf{F}^{+\intercal}(\Sigma_M^+ + \Sigma_\epsilon^+)^{-1}\Sigma_M^+(\mathbf{x}, \cdot) \\
&= \mathbf{f}(\mathbf{x}) - (\mathbf{F}^\intercal, \mathbf{F}_d^\intercal) J \begin{pmatrix} \Sigma_M(\mathbf{x}, \cdot) \\ \\ \Sigma_{M,d}(\mathbf{x}, \cdot) \end{pmatrix} \\
&= \mathbf{f}(\mathbf{x}) - \mathbf{F}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_{y,d} Q \Sigma_{y,d}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_M(\mathbf{x}, \cdot) \\
&\quad - \mathbf{F}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_M(\mathbf{x}, \cdot) + \mathbf{F}_d^\intercal Q \Sigma_{y,d}(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_M(\mathbf{x}, \cdot) \\
&\quad + \mathbf{F}^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_{y,d} Q \Sigma_{M,d}(\mathbf{x}, \cdot) - \mathbf{F}_d^\intercal Q \Sigma_{M,d}(\mathbf{x}, \cdot) \\
&= \eta(\mathbf{x}) - \gamma^\intercal Q \kappa \qquad (2.10)
\end{aligned}
$$

Substituting (2.9) and (2.10) into (2.8), we obtain

$$
\begin{aligned}
&MSE(\hat{y}^{+}(\mathbf{x})) \\
&= MSE(\hat{y}(\mathbf{x})) - \kappa^{\intercal} Q \kappa + (\eta(\mathbf{x}) - \eta(\mathbf{x})^{\intercal}(\mathbf{F}^{\intercal}(\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{F})^{-1}\eta(\mathbf{x}) \\
&\quad - \gamma^{\intercal} Q \kappa)^{\intercal}(\mathbf{F}^{\intercal}(\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{F} + \gamma^{\intercal} Q \gamma)^{-1}(\eta(\mathbf{x}) - \gamma^{\intercal} Q \kappa) \qquad (2.11)
\end{aligned}
$$

Now define $S = \mathbf{F}^{\intercal}(\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{F}$. Since $\mathbf{F}$ is assumed to have full column rank and $(\Sigma_M + \Sigma_\epsilon)^{-1}$ is positive definite by Lemma 2.1, it follows that S is positive definite and S^{-1} exists. Therefore, by the Sherman-Morrison-Woodbury formula, we have

$$
\begin{aligned}
&(\mathbf{F}^{\intercal}(\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{F} + \gamma^{\intercal} Q \gamma)^{-1} \\
=&(S + \gamma^{\intercal} Q \gamma)^{-1} \\
=&S^{-1} - S^{-1}\gamma^{\intercal}(Q^{-1} + \gamma S^{-1}\gamma^{\intercal})^{-1}\gamma S^{-1}. \qquad (2.12)
\end{aligned}
$$

Combining (2.12) and (2.11), we get

$$
\begin{aligned}
& MSE(\hat{y}^{+}(\mathbf{x})) \\
&= MSE(\hat{y}(\mathbf{x})) - \kappa^{\intercal} Q \kappa - \eta(\mathbf{x})^{\intercal} S^{-1} \gamma^{\intercal} Q \kappa - \eta(\mathbf{x})^{\intercal} S^{-1} \gamma^{\intercal} (Q^{-1} + \gamma S^{-1} \gamma^{\intercal})^{-1} \gamma S^{-1} \eta(\mathbf{x}) \\
&\quad + \eta(\mathbf{x})^{\intercal} S^{-1} \gamma^{\intercal} (Q^{-1} + \gamma S^{-1} \gamma^{\intercal})^{-1} \gamma S^{-1} \gamma^{\intercal} Q \kappa - \kappa^{\intercal} Q \gamma S^{-1} \eta(\mathbf{x}) + \kappa^{\intercal} Q \gamma S^{-1} \gamma^{\intercal} Q \kappa \\
&\quad + \kappa^{\intercal} Q \gamma S^{-1} \gamma^{\intercal} (Q^{-1} + \gamma S^{-1} \gamma^{\intercal})^{-1} \gamma S^{-1} \eta(\mathbf{x}) - \kappa^{\intercal} Q \gamma S^{-1} \gamma^{\intercal} (Q^{-1} + \gamma S^{-1} \gamma^{\intercal})^{-1} \gamma S^{-1} \gamma^{\intercal} Q \kappa \\
&= MSE(\hat{y}(\mathbf{x})) - \kappa^{\intercal} Q \kappa - \eta(\mathbf{x})^{\intercal} S^{-1} \gamma^{\intercal} (I_{kd \times kd} - (Q^{-1} + \gamma S^{-1} \gamma^{\intercal})^{-1} \gamma S^{-1} \gamma^{\intercal}) Q \kappa \\
&\quad - \kappa^{\intercal} Q (I_{kd \times kd} - \gamma S^{-1} \gamma^{\intercal} (Q^{-1} + \gamma S^{-1} \gamma^{\intercal})^{-1}) \gamma S^{-1} \eta(\mathbf{x}) \\
&\quad + \kappa^{\intercal} Q \gamma S^{-1} \gamma^{\intercal} (I_{kd \times kd} - (Q^{-1} + \gamma S^{-1} \gamma^{\intercal})^{-1} \gamma S^{-1} \gamma^{\intercal}) Q \kappa \\
&\quad - \eta(\mathbf{x})^{\intercal} S^{-1} \gamma^{\intercal} (Q^{-1} + \gamma S^{-1} \gamma^{\intercal})^{-1} \gamma S^{-1} \eta(\mathbf{x}) \qquad (2.13)
\end{aligned}
$$

Since Q and S are positive definite, the symmetric matrix $(Q^{-1} + \gamma S^{-1} \gamma^{\intercal})^{-1}$ is also positive definite. Therefore, there exists an orthogonal matrix P and a diagonal matrix Λ with positive diagonal entries such that $(Q^{-1} + \gamma S^{-1} \gamma^{\intercal})^{-1} = P \Lambda P^{\intercal}$ and $\gamma S^{-1} \gamma^{\intercal} = P \Lambda^{-1} P^{\intercal} - Q^{-1}$. Substituting these equations into (2.13), it follows that

$$
\begin{aligned}
& MSE(\hat{y}^{+}(\mathbf{x})) = MSE(\hat{y}(\mathbf{x})) - \eta(\mathbf{x})^{\intercal} S^{-1} \gamma^{\intercal} (I_{kd \times kd} - P \Lambda P^{\intercal} (P \Lambda^{-1} P^{\intercal} - Q^{-1})) Q \kappa \\
&\quad - \kappa^{\intercal} Q (I_{kd \times kd} - (P \Lambda^{-1} P^{\intercal} - Q^{-1}) P \Lambda P^{\intercal}) \gamma S^{-1} \eta(\mathbf{x})
\end{aligned}
$$

$$
\begin{aligned}
&+\kappa^{\intercal}Q(P\Lambda^{-1}P^{\intercal}-Q^{-1})(I_{kd\times kd}-P\Lambda P^{\intercal}(P\Lambda^{-1}P^{\intercal}-Q^{-1}))Q\kappa-\kappa^{\intercal}Q\kappa-\\
&\eta(\mathbf{x})^{\intercal}S^{-1}\gamma^{\intercal}P\Lambda P^{\intercal}\gamma S^{-1}\eta(\mathbf{x})\\
&=MSE(\hat{y}(\mathbf{x}))-\eta(\mathbf{x})^{\intercal}S^{-1}\gamma^{\intercal}(P\Lambda P^{\intercal}Q^{-1})Q\kappa-\kappa^{\intercal}Q(Q^{-1}P\Lambda P^{\intercal})\gamma S^{-1}\eta(\mathbf{x})\\
&\quad+\kappa^{\intercal}Q(P\Lambda^{-1}P^{\intercal}-Q^{-1})(P\Lambda P^{\intercal}Q^{-1})Q\kappa-\kappa^{\intercal}Q\kappa-\eta(\mathbf{x})^{\intercal}S^{-1}\gamma^{\intercal}P\Lambda P^{\intercal}\gamma S^{-1}\eta(\mathbf{x})\\
&=MSE(\hat{y}(\mathbf{x}))-\eta(\mathbf{x})^{\intercal}S^{-1}\gamma^{\intercal}P\Lambda P^{\intercal}\kappa-\kappa^{\intercal}P\Lambda P^{\intercal}\gamma S^{-1}\eta(\mathbf{x})+\kappa^{\intercal}Q\kappa-\kappa P\Lambda P^{\intercal}\kappa-\kappa^{\intercal}Q\kappa\\
&\quad-\eta(\mathbf{x})^{\intercal}S^{-1}\gamma^{\intercal}P\Lambda P^{\intercal}\gamma S^{-1}\eta(\mathbf{x})\\
&=MSE(\hat{y}(\mathbf{x}))-(\kappa+\gamma(S)^{-1}\eta(\mathbf{x}))^{\intercal}((Q)^{-1}+\gamma(S)^{-1}\gamma^{\intercal})^{-1}(\kappa+\gamma(S)^{-1}\eta(\mathbf{x}))\\
&\leq MSE(\hat{y}(\mathbf{x})), \qquad (2.14)
\end{aligned}
$$

□

Given a set of design points $\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}$, consider a prediction point $\mathbf{x}_0\notin\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}$. Let $\hat{y}^{+}_{k}(\mathbf{x}_0)$ be the SKG predictor constructed using Equation (1.9) based on the set of k design points $\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}$ 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 smaller MSE than $\hat{y}^{+}_{k}(\mathbf{x}_0)$.

Similar to Section 2.2, 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 $\bar{\mathbf{y}}_1^+ = (\bar{y}(\mathbf{x}_{k+1}), \bar{\mathcal{D}}^1(\mathbf{x}_{k+1}), \dots, \bar{\mathcal{D}}^d(\mathbf{x}_{k+1}))^\intercal$ and write

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

For brevity, we also define:

$$\begin{aligned}
&\Sigma_{M_{k+1}}^+ + \Sigma_{\epsilon_{k+1}}^+ = E[(\bar{\mathbf{y}}_{k+1}^+ - E[\bar{\mathbf{y}}_{k+1}^+])(\bar{\mathbf{y}}_{k+1}^+ - E[\bar{\mathbf{y}}_{k+1}^+])^\intercal], \\
&\Sigma_{M_k}^+ + \Sigma_{\epsilon_k}^+ = E[(\bar{\mathbf{y}}_k^+ - E[\bar{\mathbf{y}}_k^+])(\bar{\mathbf{y}}_k^+ - E[\bar{\mathbf{y}}_k^+])^\intercal], \\
&\Sigma_{k,1}^+ = E[(\bar{\mathbf{y}}_k^+ - E[\bar{\mathbf{y}}_k^+])(\bar{\mathbf{y}}_1^+ - E[\bar{\mathbf{y}}_1^+])^\intercal], \\
&\Sigma_{1,1}^+ = E[(\bar{\mathbf{y}}_1^+ - E[\bar{\mathbf{y}}_1^+])(\bar{\mathbf{y}}_1^+ - E[\bar{\mathbf{y}}_1^+])^\intercal], \\
&\Sigma_{M_{k+1}}^+(\mathbf{x}, \cdot) = (\mathrm{Cov}(Y(\mathbf{x}), Y(\mathbf{x}_1)), \dots, \mathrm{Cov}(Y(\mathbf{x}), Y(\mathbf{x}_k)), \\
&\qquad\qquad \mathrm{Cov}(Y(\mathbf{x}), D^1(\mathbf{x}_1)), \dots, \mathrm{Cov}(Y(\mathbf{x}), D^1(\mathbf{x}_k)), \\
&\qquad\qquad \dots, \mathrm{Cov}(Y(\mathbf{x}), D^d(\mathbf{x}_1)), \dots, \mathrm{Cov}(Y(\mathbf{x}), D^d(\mathbf{x}_k)), \\
&\qquad\qquad \mathrm{Cov}(Y(\mathbf{x}), Y(\mathbf{x}_{k+1})), \mathrm{Cov}(Y(\mathbf{x}), D^1(\mathbf{x}_{k+1})), \dots, \mathrm{Cov}(Y(\mathbf{x}), D^d(\mathbf{x}_{k+1})))^\intercal \\
&\qquad \triangleq (\Sigma_{M_k}^+(\mathbf{x}, \cdot)^\intercal, \Sigma_{M_1}^+(\mathbf{x}, \cdot)^\intercal)^\intercal, \qquad (2.15)
\end{aligned}$$

$$\begin{aligned}\mathbf{F}_{k+1}^{+} &= (\mathbf{f}(\mathbf{x}_1), \ldots, \mathbf{f}(\mathbf{x}_{k+1}), \frac{\partial \mathbf{f}(\mathbf{x}_1)}{\partial x_1}, \ldots, \frac{\partial \mathbf{f}(\mathbf{x}_{k+1})}{\partial x_1}, \ldots, \frac{\partial \mathbf{f}(\mathbf{x}_1)}{\partial x_d}, \ldots, \frac{\partial \mathbf{f}(\mathbf{x}_{k+1})}{\partial x_d})^{\intercal} \\ &= (\mathbf{f}(\mathbf{x}_1), \ldots, \mathbf{f}(\mathbf{x}_k), \frac{\partial \mathbf{f}(\mathbf{x}_1)}{\partial x_1}, \ldots, \frac{\partial \mathbf{f}(\mathbf{x}_k)}{\partial x_1}, \ldots, \frac{\partial \mathbf{f}(\mathbf{x}_1)}{\partial x_d}, \ldots, \frac{\partial \mathbf{f}(\mathbf{x}_k)}{\partial x_d}, \\ &\quad \mathbf{f}(\mathbf{x}_{k+1}), \frac{\partial \mathbf{f}(\mathbf{x}_{k+1})}{\partial x_1}, \ldots, \frac{\partial \mathbf{f}(\mathbf{x}_{k+1})}{\partial x_d})^{\intercal} \\ &\triangleq (\mathbf{F}_k^{+\intercal}, \mathbf{F}_1^{+\intercal})^{\intercal}. \end{aligned} \tag{2.16}$$

Note that in (2.15), $\Sigma_{M_k}^{+}(\mathbf{x}, \cdot)$ is a $k(d+1) \times 1$ vector and $\Sigma_{M_1}^{+}(\mathbf{x}, \cdot)$ is a $(d+1) \times 1$ vector, and in (2.16), $\mathbf{F}_k^{+}$ is a $k(d+1) \times p$ matrix and $\mathbf{F}_1^{+}$ is a $(d+1) \times p$ matrix.

Corollary 2.1. *Suppose that $\mathbf{x}_{k+1} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. For any prediction point $\mathbf{x}_0 \in \mathcal{X}$, let $\hat{y}_k^{+}(\mathbf{x}_0)$ and $\hat{y}_{k+1}^{+}(\mathbf{x}_0)$ be SKG predictors constructed using Equation (1.9). If assumption 1, 2, 3 and 4 hold, then $MSE(\hat{y}_k^{+}(\mathbf{x}_0)) \geq MSE(\hat{y}_{k+1}^{+}(\mathbf{x}_0))$.*

Proof. Proof of Corollary 1. The proof follows straightforwardly by replacing $MSE(\hat{y}(\mathbf{x}))$, $MSE(\hat{y}^{+}(\mathbf{x}))$, $(\Sigma_M^{+} + \Sigma_\epsilon^{+})$, $(\Sigma_M + \Sigma_\epsilon)$, $\Sigma_{y,d}$, $\Sigma_{d,d}$, $\Sigma_M^{+}(\mathbf{x}, \cdot)$, $\Sigma_M(\mathbf{x}, \cdot)$, and $\Sigma_{M,d}(\mathbf{x}, \cdot)$ in the proof of Theorem 2.3 with $MSE(\hat{y}_k^{+}(\mathbf{x}_0))$, $MSE(\hat{y}_{k+1}^{+}(\mathbf{x}_0))$, $(\Sigma_{M_{k+1}}^{+} + \Sigma_{\epsilon_{k+1}}^{+})$, $(\Sigma_{M_k}^{+} + \Sigma_{\epsilon_k}^{+})$, $\Sigma_{k,1}^{+}$, $\Sigma_{1,1}^{+}$, $\Sigma_{M_{k+1}}^{+}(\mathbf{x}_0, \cdot)$, $\Sigma_{M_k}^{+}(\mathbf{x}_0, \cdot)$, and $\Sigma_{M_1}^{+}(\mathbf{x}_0, \cdot)$, respectively. We omit the details. □

Similarly, when SKG predictors are constructed by applying Equation (1.11), we have the same monotonicity property.

Corollary 2.2. *Suppose that* $\mathbf{x}_{k+1} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$*. For any prediction point* $\mathbf{x}_0 \in \mathcal{X}$*, let* $\hat{y}_k^+(\mathbf{x}_0)$ *and* $\hat{y}_{k+1}^+(\mathbf{x}_0)$ *be SKG predictors constructed using Equation (1.11). If Assumptions 1, 2, 3 and 4 hold and* $\mathbf{F}_k^+$ *has full column rank, then* $MSE(\hat{y}_k^+(\mathbf{x}_0)) \geq MSE(\hat{y}_{k+1}^+(\mathbf{x}_0))$*.*

Proof. Proof of Corollary 2.The proof is identical to the proof of Theorem 2.4 with $MSE(\hat{y}_k^+(\mathbf{x}_0))$, $MSE(\hat{y}_{k+1}^+(\mathbf{x}_0))$, $(\Sigma_{M_{k+1}}^+ + \Sigma_{\epsilon_{k+1}}^+)$, $(\Sigma_{M_k}^+ + \Sigma_{\epsilon_k}^+)$, $\Sigma_{k,1}^+$, $\Sigma_{1,1}^+$, $\Sigma_{M_{k+1}}^+(\mathbf{x}_0, \cdot)$, $\Sigma_{M_k}^+(\mathbf{x}_0, \cdot)$, $\Sigma_{M_1}^+(\mathbf{x}_0, \cdot)$, $\mathbf{F}_{k+1}^+$, $\mathbf{F}_k^+$, $\mathbf{F}_1^+$ replacing $MSE(\hat{y}(\mathbf{x}))$, $MSE(\hat{y}^+(\mathbf{x}))$, $(\Sigma_M^+ + \Sigma_\epsilon^+)$, $(\Sigma_M + \Sigma_\epsilon)$, $\Sigma_{y,d}$, $\Sigma_{d,d}$, $\Sigma_M^+(\mathbf{x}, \cdot)$, $\Sigma_M(\mathbf{x}, \cdot)$, $\Sigma_{M,d}(\mathbf{x}, \cdot)$, $\mathbf{F}^+$, $\mathbf{F}$, $\mathbf{F}_d$. □

3 An Adaptive Sequential Kriging Approach

3.1 Motivation

In general, the experiment design for kriging metamodels can be classified into three categories: single-stage method, simple sequential methods without adaption, and sequential methods with adaption to data. Single-stage experiment design fixes the number of design points n in advance, then collects the output performance measures and optimizes all n design points simultaneously according to the chosen criteria. The single-stage method is easy to implement and computationally efficient. However, there is not a rigorous rule about how to determine the predefined n design points. The most commonly used method is Latin hypercube sampling(LHS). In this method, a very large number of design points n may be required when the sample space is large and continuous. The drawback of such type of sampling method is that it fails to consider the specific characteristics of the response surface. Therefore it is possible that too much computation budget is wasted on regions that have little or no additional information about the response surface. On the other hand, there may exist crucial characteristics and useful information of the response surface in the unsampled regions.

Different from single-stage method, sequential methods do not predefine the number of design points at the beginning. Instead, only a small set of initial design points need to be specified in advance. Then at each stage, a new design point is selected based on a certain chosen criteria. The sequential method's adaptability depends on the selection criteria. If the criteria do not depend on information of the unknown response surface, it is a simple sequential design and it can be stopped as soon as either the prediction performance is satisfied or computational resources run out. If the selection criteria contain updated information of the response surface, it is a sequential sampling method with adaption to the metamodel. It is not surprising that sequential methods with adaption usually have better performance due to their ability of adaptively updating the parameters in a metamodel. However, sometimes it can be time consuming to select new design points sequentially. In addition, the validity of sequential methods is more difficult to be justified theoretically.

Developing sequential adaptive methods for simulation metamodels is still an open question and more effort is needed in this topic. Several sequential methods have been proposed in recent years ([9], [13], [36], [37], [38] and [39]). Among all these methods, only two dynamic approaches are comparable to our work: sequential sampling maximum MSE of the predictor ([9], [10]) and sequential exploration in the sample space ([39]). The first approach designs selection criteria solely based on metrics of uncertainty of the predictor, like MSE or IMSE. It was claimed that the prediction accuracy can be improved by selecting design points that can reduce both the intrinsic and extrinsic quantity of uncertainty of the metamodel. For the second approach, the underlying philosophy relies on balancing the exploration and exploitation tradeoff of the simulation budget allocation. However, for both approaches, little rigorous theoretical results on either the quantity of uncertainty reduction or prediction accuracy are available.

To better accommodate the theoretical foundation of sequential sampling design in SK/SKG, we propose a novel sequential sampling algorithm based on the monotonicity properties of SK/SKG established in the previous sections. In our proposed algorithm, we consider both the strategy of design point selection and simulation budget allocation. Moreover, in order to guarantee the overall model prediction accuracy of our algorithm, a detailed theoretical proof is provided in the idealized setting when the number of design points goes to infinity.

3.2 Existing Sequential Design Algorithms

3.2.1 Sequential Adaptive Approaches based on MSE or IMSE

Under the deterministic kriging metamodels, it has been stated in [9] that the most widely-used criteria for sequential experiment design are the Integrated Mean Square Error (IMSE)

$$IMSE(k) \triangleq \int_{\mathbf{x}\in\mathcal{X}} MSE(\tilde{y}_k(\mathbf{x}))d\mathbf{x} \tag{3.1}$$

and the Maximum Mean Square Error

$$max_{\mathbf{x}\in\mathcal{X}} MSE(\tilde{y}_k(\mathbf{x}))d\mathbf{x}. \tag{3.2}$$

One typical sequential algorithm based on IMSE in ordinary or universal kriging is to use simple one-step look-ahead schemes ([9]). This sequential approach simply searches the design point that attains the largest reduction in IMSE over the whole region. However, to our best knowledge, little theoretical analyses, such as convergence analysis and performance guarantee when the number of design points goes to infinity, are available for such sequential algorithm. In particular, it is not clear how much uncertainty has been reduced when an optimal design point is and whether such sequential algorithm can be applied in stochastic settings.

The basic idea of sequential approaches based on MSE is to select the design point that maximizes the current MSE estimate. Some variations of MSE are also used as a selection criterion, like the confidence interval proposed in [37] and the relative error proposed in [38]. However, when the space of design points is continuous, multiple maximum MSE estimators may exist in the early stage of a sequential fitting procedure, making it impossible to select a unique design point. The following example is an illustration of such a scenario.

Consider the unknown response surface as a deterministic function with artificial noise. We use stochastic kriging with gradient estimators to model its surface. Let

$$y(x) = exp(-1.4x)cos(\frac{7\pi x}{2}) + \epsilon, -3 \leq x \leq 0,$$

where the noise $\epsilon \sim \mathcal{N}(0, 1)$. Assume the gradient can be estimated with noise, identified by

$$g(x) = y'(x) + \zeta,$$

where $\zeta \sim \mathcal{N}(0,1)$, and the initial design points are given by Latin hypercube designs (LHD) with good space-filling properties, i.e. $\mathcal{I} = \{-3, -2.5, -2, -1.5, -1, -0.5, 0\}$. The number of replications for simulations at each design point is 30.

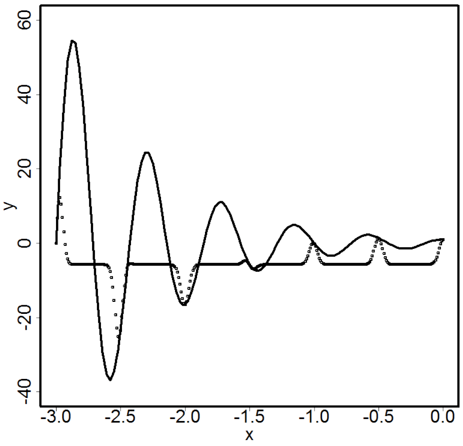


Figure 1: Comparisons between the fitted response surface and the true surface

It can be seen from Figure 1 that the fitted surface is only close to the true surface in the neighbourhood of the design points. However, the fitted surface is almost flat in the area between any of two consecutive design

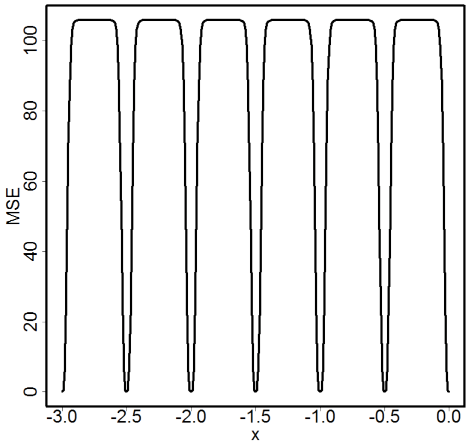


Figure 2: The MSE estimator of the fitted predictor

points. This is because no information from the true response surface is available from the initial design points. Figure 2 shows the MSE estimator of the predictor as a function of the design points. The MSEs of the predictor at the initial design points are reduced to zero. While the MSE estimators are relatively high in the unexplored area. It is interesting to see that the MSEs achieve almost the same maximum value in the unexplored area between two consecutive points. Therefore, it is not clear how to sample a unique point that maximize the MSE estimator.

In general, the sequential procedures based on MSE and IMSE criteria are intuitively easy to understand. However, few theoretical analyses have been conducted with regard to the explicit reduction of MSE/IMSE, the role of intrinsic/extrinsic noise and the uncertainty measure performance for such procedures.

3.2.2 Adaptive Exploration-Exploitation Sampling Algorithm

Different from traditional sequential approaches based on MSE and IMSE criterion, the adaptive exploration-exploitation sampling algorithm (AEES) proposed by Ajdari and Mahlooji ([39]) focuses on decomposing the sample space and then select a new design point according to the principle of balancing the exploration and exploitation. The philosophy of exploration is to focus on covering the entire search space with the least possible unsampled regions. One typical example is space-filling design. However, exploitative methods primarily focus on regions with more interesting characteristics than others. These exploitative methods attempt to skip regions that contain no significant information about characteristics of the surface but in favor of those that are considered to be more informative. Thus it is not surprising to see that there is a conflict between exploration and exploitation methods due to the limitation of simulation budgets. One of the disadvantages associated with exploitation-based methods is that the sample points may become clustered in some regions while other parts remain unsampled. On the other hand, another disadvantage of exploration-based methods is that the simulation efforts may be spent in regions that are not informative of interest. The method proposed by Ajdari and Mahlooji combines the capabilities of both exploration and exploitation methods. It aims at using the advantage of both the exploration and exploitation methods while avoiding the drawbacks associated with each one of them.

The novelty of the proposed algorithm in [39] lies in its method for selecting the next design point, called Delaunay-Hybrid Adaptive Sequential Design (DHASD). In particular, when selecting the new design point, the sample space is decomposed into a mesh of triangles upon the current design points by the Delaunay Triangulation algorithm. The interior of each triangle is treated as an unsampled region. Then two criteria are considered to represent the potential of each triangle for the subsequent investigation based on exploration and exploitation capabilities. One criterion denoted by the exploration score is related to the space-filling characteristic of the experiment design and guaranties a good coverage of the sample space. On the other hand, the exploitation criterion represents the degree of capturing the response surface's information and thus improving the accuracy of the metamodel. Finally the two criteria are combined in an adaptive manner so that the triangle that maximize the total score is the most potential region. As a result, the centroid of the triangle is selected as the new design point.

The idea of AEES is creative. However, one main concern about AEES algorithm is that no characteristic of the metamodel is involved in the selection criteria. The next point is only selected based on the information of the unknown response surface. In other words, we may change the metamodel to any other appropriate ones but keep the same selection criteria. The algorithm will still work. Therefore, the AEES is general to various kinds of metamodels but probably is not the optimal adaptive sequential sampling approach for SK/SKG metamodels. In the next section, we will propose a novel adaptive sequential sampling approach that combines both the structural information of the metamodel and the philosophy of exportation-exploitation.

3.3 Adaptive Sequential Algorithm Design

Based on the monotonicity analysis of the MSE estimators in SK/SKG metamodels, we propose an adaptive sequential sampling method for the SK/SKG metamodeling. Suppose we have selected a set of design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ and a set of 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 (1.11). Our goal is to provide a global fit of an unknown response surface. Therefore it is hoped that $\tilde{y}_k(\mathbf{x})$ would be more accurate to represent the true surface after updating the parameters in SK/SKG metamodels iteratively. To evaluate the overall quality of the prediction of the model over the design space, the integrated MSE (IMSE)

$$IMSE(k) \triangleq \int_{\mathbf{x}\in\mathcal{X}} MSE(\tilde{y}_k(\mathbf{x}))d\mathbf{x} \tag{3.3}$$

serves as a useful criterion and a small IMSE is expected for an accurate SK/SKG predictor.

When the total simulation budget N is fixed and the number of design points K is specified, a two-stage sequential design is proposed by Ankenman et al. (2010) under the assumption that model parameters are known. In the beginning, a small number m of design points are determined for SK model fitting in advance. The remaining $K - m$ number of design points and allocation of replications are selected subsequently to minimize the estimated IMSE. Claiming that choosing the remaining design points in a space-filling way would avoid solving a high-dimensional non-linear optimization problem, the authors focused their research on how to allocate the simulation replications.

However, when the model parameters are unknown, the underlying assumption of the two-stage sequential design is violated. Thus, the superiority of the two-stage design can not be attained under general circumstances. In order to develop an more efficient strategy, the model parameters should be estimated via the information carried by the design points and the corresponding simulation outputs. Taking these into consideration, a truly optimal design should: (1) sequentially estimate model parameters incorporating simulation outputs at all previously generated points, and (2) minimize the estimated IMSE after all of the budget N has been allocated. However, developing such an optimal strategy would involve solving a stochastic dynamic programming problem, which is computationally expensive.

Motivated by the monotonicity results derived in Section 2, we consider an adaptive sequential approach that myopically maximizes the difference between successive IMSEs at each iteration. Instead of trying to derive a N-step bellman equation, we focused on researching the changes of the model predictor's uncertainty by looking one step ahead. In particular, let $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ be the set of design points and let $\{n_1, \ldots, n_k\}$ be the set of simulation replications on each design point. To tackle the problem of selecting the next iteration design point, we consider the point $\mathbf{x}_{k+1}$ and allocation of replications n_{k+1} that achieve the maximum reduction in the IMSE at each iteration:

$$\begin{aligned}(\mathbf{x}_{k+1}, n_{k+1}) &= \operatorname*{arg\,max}_{\mathbf{x}\in\mathcal{X},\, n\geq 1}\Big(IMSE(k) - IMSE(k+1)\Big)\\ &= \operatorname*{arg\,max}_{\mathbf{x}\in\mathcal{X},\, n\geq 1}\int_{\mathbf{x}_0\in\mathcal{X}}\Big(MSE(\tilde{y}_k(\mathbf{x}_0)) - MSE(\tilde{y}_{k+1}(\mathbf{x}_0))\Big)d\mathbf{x}_0, \quad (3.4)\end{aligned}$$

where $MSE(\tilde{y}_{k+1}(\mathbf{x}))$ is considered as a function of the new location $\mathbf{x}$ given that the kriging parameters are estimated in the $(k)_{th}$ iteration and the number of replications n allocated to $\mathbf{x}$. The $IMSE(k+1)$ is interpreted as the overall uncertainty of the metamodel with the $(k+1)_{th}$ randomly selected design point added under the parameters settings of the $(k)_{th}$ iteration. So there is no need to estimate the IMSEs directly in (3.4) because the difference between MSEs is given in the proofs of Theorem 2.2 and Corollary 2.2. In particular, the reduction in MSE estimator when a new point is added into SK/SKG metamodel are quantified with the formulas. Therefore, we are able to calculate the difference of IMSEs in (3.4). The explicit formulas of the difference in IMSEs are given as follows. We list the difference for SK and SKG metamodels, respectively.

$$\mathbf{x}_{k+1} \triangleq \begin{cases} \underset{\mathbf{x}\in\mathcal{X}}{\arg\max} \quad \int_{\mathbf{x}_0\in\mathcal{X}} \frac{(\phi(\mathbf{x}_0)+\eta_k(\mathbf{x}_0)^\intercal W^{-1}\psi^\intercal)^2}{\Phi^{-1}+\psi W^{-1}\psi^\intercal} d\mathbf{x}_0 & \text{SK metamodel} \\ \underset{\mathbf{x}\in\mathcal{X}}{\arg\max} \quad \int_{\mathbf{x}_0\in\mathcal{X}} (\kappa^+ + \gamma^+(S^+)^{-1}\eta^+)^\intercal((Q^+)^{-1} + \gamma^+(S^+)^{-1}\gamma^{+\intercal})^{-1} & \\ \qquad (\kappa^+ + \gamma^+(S^+)^{-1}\eta^+) d\mathbf{x}_0 & \text{SKG metamodel} \end{cases} \tag{3.5}$$

where $\kappa^+, \gamma^+, S^+, \eta^+, Q^+$ are given as follows.

$$\begin{aligned} \kappa^+ &= \Sigma_{k,1}^{+}{}^\intercal(\Sigma_{M_k}^+ + \Sigma_{\epsilon_k}^+)^{-1}\Sigma_{M_k}^+(\mathbf{x}_0,\cdot) - \Sigma_{M_1}^+(\mathbf{x}_0,\cdot) \\ \gamma^+ &= \Sigma_{k,1}^{+}{}^\intercal(\Sigma_{M_k}^+ + \Sigma_{\epsilon_k}^+)^{-1}\mathbf{F}_k^+ - \mathbf{F}_1^+ \\ S^+ &= \mathbf{F}_k{}^{+\intercal}(\Sigma_{M_k}^+ + \Sigma_{\epsilon_k}^+)^{-1}\mathbf{F}_k^+ \\ \eta^+ &= \mathbf{f}(\mathbf{x}_0) - \mathbf{F}_k{}^{+\intercal}(\Sigma_{M_k}^+ + \Sigma_{\epsilon_k}^+)^{-1}\Sigma_{M_k}^+(\mathbf{x}_0,\cdot) \\ Q^+ &= (\Sigma_{1,1}^+ - \Sigma_{k,1}^{+}{}^\intercal(\Sigma_{M_k}^+ + \Sigma_{\epsilon_k}^+)^{-1}\Sigma_{k,1}^+)^{-1} \end{aligned} \tag{3.6}$$

The main benefit here is that it allows the model parameters (and hence the IMSE in (3.4)) to be estimated incrementally using all data collected up to the current iteration. However, a difficulty associated with (3.4) is that we can no longer maintain the budget constraint $\sum_{k=1}^{K} n_k = N$, because the optimal n_{k+1} that solves (3.4) is always given by $n_{k+1} = \infty$ by looking into the proof in Theorem 2.2 (Corollary 2.2). Therefore, we instead consider the alternative setting of attaining a desired IMSE target $\varepsilon > 0$. In Section 3.4, we show that in 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 $\mathbf{x}_i$, the number of simulation replications at $\mathbf{x}_i$ and the intrinsic-extrinsic noise variance ratio $\frac{V(\mathbf{x}_i)}{\tau^2}$

$$MSE(\tilde{y}_k(\mathbf{x}_i)) \leq \frac{V(\mathbf{x}_i)}{n_i + \frac{V(\mathbf{x}_i)}{\tau^2}}. \tag{3.7}$$

In addition, the MSE is continuous in the sense that $MSE(\tilde{y}_k(\mathbf{x}))$ will stay close to $MSE(\tilde{y}_k(\mathbf{x}_i))$ as long as the new location $\mathbf{x}$ is sufficiently close to $\mathbf{x}_i$. Consequently, if $MSE(\tilde{y}_k(\mathbf{x}_i)) < \frac{\varepsilon}{|\mathcal{X}|}$ at all sampled design points $\mathbf{x}_i$, where $|\mathcal{X}|$ is the volume of the design space $\mathcal{X}$, then it is able to ensure the overall IMSE to fall below a given threshold ε (as the number of design points increases). This, together with (3.7), leads to the condition $n_i > \frac{V(\mathbf{x}_i)(\tau^2|\mathcal{X}|-\varepsilon)}{\varepsilon\tau^2}$, suggesting that n_i should be chosen proportional to the intrinsic variance at $\mathbf{x}_i$.

It is assumed in the above discussion that the intrinsic variance function $V(\mathbf{x})$ should be known. However when it is unknown in practice, two approaches are usually recommended to estimate the intrinsic variance. One is introducing a second SK metamodel to account for both extrinsic and intrinsic uncertainty in estimating V, the other is to apply an ordinary kriging metamodel. When introducing a second SK metamodel, suppose we have obtained n_i simulation replications at design points $\mathbf{x}_i$, i.i.d. noisy observations of $V(\mathbf{x}_i)$ can be obtained via a batch means method (e.g., [40]) by splitting the n_i output performance measures $y_j(\mathbf{x}_i)$ into m_i batches of equal size and then computing the sample variance for each batch (alternatively, a bootstrap resampling approach can be applied when n_i is small; see, e.g., [41]). Then the m_i estimates of $V(\mathbf{x}_i)$ can be used in (1.12) to construct an optimal MSE predictor $\hat{V}(\mathbf{x})$. By using a new SK model, it is able to consider the intrinsic noise when estimating the measures of $V(\mathbf{x}_i)$. As a result, it can provide a relatively more accurate estimate of the intrinsic variance. However, fitting a second SK metamodel that accounts for $V(\mathbf{x})$ sometimes may be overqualified. Another difficulty in using batch mean method is that it is not easy to choose a suitable batch size because the n_i may vary at design points. Such issues can be avoided by using an ordinary kriging at the price of getting a less accurate estimation.

In our computational results reported in Section 4, we use the approach outlined in [13]. Different from incorporating another SK metamodel, it 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. Since $V(\mathbf{x})$ is not directly observable, the intrinsic variance $V(\mathbf{x}_i)$ at a point $\mathbf{x}_i$ is replaced by its sample variance computed using the n_i simulation replications at $\mathbf{x}_i$. The estimates of $V(\mathbf{x}_i)$ are then used in (1.7) 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 (β, τ^2, θ) can be estimated in the way described in [13] by constructing the log-likelihood function and then applying a standard non-linear optimization algorithm to search for the maximum likelihood estimators of (β, τ^2, θ). The same techniques can also be used to estimate the model parameters in the SKG framework; we refer the readers to [14] for more details.

To summarize, we propose the following sampling 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 $k = 0$.

Step 1: Fit $\hat{V}$ (and $\hat{V}_r$) 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

$$\mathbf{x}_{k+1} = \arg\max_{\mathbf{x} \in \mathcal{X}} \quad \widehat{IMSE}(k) - \widehat{IMSE}(k+1),$$

where $\widehat{IMSE}(k)$ is an estimator of $IMSE(k)$ when (unknown) model parameters are replaced by their corresponding estimators. Note that there is no need to estimate $\widehat{IMSE}(k+1)$ in the k_{th} iteration, we solve the candidate point $\mathbf{x}_{k+1}$ through (3.5).

Step 3: Allocate $n_{k+1} > \frac{V(\mathbf{x}_{k+1})(\tau^2|\mathcal{X}|-\varepsilon)}{\varepsilon\tau^2}$ replications to $\mathbf{x}_{k+1}$ and collect output performance measures at $\mathbf{x}_{k+1}$.

Step 4: If $\widehat{IMSE}(k+1) \leq \varepsilon$, then terminate; otherwise set $k = k+1$ and go to **step 1**.

We remark that during the initialization step of ASK, the desired accuracy ε could instead be specified using the average IMSE (AIMSE), i.e., IMSE normalized by the volume of the domain $|\mathcal{X}|$, in which case the choice of n_{k+1} at Step 3 becomes $n_{k+1} > \frac{V(\mathbf{x}_{k+1})(\tau^2-\varepsilon)}{\varepsilon\tau^2}$.

3.4 Theoretical Results

In this section, we justify the validity of the proposed ASK algorithm in an ideal setting. Suppose that all the model parameters are known, based on the monotonicity properties derived in Section 2, the IMSE of the resulting predictor can be smaller than a given threshold if the design points and the simulation replications are sequentially determined by the ASK procedure.

We state a list of lemmas and corollaries to establish our main result. In particular, we will show that under both the SK and SKG frameworks and when either β 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 (3.7); 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. Not surprisingly, these results are consistent with our understanding of deterministic kriging models, where it is well known that the predictor variance is zero at all design locations.

Lemma 3.1. *Given a set of design points* $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$*, let* $\hat{y}(\mathbf{x})$ *be the SK* *predictor constructed using (1.9). 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 $n_i > \frac{V(\mathbf{x}_i)(\tau^2-\varepsilon)}{\varepsilon\tau^2}$, then (a) $MSE(\hat{y}(\mathbf{x}_i)) < \varepsilon$; (b) there exists an $r_i > 0$ such that $MSE(\hat{y}(\mathbf{x})) < \varepsilon$ for all $\mathbf{x} \in B_{r_i}(\mathbf{x}_i) \cap \mathcal{X}$.*

To prove claim (a), We consider there are $k > 1$ design points in kriging model. Let $\hat{y}_{-i}(\mathbf{x})$ be the predictor obtained from the set of $k-1$ design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ by excluding $\mathbf{x}_i$. Thus, it follows from the proof of Theorem 2.1 that

$$\begin{aligned} MSE(\hat{y}(\mathbf{x}_i)) &= MSE(\hat{y}_{-i}(\mathbf{x}_i)) - \frac{\phi(\mathbf{x}_i)^2}{MSE(\hat{y}_{-i}(\mathbf{x}_i)) + \Sigma_\epsilon(\mathbf{x}_i, \mathbf{x}_i)} \\ &= MSE(\hat{y}_{-i}(\mathbf{x}_i)) - \frac{MSE(\hat{y}_{-i}(\mathbf{x}_i))^2}{MSE(\hat{y}_{-i}(\mathbf{x}_i)) + \Sigma_\epsilon(\mathbf{x}_i, \mathbf{x}_i)} \\ &= \frac{MSE(\hat{y}_{-i}(\mathbf{x}_i))\Sigma_\epsilon(\mathbf{x}_i, \mathbf{x}_i)}{MSE(\hat{y}_{-i}(\mathbf{x}_i)) + \Sigma_\epsilon(\mathbf{x}_i, \mathbf{x}_i)} \\ &\leq \frac{\tau^2\Sigma_\epsilon(\mathbf{x}_i, \mathbf{x}_i)}{\tau^2 + \Sigma_\epsilon(\mathbf{x}_i, \mathbf{x}_i)} \end{aligned} \tag{3.8}$$

which, when combined with the fact $n_i > \frac{V(\mathbf{x}_i)(\tau^2-\varepsilon)}{\tau^2\varepsilon}$, shows that

$$MSE(\hat{y}(\mathbf{x}_i)) \leq \frac{\tau^2\Sigma_\epsilon(\mathbf{x}_i, \mathbf{x}_i)}{\tau^2 + \Sigma_\epsilon(\mathbf{x}_i, \mathbf{x}_i)} < \varepsilon. \tag{3.9}$$

Now define $\varepsilon' \triangleq \frac{\tau^2\Sigma_\epsilon(\mathbf{x}_i,\mathbf{x}_i)}{\tau^2+\Sigma_\epsilon(\mathbf{x}_i,\mathbf{x}_i)} < \varepsilon$ and consider a prediction point in the vicinity of $\mathbf{x}_i$, i.e., $\mathbf{x} \in B_{r_i}(\mathbf{x}_i) \cap \mathcal{X}$ for some $r_i > 0$. The MSE of $\hat{y}(\mathbf{x})$ can be expressed as

$$
\begin{aligned}
MSE(\hat{y}(\mathbf{x})) &= \tau^2 - \Sigma_M(\mathbf{x}, \cdot)^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \Sigma_M(\mathbf{x}, \cdot) \\
&= \tau^2 - (\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot) + \Sigma_M(\mathbf{x}_i, \cdot))^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \\
&\qquad (\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot) + \Sigma_M(\mathbf{x}_i, \cdot)) \\
&= \tau^2 - (\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot))^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} (\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot)) - \\
&\qquad 2(\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot))^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \Sigma_M(\mathbf{x}_i, \cdot) - \\
&\qquad \Sigma_M(\mathbf{x}_i, \cdot)^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \Sigma_M(\mathbf{x}_i, \cdot) \\
&= MSE(\hat{y}(\mathbf{x}_i)) - (\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot))^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} (\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot)) - \\
&\qquad 2(\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot))^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \Sigma_M(\mathbf{x}_i, \cdot) \qquad (3.10) \\
&\leq MSE(\hat{y}(\mathbf{x}_i)) - 2(\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot))^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \Sigma_M(\mathbf{x}_i, \cdot), \qquad (3.11) \\
&\leq MSE(\hat{y}(\mathbf{x}_i)) + 2\|\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot)\| \|(\Sigma_M + \Sigma_\epsilon)^{-1} \Sigma_M(\mathbf{x}_i, \cdot)\|, \\
&\qquad (3.12)
\end{aligned}
$$

where inequality (3.11) follows because the second term in (3.10) is non-negative (due to the positive definiteness of $(\Sigma_M + \Sigma_\epsilon)^{-1}$), and the last inequality (3.12) follows from the Cauchy-Schwarz inequality. From the discussion after Assumption 1, we know that $(\Sigma_M + \Sigma_\epsilon)^{-1} \Sigma_M(\mathbf{x}_i, \cdot)$ is a $k \times 1$

vector with bounded Euclidean norm $L_k \triangleq \|(\Sigma_M + \Sigma_\epsilon)^{-1}\Sigma_M(\mathbf{x}_i, \cdot)\|$. In addition, the jth element of the vector $\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot)$ can be written as $\tau^2(R(d(\mathbf{x}, \mathbf{x}_j), \boldsymbol{\theta}) - R(d(\mathbf{x}_i, \mathbf{x}_j), \boldsymbol{\theta}))$. Thus, since $(\varepsilon - \varepsilon')/2 > 0$, the continuity of the correlation $R(d, \boldsymbol{\theta})$, together with the triangle inequality $|d(\mathbf{x}, \mathbf{x}_j) - d(\mathbf{x}_i, \mathbf{x}_j)| \le d(\mathbf{x}, \mathbf{x}_i)$, suggests there exists an r_i (depending on k) such that $|d(\mathbf{x}, \mathbf{x}_j) - d(\mathbf{x}_i, \mathbf{x}_j)| \le d(\mathbf{x}, \mathbf{x}_i) < r_i$ implies $\tau^2(R(d(\mathbf{x}, \mathbf{x}_j), \boldsymbol{\theta}) - R(d(\mathbf{x}_i, \mathbf{x}_j), \boldsymbol{\theta})) \le \frac{\varepsilon - \varepsilon'}{4L_k\sqrt{k}}$. This shows that the second term on the right-hand-side of (3.12) is bounded from above by $(\varepsilon - \varepsilon')/2$ for all $\mathbf{x} \in B_{r_i}(\mathbf{x}_i) \cap \mathcal{X}$. Finally, the proof is completed by noting that $MSE(\hat{y}(\mathbf{x}_i)) \le \frac{\tau^2\Sigma_\epsilon(\mathbf{x}_i, \mathbf{x}_i)}{\tau^2 + \Sigma_\epsilon(\mathbf{x}_i, \mathbf{x}_i)} = \varepsilon'$ (see (3.9)).

Corollary 3.1. *Given a set of design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, let $\hat{y}(\mathbf{x})$ be the SK predictor constructed using (1.11). For any $\mathbf{x}_i \in \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ and $\varepsilon > 0$, if Assumption 1 holds and $n_i > \frac{V(\mathbf{x}_i)(\tau^2 - \varepsilon)}{\varepsilon\tau^2}$, then (a) $MSE(\hat{y}(\mathbf{x}_i)) < \varepsilon$; (b) there exists an $r_i > 0$ such that $MSE(\hat{y}(\mathbf{x})) < \varepsilon$ for all $\mathbf{x} \in B_{r_i}(\mathbf{x}_i) \cap \mathcal{X}$.*

As in the proof of Lemma 3.1, part (a) amounts to showing that $MSE(\hat{y}(\mathbf{x}_i))$ at a given design point $\mathbf{x}_i$ is bounded by the variance of the averaged intrinsic noises at $\mathbf{x}_i$. Consider that there are $k > 1$ design points in kriging model, a closer inspection of the proof of Theorem 2.2 shows that (3.8) still holds when the MSE is given in form of (1.12). In particular, let

$\hat{y}_{-i}(\mathbf{x})$ be the SK constructed from (1.11) by using the set of design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_k\}$, we have

$$\begin{aligned} MSE(\hat{y}(\mathbf{x}_i)) &= \frac{MSE(\hat{y}_{-i}(\mathbf{x}_i))\Sigma_\epsilon(\mathbf{x}_i, \mathbf{x}_i)}{MSE(\hat{y}_{-i}(\mathbf{x}_i)) + \Sigma_\epsilon(\mathbf{x}_i, \mathbf{x}_i)} \\ &\le \frac{\tau^2 \Sigma_\epsilon(\mathbf{x}_i, \mathbf{x}_i)}{\tau^2 + \Sigma_\epsilon(\mathbf{x}_i, \mathbf{x}_i)} < \varepsilon. \end{aligned}$$

To show part (b), we consider a prediction point $\mathbf{x}$ contained in the open ball $B_{r_i}(\mathbf{x}_i) \cap \mathcal{X}$ and write the MSE of the predictor at $\mathbf{x}$ as

$$\begin{aligned} &MSE(\hat{y}(\mathbf{x})) \\ &= \tau^2 - \Sigma_M(\mathbf{x}, \cdot)^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \Sigma_M(\mathbf{x}, \cdot) + \eta(\mathbf{x})^\intercal (\mathbf{1}_k^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \mathbf{1}_k)^{-1} \eta(\mathbf{x}) \\ &= \tau^2 - (\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot) + \Sigma_M(\mathbf{x}_i, \cdot))^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} (\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot) + \Sigma_M(\mathbf{x}_i, \cdot)) \\ &\qquad + (\eta(\mathbf{x}_i) - \mathbf{1}_k^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} (\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot)))^\intercal (\mathbf{1}_k^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \mathbf{1}_k)^{-1} \\ &\qquad\quad (\eta(\mathbf{x}_i) - \mathbf{1}_k^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} (\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot))) \\ &= \tau^2 - \Sigma_M(\mathbf{x}_i, \cdot)^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \Sigma_M(\mathbf{x}_i, \cdot) + \eta(\mathbf{x}_i)^\intercal (\mathbf{1}_k^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \mathbf{1}_k)^{-1} \eta(\mathbf{x}_i) \\ &\qquad - 2(\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot))^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \Sigma_M(\mathbf{x}_i, \cdot) \\ &\qquad - (\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot))^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} (\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot)) \\ &\qquad - 2\eta(\mathbf{x}_i)^\intercal (\mathbf{1}_k^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \mathbf{1}_k)^{-1} \mathbf{1}_k^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} (\Sigma_M(\mathbf{x}, \cdot) - \Sigma_M(\mathbf{x}_i, \cdot)) \end{aligned}$$

$$
\begin{aligned}
&+ (\Sigma_M(\mathbf{x},\cdot) - \Sigma_M(\mathbf{x}_i,\cdot))^\intercal \mathbf{U}_k (\Sigma_M(\mathbf{x},\cdot) - \Sigma_M(\mathbf{x}_i,\cdot)) \\
= {}& MSE(\hat{y}(\mathbf{x}_i)) - (\Sigma_M(\mathbf{x},\cdot) - \Sigma_M(\mathbf{x}_i,\cdot))^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} (\Sigma_M(\mathbf{x},\cdot) - \Sigma_M(\mathbf{x}_i,\cdot)) \\
&- 2(\Sigma_M(\mathbf{x},\cdot) - \Sigma_M(\mathbf{x}_i,\cdot))^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \Sigma_M(\mathbf{x}_i,\cdot) \\
&- 2\eta(\mathbf{x}_i)^\intercal (\mathbf{1}_k^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \mathbf{1}_k)^{-1} \mathbf{1}_k^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} (\Sigma_M(\mathbf{x},\cdot) - \Sigma_M(\mathbf{x}_i,\cdot)) \\
&+ (\Sigma_M(\mathbf{x},\cdot) - \Sigma_M(\mathbf{x}_i,\cdot))^\intercal \mathbf{U}_k (\Sigma_M(\mathbf{x},\cdot) - \Sigma_M(\mathbf{x}_i,\cdot)) \\
\le {}& MSE(\hat{y}(\mathbf{x}_i)) - 2(\Sigma_M(\mathbf{x},\cdot) - \Sigma_M(\mathbf{x}_i,\cdot))^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \Sigma_M(\mathbf{x}_i,\cdot) \\
&- 2\eta(\mathbf{x}_i)^\intercal (\mathbf{1}_k^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} \mathbf{1}_k)^{-1} \mathbf{1}_k^\intercal (\Sigma_M + \Sigma_\epsilon)^{-1} (\Sigma_M(\mathbf{x},\cdot) - \Sigma_M(\mathbf{x}_i,\cdot)) \\
&+ (\Sigma_M(\mathbf{x},\cdot) - \Sigma_M(\mathbf{x}_i,\cdot))^\intercal \mathbf{U}_k (\Sigma_M(\mathbf{x},\cdot) - \Sigma_M(\mathbf{x}_i,\cdot)), \qquad (3.13)
\end{aligned}
$$

where we have defined $\mathbf{U}_k = (\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{1}_k(\mathbf{1}_k^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}\mathbf{1}_k)^{-1}\mathbf{1}_k^\intercal(\Sigma_M + \Sigma_\epsilon)^{-1}$, and the last inequality holds because $(\Sigma_M + \Sigma_\epsilon)^{-1}$ is positive definite by Lemma 2.1. It is easy to observe that $\mathbf{U}_k$ is a symmetric positive semi-definite matrix. Therefore, the last term on the right-hand-side of (3.13) is bounded from above by $\lambda_{\mathbf{U}_k}\sum_{j=1}^{k}(\Sigma_M(\mathbf{x},\mathbf{x}_j) - \Sigma_M(\mathbf{x}_i,\mathbf{x}_j))^2$, where $\lambda_{\mathbf{U}_k} \ge 0$ is the largest eigenvalue of $\mathbf{U}_k$. The rest of the proof follows straightforwardly by invoking the continuity of $R(d,\boldsymbol{\theta})$ and the triangle inequality, and then employing a similar argument as given in the proof of Lemma 3.1; we omit the details.

Lemma 3.2. *Given a set of design points* $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$*, let* $\hat{y}^+(\mathbf{x})$ *be the SKG predictor constructed using (1.9). For any* $\mathbf{x}_i \in \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ *and* $\varepsilon > 0$*, if Assumptions 1, 2, 3, 4 hold and* $n_i > \frac{V(\mathbf{x}_i)(\tau^2-\varepsilon)}{\varepsilon\tau^2}$*, then (a)* $MSE(\hat{y}^+(\mathbf{x}_i)) < \varepsilon$*; (b) there exists an* $r_i > 0$ *such that* $MSE(\hat{y}^+(\mathbf{x})) < \varepsilon$ *for all* $\mathbf{x} \in B_{r_i}(\mathbf{x}_i) \cap \mathcal{X}$*.*

Proof. Proof of Lemma 3. Let $\hat{y}(\mathbf{x})$ be the SK predictor constructed using (1.9) based on $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. Lemma 3.1 shows that for any given $\varepsilon > 0$, $MSE(\hat{y}(\mathbf{x}_i)) < \varepsilon$ and there exists an $r_i > 0$ such that $MSE(\hat{y}(\mathbf{x})) < \epsilon$ for all $\mathbf{x} \in B_{r_i}(\mathbf{x}_i) \cap \mathcal{X}$. When gradient information is available, Theorem 2.3 shows that $\hat{y}^+(\mathbf{x})$ always improves the performance of $\hat{y}(\mathbf{x})$ in the sense that $MSE(\hat{y}^+(\mathbf{x})) \leq MSE(\hat{y}(\mathbf{x}))$. Consequently, we obtain $MSE(\hat{y}^+(\mathbf{x}_i)) < \varepsilon$ and $MSE(\hat{y}^+(\mathbf{x})) < \varepsilon$ for all $\mathbf{x} \in B_{r_i}(\mathbf{x}_i) \cap \mathcal{X}$. □

Corollary 3.2. *Given a set of design points* $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$*, let* $\hat{y}^+(\mathbf{x})$ *be the SKG predictor constructed using (1.11). For any* $\mathbf{x}_i \in \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ *and* $\varepsilon > 0$*, if Assumptions 1, 2, 3, 4 hold and* $n_i > \frac{V(\mathbf{x}_i)(\tau^2-\varepsilon)}{\varepsilon\tau^2}$*, then (a)* $MSE(\hat{y}^+(\mathbf{x}_i)) < \varepsilon$*; (b) there exists an* $r_i > 0$ *such that* $MSE(\hat{y}^+(\mathbf{x})) < \varepsilon$ *for all* $\mathbf{x} \in B_{r_i}(\mathbf{x}_i) \cap \mathcal{X}$*.*

Proof. Proof of Corollary 4. Follows directly from Corollary 3.1 and Theorem 2.4. □

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 very 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 theorem:

Theorem 3.1. *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, 2, 3, and 4 hold and the number of simulation replications $n_i > \frac{V(\mathbf{x}_i)(\tau^2|\mathcal{X}|-\varepsilon)}{\varepsilon\tau^2}$ for all i, then $\lim_{k\to\infty} IMSE(k) \leq \varepsilon$.*

Proof. Proof of Theorem 5. We only consider the SK framework when the predictor $\hat{y}_k(\mathbf{x})$ is obtained using (1.11); the same result can be seen to hold in the SKG case by applying Theorem 2.4.

For a given sequence $\{n_i\}$, the sequence of design points $\mathbf{x}_1, \mathbf{x}_2, \ldots$ is completely determined by the ASK method. From Theorem 2.2, it is easy to see that the sequence of IMSEs of the optimal predictors $\hat{y}_k(\mathbf{x})$ is monotonically non-increasing. Therefore, it follows from the monotone convergence theorem that $\lim_{k\to\infty} IMSE(k) = c$ for some constant $c \geq 0$.

We now proceed by contradiction and assume $\lim_{k\to\infty} IMSE(k) = c > \varepsilon$. This hypothesis, together with the monotonicity of $MSE(\hat{y}_k(\mathbf{x}))$ (Theorem 2.2), implies there exists a subset $\Omega \subseteq \mathcal{X}$ with positive Lebesgue measure and non-empty interior (due to the continuity of MSE) such that

$$MSE(\hat{y}_k(\mathbf{x})) > \frac{c+\varepsilon}{2|\mathcal{X}|} \quad \forall\, \mathbf{x} \in \Omega \text{ and } \forall\, k. \tag{3.14}$$

Let $\mathbf{x}^* \in \Omega$ be such that $B_r(\mathbf{x}^*) \subseteq \Omega$ for some $r > 0$. Clearly, $\mathbf{x}^*$ has never been generated by the algorithm, i.e, $\mathbf{x}^* \notin \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$; otherwise the condition that $n_i > \frac{V(\mathbf{x}_i)(\tau^2|\mathcal{X}|-\varepsilon)}{\varepsilon\tau^2}$ and Corollary 3.1 part (a) would lead to $MSE(\hat{y}_k(\mathbf{x}^*)) < \frac{\varepsilon}{|\mathcal{X}|} < \frac{c+\varepsilon}{2|\mathcal{X}|}$, which contradicts (3.14).

Fix an iteration $k \geq 1$, and let $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ be the set of current design points. Let $\hat{y}_k^*(\mathbf{x})$ be the optimal SK predictor constructed using the design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{x}^*\}$ with $n^* > \frac{V(\mathbf{x}^*)(\tau^2|\mathcal{X}|-\varepsilon)}{\varepsilon\tau^2}$ simulation replications allocated to $\mathbf{x}^*$. We have from Corollary 3.1 that there exists an r' such that $MSE(\hat{y}_k^*(\mathbf{x})) < \frac{\varepsilon}{|\mathcal{X}|}$ for all $\mathbf{x} \in B_{r^*}(\mathbf{x}^*)$, where $r^* = \min\{r, r'\}$.

Since $\lim_{k\to\infty} IMSE(k) = c$, there exist an $N' > 0$ such that $c \leq IMSE(k') \leq c + \frac{c-\varepsilon}{4|\mathcal{X}|}|B_{r^*}(\mathbf{x}^*)|$ whenever $k' \geq \max\{N', k\}$. Given the set of design points generated by ASK up to the k'th iteration, i.e., $\{\mathbf{x}_1, \ldots, \mathbf{x}_{k'}\}$, let $\mathbf{x}_{k'+1}$ be the next design point chosen by the algorithm and $\hat{y}^*_{k'+1}(\mathbf{x})$ be the optimal SK predictor constructed using points $\{\mathbf{x}_1, \ldots, \mathbf{x}_{k'}, \mathbf{x}^*\}$ with n^* simulation replications at $\mathbf{x}^*$. Note that since $\{\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{x}^*\} \subseteq \{\mathbf{x}_1, \ldots, \mathbf{x}_{k'}, \mathbf{x}^*\}$, a repeated application of the monotonicity property Theorem 2.2 indicates that

$$MSE(\hat{y}^*_{k'+1}(\mathbf{x})) \leq MSE(\hat{y}^*_k(\mathbf{x})) < \frac{\varepsilon}{|\mathcal{X}|} \;\; \forall \mathbf{x} \in B_{r^*}(\mathbf{x}^*). \tag{3.15}$$

Let $IMSE^*(k'+1)$ be the integrated MSE of $\hat{y}^*_{k'+1}(\mathbf{x})$. Since $\mathbf{x}_{k'+1}$ minimizes $IMSE(k'+1)$, we have

$$\begin{aligned}
IMSE(k'+1) &\leq IMSE^*(k'+1) \\
&= \int_{B_{r^*}(\mathbf{x}^*)} MSE(\hat{y}^*_{k'+1}(\mathbf{x}))d\mathbf{x} + \int_{\mathcal{X}\backslash B_{r^*}(\mathbf{x}^*)} MSE(\hat{y}^*_{k'+1}(\mathbf{x}))d\mathbf{x} \\
&< \frac{\varepsilon|B_{r^*}(\mathbf{x}^*)|}{|\mathcal{X}|} + \int_{\mathcal{X}\backslash B_{r^*}(\mathbf{x}^*)} MSE(\hat{y}_{k'}(\mathbf{x}))d\mathbf{x} \quad \text{by (3.15) and Theorem 2.2} \\
&= \frac{\varepsilon|B_{r^*}(\mathbf{x}^*)|}{|\mathcal{X}|} + IMSE(k') - \int_{B_{r^*}(\mathbf{x}^*)} MSE(\hat{y}_{k'}(\mathbf{x}))d\mathbf{x} \\
&< \frac{\varepsilon|B_{r^*}(\mathbf{x}^*)|}{|\mathcal{X}|} + c + \frac{c-\varepsilon}{4|\mathcal{X}|}|B_{r^*}(\mathbf{x}^*)| - \frac{c+\varepsilon}{2|\mathcal{X}|}|B_{r^*}(\mathbf{x}^*)| \quad \text{by (3.14)} \\
&= c - \frac{c-\varepsilon}{4|\mathcal{X}|}|B_{r^*}(\mathbf{x}^*)|
\end{aligned}$$

$$< c.$$

This contradicts the fact $\lim_{k\to\infty} IMSE(k) = c$. Hence, we must have $\lim_{k\to\infty} IMSE(k) \leq \varepsilon$. □

So far, we have introduced all of our main theoretical results. In the next section, we will justify the monotonicity property of SK/SKG metamodels and the validity of ASK by running the model fitting process on several numerical examples.

4 Numerical Experiments and Comparisons

4.1 Introduction

The main work in this section is to test our theoretical results in Section 2 and Subsection 3.4. In the first place, the monotonicity properties of the MSE estimators in the SK/SKG metamodels is tested through numerical examples. We show that under the setting of fixed model parameters, the IMSE estimator of the SK/SKG predictor is a monotone decreasing function of the number of design points, which is theoretically proved in the theorems and corollaries of Section 2. In addition, we also test the effectiveness of the proposed ASK procedure. It is interesting to see how the true uncertainty measure changes when the model parameters are updated in each iteration.

To illustrate the monotonicity properties and the effectiveness of ASK, we prepared two sets of examples: an M/M/1 queue and four deterministic functions with added noise. In all of the numerical experiments, the SK (SKG) predictors are constructed using an (unknown) constant trend model with a Gaussian correlation function $R_M(d(\mathbf{x}_i, \mathbf{x}_j), \theta) = \exp(-\theta(\mathbf{x}_i - \mathbf{x}_j)^2)$. The variance functions V and V_r are fitted using the ordinary kriging model assuming the same model structure. We will introduce the two sets of examples before providing further details of the experiments design.

4.1.1 Example A: M/M/1 queue

This example is taken from [13]. 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. Clearly, the elementary queueing theory shows that $f(x) = \frac{x}{1-x}$. 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 $\bar{f}(x) = \frac{1}{t}\int_0^t N_s(x)ds$ by performing a single (but very long) simulation run, where N_s is the number of observed customers in system at time s. The variance of the estimator can be approximated by $Var[\bar{f}(x)] \approx \frac{2x(1+x)}{t(1-x)^4}$ when t is large (see, e.g., [13]).

4.1.2 Example B: Deterministic Functions with Added Noise

The following benchmark functions, which have been previously studied in e.g., [42] and [43], are used in our experiments:

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

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

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

(4) $y(\mathbf{x}) = Y(\mathbf{x}) + \epsilon(\mathbf{x})$, $\mathbf{x} = (x_1, x_2)^\intercal \in [-4,4] \times [-4,4]$, where $Y(\mathbf{x}) = 1 + \frac{x_1^2}{4000} + \frac{x_2^2}{4000} - cos(x_1) cos(\frac{x_2}{\sqrt{(2)}})$ and $\epsilon(\mathbf{x}) \sim \mathcal{N}(0, V(\mathbf{x}))$.

In all examples, we assume that noisy gradient estimates $\mathcal{D}_j^r(\mathbf{x}) = \frac{\partial Y(\mathbf{x})}{\partial x_r} + \zeta_j^r(\mathbf{x})$ are available at $\mathbf{x}$ on the jth simulation replication, where $\zeta_j^r(\mathbf{x}) \sim \mathcal{N}(0, V(\mathbf{x}))$ for $r = 1, \ldots, d$ and d is the problem dimension. Note that for simplicity, the same variance function $V(\mathbf{x})$ is used for both ϵ and ζ.

4.2 Test the Monotonicity Properties of the MSE Estimator in SK/SKG Metamodel

To quantify the overall quality of a SK/SKG predictor over the entire domain, we use the IMSE (3.3) as the measure of performance. We consider the following simple sequential version of a space-filling scheme based on quasi-Monte Carlo sampling to test the monotonicity properties.

Step 0: Specify the total number of design points N, a set of initial space-filling design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ over $\mathcal{X}(k < N)$, and the number of simulation replications n_0 at each design point.

Step 1: Collect output performance measures (including gradient estimator when using the SKG) at each $\mathbf{x}_i$. Fit an initial SK/SKG model as discussed in Ankenman, Nelson, and Staum (2010) and fix the parameters of the model.

Step 2: Choose a new design point $\mathbf{x}_{k+1}$ based on quasi-Monte Carlo sampling. Perform n_0 independent simulation runs at $\mathbf{x}_{k+1}$ and collect output performance measures (including gradient performance measure when using SKG). Compute the IMSE of the SK/SKG predictor with $k+1$ design points.

Step 3: If the current number of design points exceeds N, then terminate; otherwise set $k = k + 1$ and go to **step 2**.

Note that in all examples of this section, we set $N = 15$, the number of initial points to 5, and the number of simulation replications $n_0 = 30$. We also run 30 simulation replications to estimate the gradient information in deterministic examples with added noise. Also, 30 replicates of the above procedure are run to ensure the influence of variation will be minimized. To examine the performance of IMSE, we plot the mean IMSE (averaged over

30 independent runs) versus the number of added design points.

4.2.1 Performance of Monotonicity in M/M/1 Queue Example

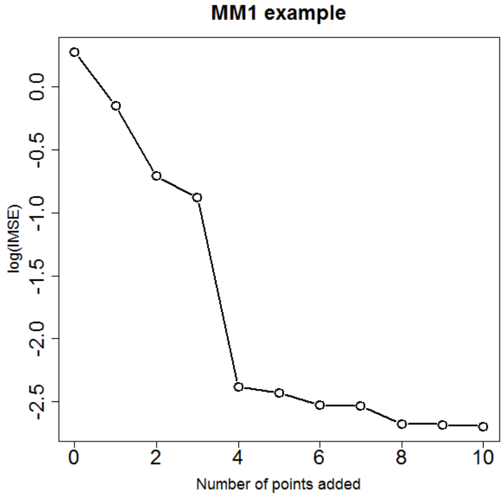


Figure 3: Validate the monotonicity property in M/M/1 queueing example with SK metamodel

In Figure 3, the mean IMSE in logarithm scale is plotted against the number of added design points. Apparently, the IMSE measure monotonically decreases as the number of design points increases, which conforms well with our expectation based on the established theoretical proofs. Notice that since these results are based on space-filling designs, the difference between two consecutive IMSEs does not achieve the maximum amount of reduction. If a more sophisticated selection of design points is applied, e.g., the selection criteria in ASK procedure, better performance shall be expected.

4.2.2 Performance of Monotonicity in the Deterministic Examples

Both SK and SKG metamodels are applied for each deterministic example in this section. Moreover, we consider two types of the intrinsic noise variance functions: (i) $V(\mathbf{x}) = 1$, and (ii) $V(\mathbf{x}) = 2(\mathbf{x}^{\intercal}\mathbf{x} + 1)$

In each of the test examples, we can see from Figure 4 to Figure 7 that the mean IMSE curve (of log scale) is monotonically decreasing as the number of new design points increases. Notice that in each plot, although the two mean IMSE curves show similar shapes and trends, curve obtained with type II noise consistently has larger IMSEs compare to those obtained with type I noise. This suggests that the intrinsic variance probably plays a crucial rule on the prediction performance of SK models when design points are fixed. Besides, by comparing the two plots in each figure, we can see that the mean IMSE obtained in SKG model is smaller than that obtained in SK model when they use the same type of intrinsic noise. It indicates that the MSE/IMSE of the predictor can be decreased when the gradient information is available, which conforms with the results in Theorem 2.3 and 2.4.

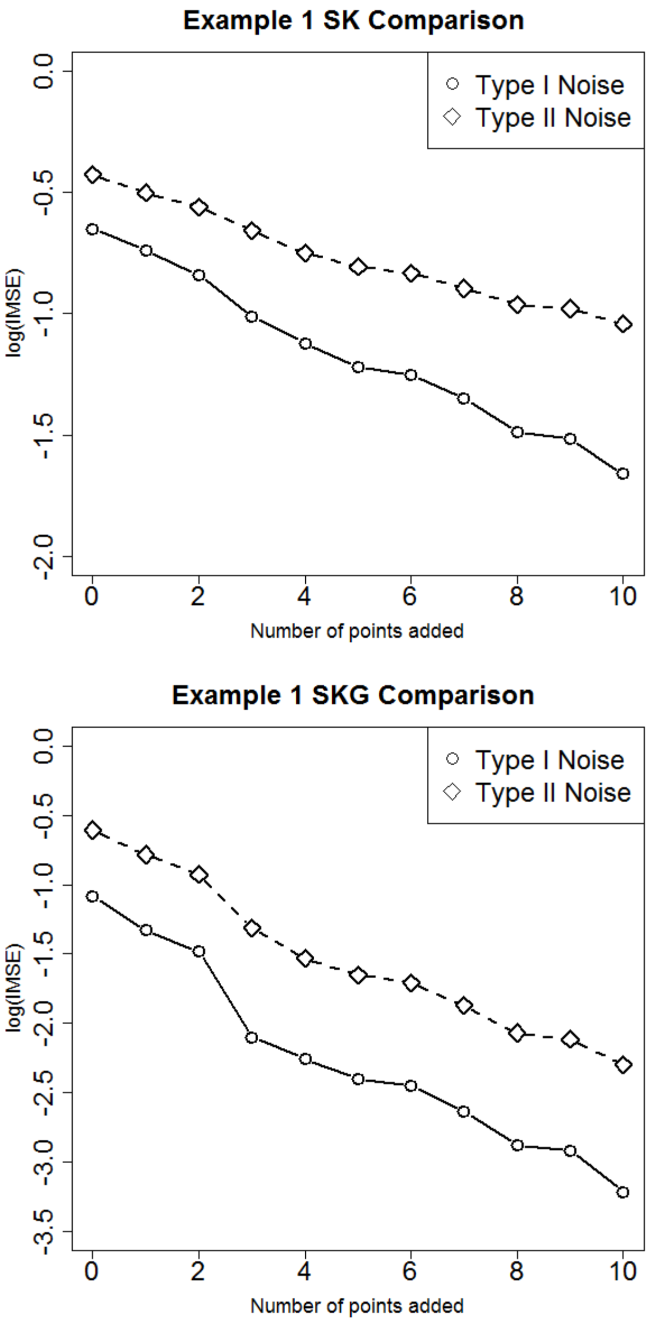


Figure 4: Validate the monotonicity property of Example 1 with SK and SKG metamodels

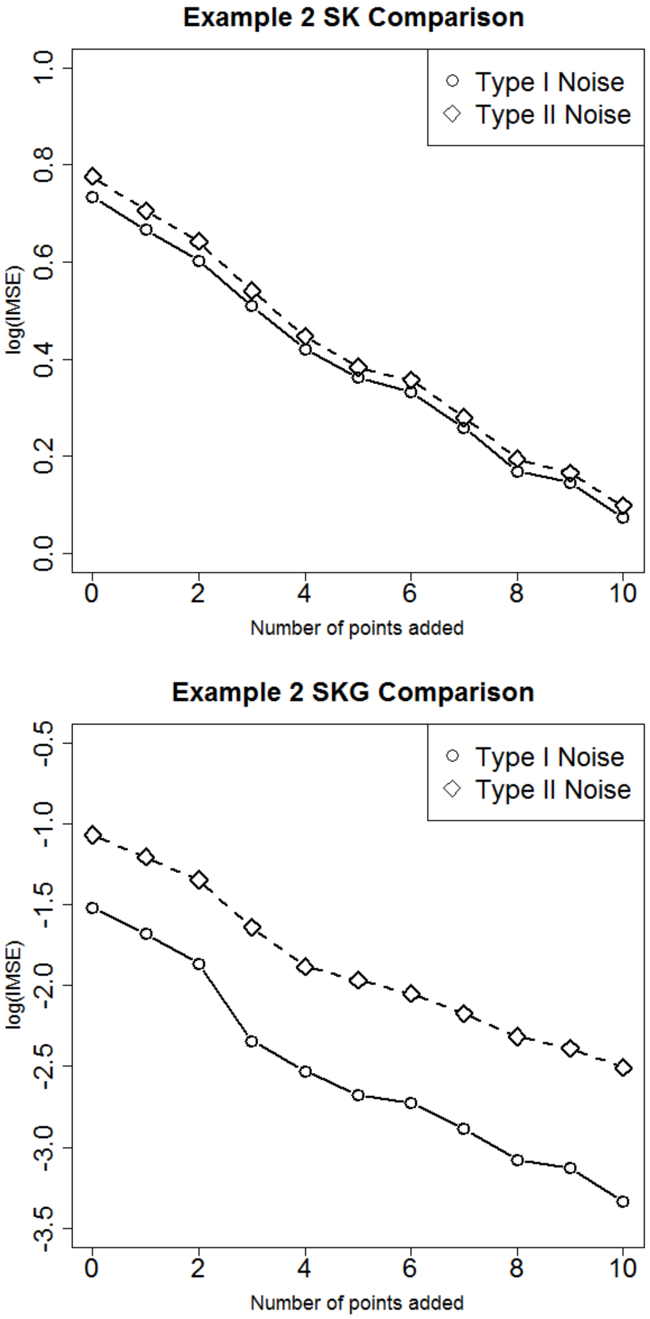


Figure 5: Validate the monotonicity property of Example 2 with SK and SKG metamodels

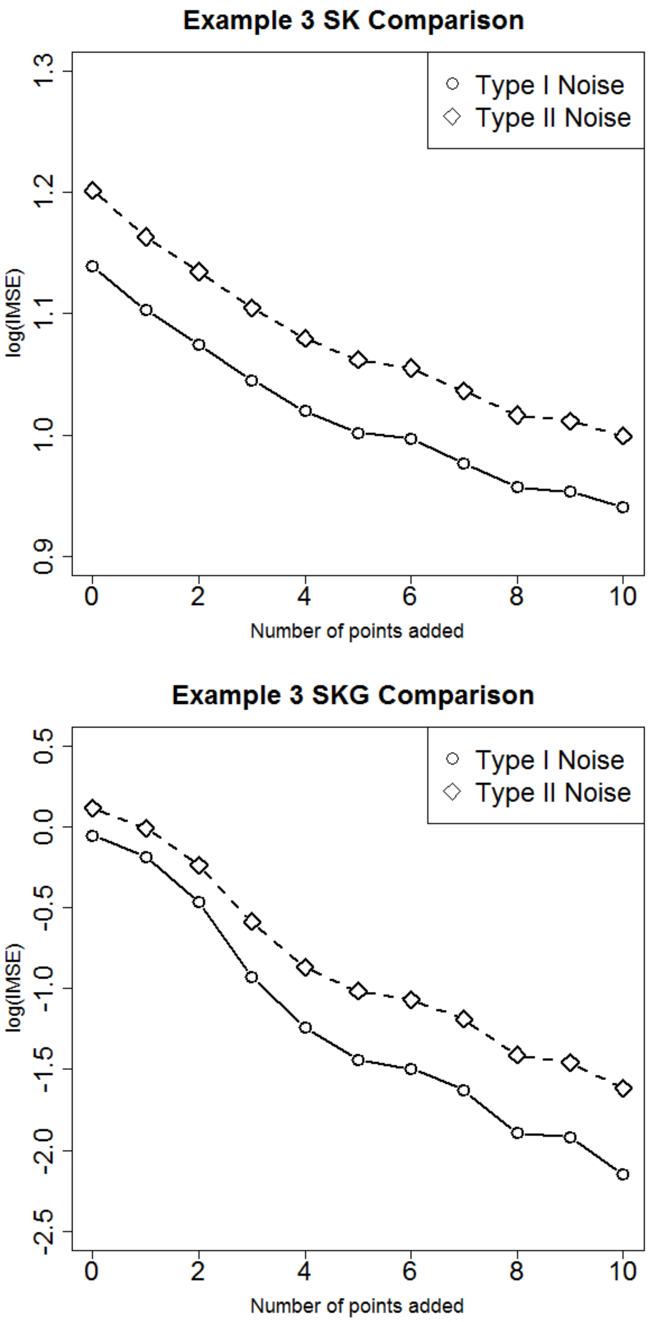


Figure 6: Validate the monotonicity property of Example 3 with SK and SKG metamodels

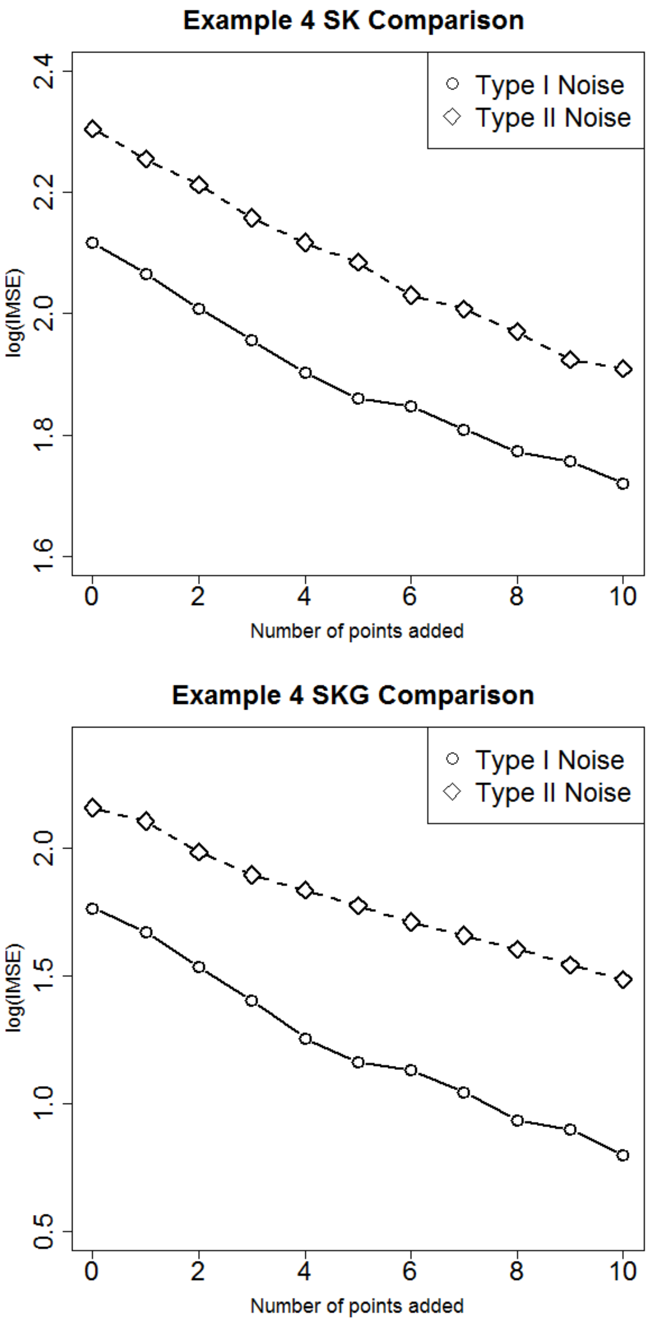


Figure 7: Validate the monotonicity property of Example 4 with SK and SKG metamodels

4.3 Test the Effectiveness of the ASK procedure

To illustrate the effectiveness of the ASK procedure, we also consider the M/M/1 queue example and the deterministic examples. In both cases, the performance of ASK is compared with those of the sequential MSE (SMSE) approach and the AEES approach which are introduced in Section 3.2. We use the same simulation replication rule in SMSE and ASK. The global optimization algorithm proposed in [44] is used to solve the new point selection problem in both SMSE and ASK procedure.

Since the true response curves of the test functions are known, the true average integrated square error (AISE) is employed to test the practical accuracy of the predictor in this section, defined by

$$AISE = \frac{\int_{x \in \mathcal{X}} (f(\mathbf{x}) - \hat{y}(\mathbf{x}))^2 d\mathbf{x}}{|\mathcal{X}|} \tag{4.1}$$

The AISE is a measure aiming at compareing the true response value with the estimated response value in a metamodel predictor. The integral in AISE is estimated using standard Gaussian quadrature in our implementation.

4.3.1 Comparisons in the M/M/1 Queueing Example

In the implementation of ASK for M/M/1 example, the following settings are used:

- IMSE target: $\varepsilon = 0.1|\mathcal{X}|$
- Initial space-filling design points: 5 points over $[0.05, 0.95]$
- Independent simulation runs on initial design points: 30
- Length of time units: $t = 1000$ (assuming that $\bar{f}(x)$ is unbiased with the large t)

For each of the three procedures, 30 independent replicates are run. All the comparisons about the three algorithms are based on the same number of design points.

Figure 8 shows the mean AISE (in log scale) obtained in each procedure as a function of the number of added design points. Figure 9 plots the mean response surface (measured in log scale) predicted by the SK predictor when the procedure is terminated for all three algorithms. As it is shown in Figure 8, the sequential procedure is stopped when the number of design points reaches 13. The means and standard errors of AISE values obtained from

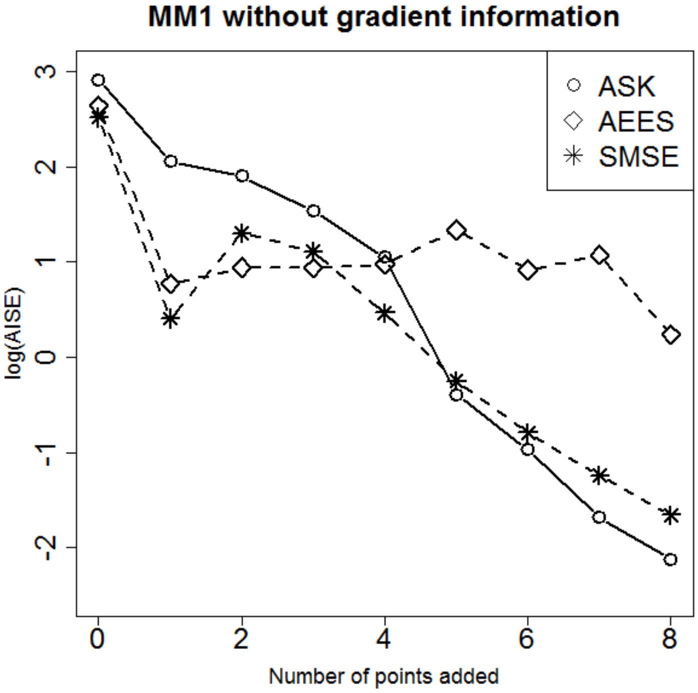


Figure 8: Sequential AISEs obtained by ASK, AEES and SMSE procedures in M/M/1 queueing example

30 independent runs are listed in Table 1. The total number of simulation replications associated with each procedure is provided in Table 2.

It is shown in Figure 8 that both AEES and SMSE receive a larger initial reduction in AISE than ASK. A possible reason behind this is that the initial estimates of model parameters are not accurate, making ASK not able to locate the point that achieves the maximum reduction in the true IMSE. However, as the number of design points increases, the estimates of model

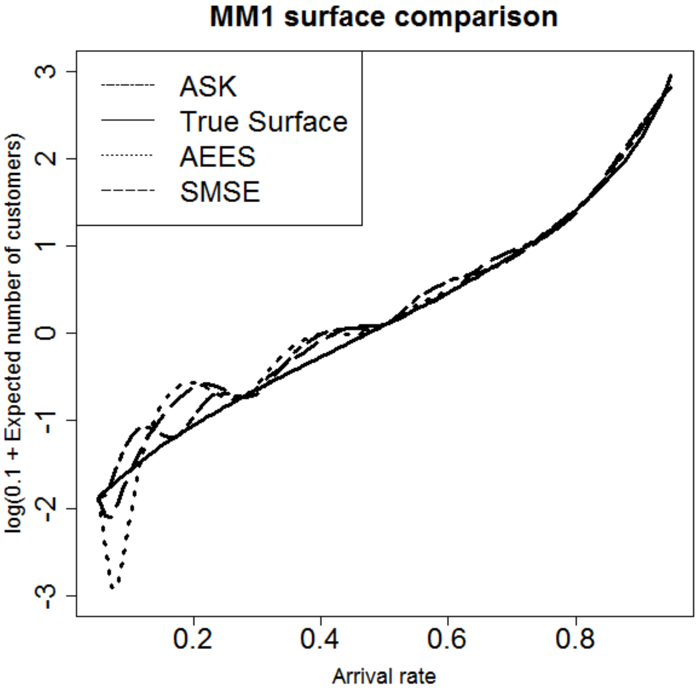


Figure 9: The final fitted response surfaces obtained by ASK, AEES and SMSE procedures for the M/M/1 queueing example

parameters are more accurate and thus making the AISE of ASK consistently decreasing. After adding 5 design points, the performance of ASK is superior to AEES's and SMSE's. Figure 8 shows that, among all the three sequential procedures, only the mean AISE curve of ASK is monotonically decreasing. There is a temporary increase in SMSE with 2 added design points. A long period of irregular rising and falling are shown in the curve of AEES. We believe that it is because the model parameters are constantly estimated and updated in each iteration based on the current information. Since the selection rule of ASK is derived to maximize the reduction in estimated IMSE, it also tends to minimize the true IMSE. However, both SMSE and AEES are not designed to optimize IMSE. So, it is no wonder that updating the model parameter results in a temporary increase in the actual AISE estimated by SMSE and AEES.

In the final output, ASK achieves the minimum AISE, followed by SMSE and then AEES. It is worth mentioning that both ASK and SMSE share the same simulation replication rules. The difference lies in the way of choosing new design points. It can be seen from Table 1 and 2 that ASK achieves a smaller AISE and less number of simulation replications than SMSE when they terminate. Therefore, although simulation replication relocation rule is of great importance, a sophisticated selection of design points is crucial since it may help significantly elevate the quality of the predictor in SK models.

Table 1: AISE (mean ± standard error) obtained by ASK, AEES, and SMSE on the M/M/1 queueing example. All results are based on 30 independent runs.

ASK	AEES	SMSE
0.12 ± 0.02	1.27 ± 0.39	0.19 ± 0.02

Table 2: The number of function evaluations (mean ± standard error) obtained by ASK, AEES, and SMSE on the M/M/1 queueing example. All results are based on 30 independent runs.

ASK	AEES	SMSE
894.43 ± 26.41	1193.32 ± 23.34	1011.12 ± 46.22

4.3.2 Comparisons in the Deterministic Examples

When implementing ASK, AEES and SMSE in both SK and SKG settings, we have the following settings:

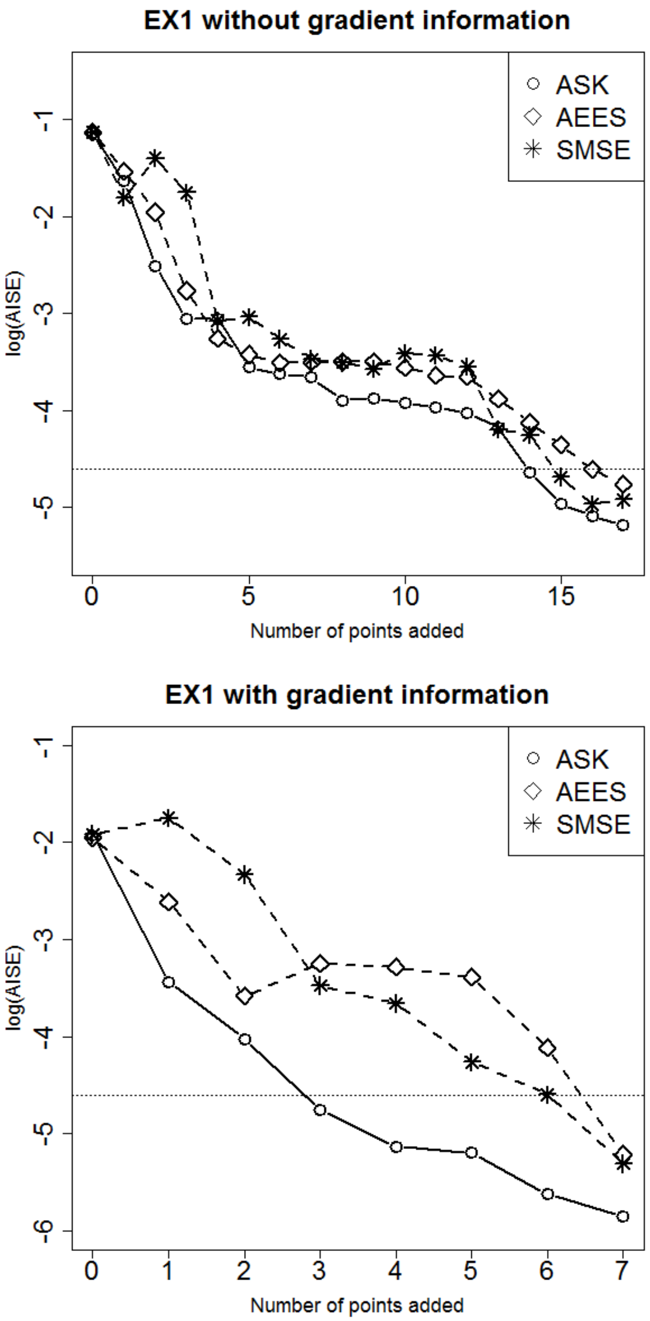
- Type of intrinsic noise variance: $V(\mathbf{x}) = 0.1|f(\mathbf{x})|$
- IMSE target: $\epsilon = 0.01|\mathcal{X}|$ (AISE target 0.01)
- Number of initial space-filling design points: 10 under SK model and 5 under SKG model

We use the AISE as a stopping criterion and continue running each algorithm before the AISE drops below 0.01. The sequential performance of these algorithms are represented by the mean AISE curves (in log scale) and are plotted in Figure 10 to Figure 13. The obtained AISE and the number of additional design points required by all three algorithms are listed in Table 3. The number of simulation replications for each test example when the algorithm attains the desired AISE level is reported in Table 4.

We find that the AISEs obtained by the three algorithms are reasonably close to the prescribed AIMSE threshold in all cases. In particular, the actual mean AISEs calculated by ASK are smaller than 0.01 in most cases and those AISE values that are larger than the AIMSE threshold are more likely to occur when the number of design points is small. Therefore, we suggest that as long as the number of design points is not too small, there is no significant difference in using estimated IMSE as a stopping criterion to evaluate the performance of predictor.

Based on the plots and tables, the ASK procedure is able to attain the target AISE level with fewer numbers of design points compared with the two competing methods in all test cases. In particular, under the SK framework, ASK demonstrates superior performances over AEES and SMSE in terms of both the number of simulation replications required and the number of design points used. In the scenarios when gradient information is employed, ASK generally demonstrates similar or better performance compared with AEES and SMSE. In addition, we see that under the SKG framework, the number of design points required by ASK to attain the AISE target is dramatically smaller that in the SK cases (8 to 12 in SKG v.s. 21 to 27 in SK). The result explicitly demonstrates the advantage of incorporating gradient information for prediction performance enhancement of stochastic kriging metamodels.

To conclude, we claim that the proposed ASK procedure outperforms the other two algorithms since ASK reaches the target accuracy threshold with the least number of design points and the almost the least number of simulation replications in almost all test cases.



EX1 without gradient information

EX1 with gradient information

Figure 10: Sequential logarithmic scaled AISE comparisons obtained by ASK, AEES and SMSE procedures for deterministic example 1

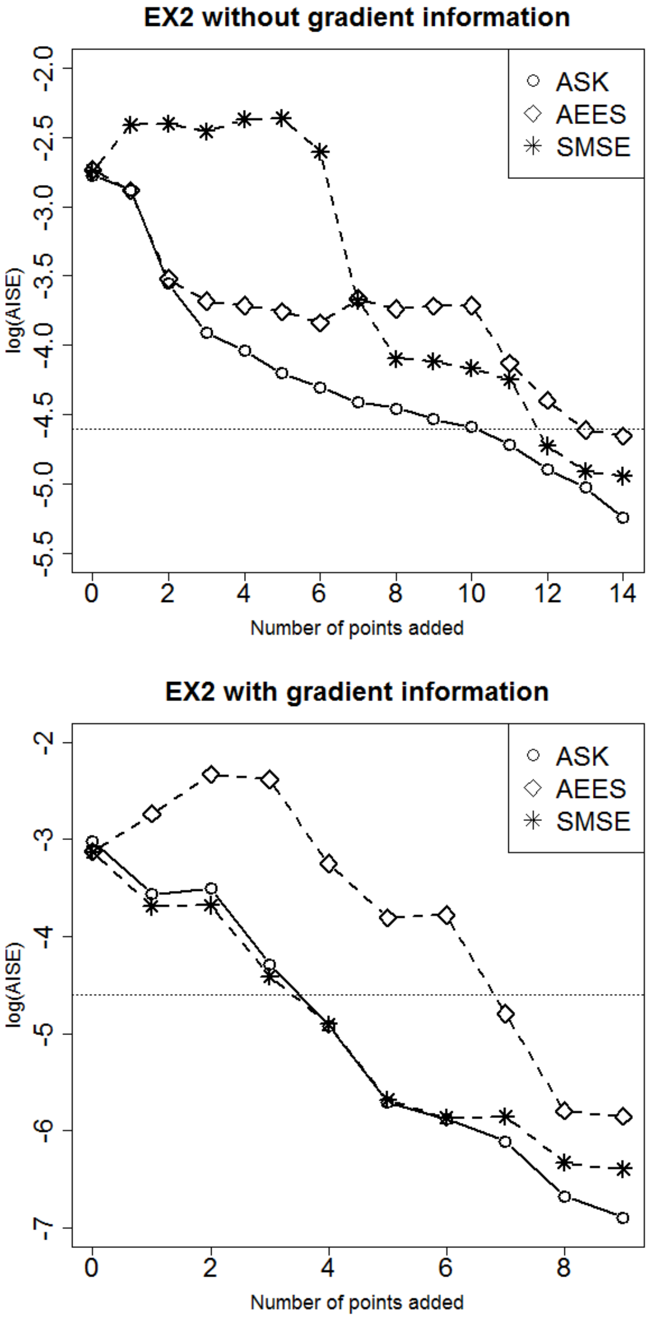


Figure 11: Sequential logarithmic scaled AISE comparisons obtained by ASK, AEES and SMSE procedures for deterministic example 2

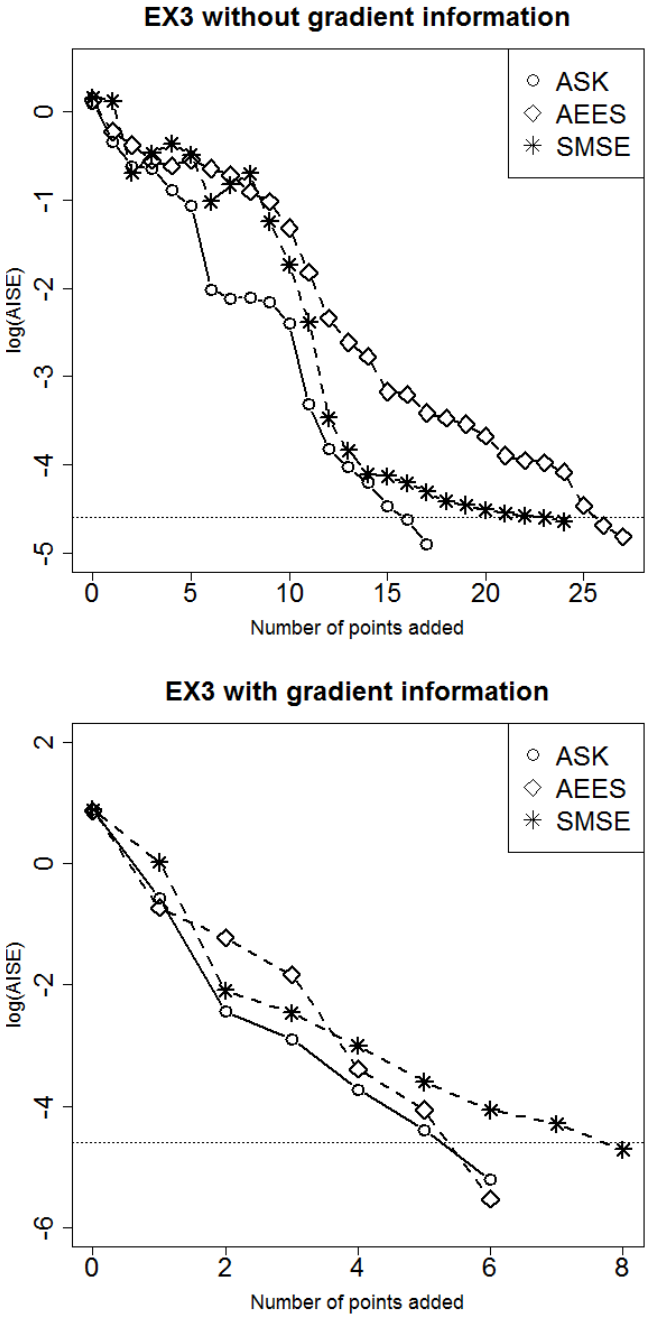


Figure 12: Sequential logarithmic scaled AISE comparisons obtained by ASK, AEES and SMSE procedures for deterministic example 3

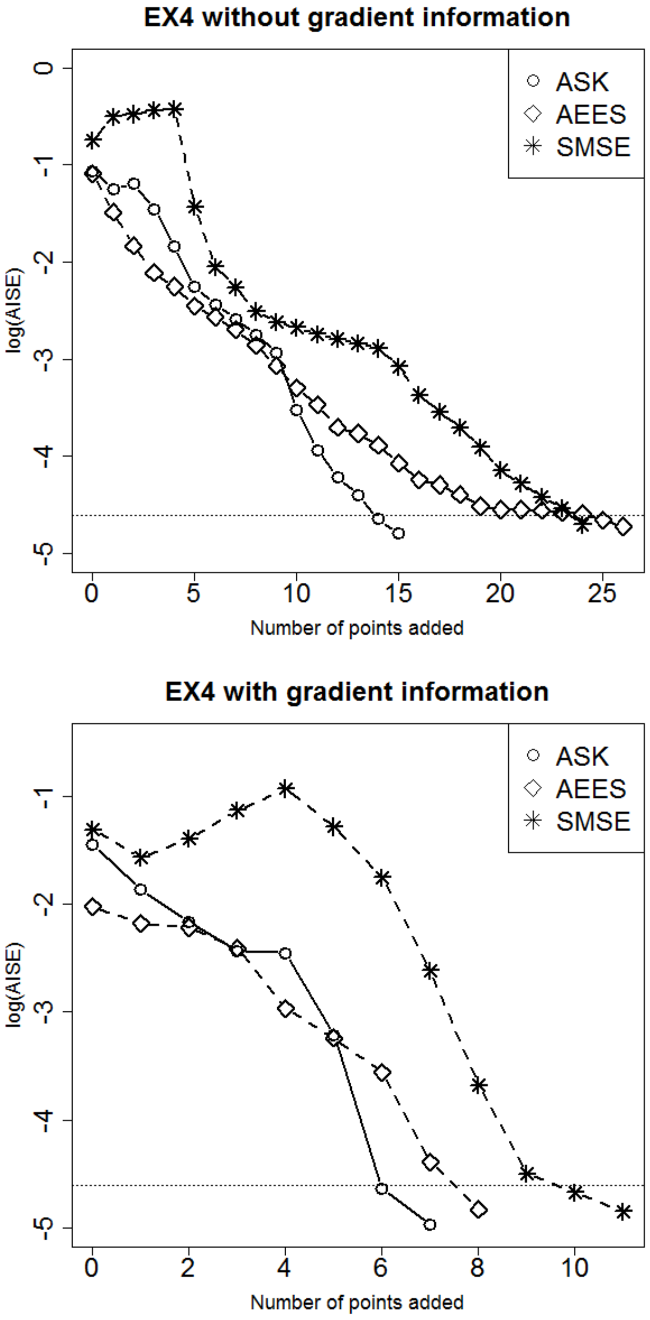


Figure 13: Sequential logarithmic scaled AISE comparisons obtained by ASK, AEES and SMSE procedures for deterministic example 4

Table 3: AISE (mean ± standard error)(Number of added points) obtained by ASK, AEES, and SMSE when the estimated AIMSE reaches 0.01 under both SK and SKG frameworks. All results are based on 30 independent runs.

Test Frwk.	Test Fcn.	ASK	AEES	SMSE
SK	Ex.(1)	2.1e-2 ± 2.4e-3 (10)	2.0e-2 ± 1.1e-3 (13)	2.8e-2 ± 2.3e-3 (10)
	Ex.(2)	1.2e-2 ± 1.7e-3 (7)	2.6e-2 ± 3.7e-3 (7)	7.0e-3 ± 3.3e-4 (15)
	Ex.(3)	8.9e-3 ± 6.4e-4 (17)	6.4e-3 ± 4.0e-4 (29)	1.2e-2 ± 2.7e-3 (20)
	Ex.(4)	7.4e-3 ± 5.5e-4 (16)	8.9e-3 ± 6.5e-4 (26)	8.4e-3 ± 4.8e-4 (25)
SKG	Ex.(1)	5.9e-3 ± 2.5e-4 (4)	3.7e-3 ± 5.5e-4 (8)	1.0e-2 ± 1.7e-3 (6)
	Ex.(2)	3.0e-2 ± 5.8e-4 (2)	2.3e-2 ± 1.5e-3 (6)	2.5e-2 ± 3.8e-3 (3)
	Ex.(3)	5.4e-3 ± 1.1e-3 (6)	1.8e-3 ± 2.9e-4 (8)	8.9e-3 ± 9.4e-4 (8)
	Ex.(4)	6.5e-3 ± 4.4e-4 (7)	5.4e-3 ± 6.2e-4 (9)	9.3e-3 ± 8.5e-4 (10)

Table 4: Total number of simulation replications (mean ± standard error)(Number of added points) required by ASK, AEES, and SMSE to reach an AISE of at least 0.01 under both SK and SKG frameworks. All results are based on 30 independent runs.

Test Frwk.	Test Fcn.	ASK	AEES	SMSE
SK	Ex.(1)	450.43 ± 5.74 (15)	4710 ± 23.01 (17)	810 ± 4.53 (16)
	Ex.(2)	355.83 ± 6.32 (11)	2240.3 ± 90.00 (14)	384.83 ± 5.46(12)
	Ex.(3)	575.9 ± 12.40 (17)	1878.4 ± 76.56 (27)	734.43 ± 16.15 (24)
	Ex.(4)	468.27 ± 8.34 (15)	1303.33 ± 26.87 (26)	540.83 ± 9.00 (24)
SKG	Ex.(1)	175.6 ± 2.54 (3)	817.4 ± 38.23 (7)	556.43 ± 20.12 (7)
	Ex.(2)	191.45 ± 2.76 (4)	1392.6 ± 2.14(7)	168.3 ± 1.13(4)
	Ex.(3)	220.34 ± 5.80 (6)	730.64 ± 5.03 (6)	302.43 ± 7.45(8)
	Ex.(4)	235.3 ± 7.7 (7)	771.43 ± 20.3 (8)	251.32 ± 4.54(11)

5 Summary and Future Work

In our research, we focus on investigating the performance of SK/SKG metamodels in a sequential setting. To do this, we first conducted theoretical analyses about how the MSE of SK/SKG predictor is changing in a sequential procedure. Our findings indicate that the MSE measure is monotonically non-increasing as we keep adding the number of design points when the model parameters are either fixed or given. In addition, under the same parameters setting, the availability of gradient information can lead to a significant decrease in the MSE measure. With these findings, we can not only complement existing results in the (stochastic) kriging literature, but also develop sequential sampling procedures under both SK and SKG frameworks. Motivated by this work, we designed a novel adaptive sequential kriging method for adaptively selecting the design point and simulation replications. We also developed the theory to justify our proposed algorithm when all model parameters are known. Our numerical experiments not only justified the monotonicity properties of the MSE estimator of SK/SKG predictor, but also indicate the superiority of ASK among several existing methods in terms of achieving the same level of accuracy with fewer design points and number of simulation replications.

Our work can be extended in the following directions. First of all, our theoretical analyses are performed under the assumption that the model parameters are given. We expect to obtain similar results when the model parameters are updated in each iteration. To do this, more complicated statistical analyses regarding to the updated information about model parameters may be involved. Another promising research topic is to develop sequential optimization algorithm using the SK/SKG framework. Successful methods like EGO developed such sequential optimization algorithm based on ordinary kriging. Compared with ordinary kriging, SK/SKG is a more powerful tool that can accurately approximate the response surface in stochastic framework. Therefore, with our theoretical analyses of SK/SKG metamodels, we expect to combine them with existing stochastic simulation optimization algorithms in the future.

References

[1] Lei Gu. A comparison of polynomial based regression models in vehicle safety analysis. In *ASME Design Engineering Technical Conferences, ASME Paper No.: DETC/DAC-21083*, 2001.

[2] G Gary Wang and S Shan. Review of metamodeling techniques in support of engineering design optimization. *Journal of Mechanical Design*, 129(4):370–380, 2007.

[3] Mario Chica-Olmo and Juan A Luque-Espinar. Applications of the local estimation of the probability distribution function in environmental sciences by kriging methods. *Inverse Problems*, 18(1):25, 2002.

[4] S Bocchi, A Castrignano, F Fornaro, and T Maggiore. Application of factorial kriging for mapping soil variation at field scale. *European Journal of Agronomy*, 13(4):295–308, 2000.

[5] Tomislav Hengl, Gerard BM Heuvelink, and Alfred Stein. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, 120(1):75–93, 2004.

[6] A Marsh LaVenue and John F Pickens. Application of a coupled adjoint sensitivity and kriging approach to calibrate a groundwater flow model. *Water Resources Research*, 28(6):1543–1569, 1992.

[7] Bin Zou, J Gaines Wilson, F Benjamin Zhan, and Yongnian Zeng. An emission-weighted proximity model for air pollution exposure assessment. *Science of the total environment*, 407(17):4939–4945, 2009.

[8] Georges Matheron. Principles of geostatistics. *Economic geology*, 58(8):1246–1266, 1963.

[9] Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.

[10] Thomas J Santner, Brian J Williams, and William I Notz. *The design and analysis of computer experiments*. Springer Science & Business Media, 2013.

[11] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[12] Deng Huang, TT Allen, WI Notz, and RA Miller. Sequential kriging optimization using multiple-fidelity evaluations. *Structural and Multidisciplinary Optimization*, 32(5):369–382, 2006.

[13] Bruce Ankenman, Barry L Nelson, and Jeremy Staum. Stochastic kriging for simulation metamodeling. *Operations research*, 58(2):371–382, 2010.

[14] Xi Chen, Bruce E Ankenman, and Barry L Nelson. Enhancing stochastic kriging metamodels with gradient estimators. *Operations Research*, 61(2):512–528, 2013.

[15] Erik Vanmarcke. *Random fields: analysis and synthesis.* World Scientific, 2010.

[16] Noel Cressie. *Statistics for spatial data.* John Wiley & Sons, 2015.

[17] Michael L Stein. *Interpolation of spatial data: some theory for kriging.* Springer Science & Business Media, 2012.

[18] Wim Van Beers and Jack PC Kleijnen. Kriging interpolation in simulation: a survey. In *Simulation Conference, 2004. Proceedings of the 2004 Winter*, volume 1. IEEE, 2004.

[19] Xi Chen, Kai Wang, and Feng Yang. Stochastic kriging with qualitative factors. In *Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World*, pages 790–801. IEEE Press, 2013.

[20] Michael C Fu. What you should know about simulation and derivatives. *Naval Research Logistics (NRL)*, 55(8):723–736, 2008.

[21] Michael C Fu. Gradient estimation. *Handbooks in operations research and management science*, 13:575–616, 2006.

[22] Paul Glasserman. *Gradient estimation via perturbation analysis*. Springer Science & Business Media, 1991.

[23] Pierre L'Ecuyer. A unified view of the ipa, sf, and lr gradient estimation techniques. *Management Science*, 36(11):1364–1383, 1990.

[24] James C Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

[25] Kai-Tai Fang, Runze Li, and Agus Sudjianto. *Design and modeling for computer experiments*. CRC Press, 2005.

[26] Jeremy Staum. Better simulation metamodeling: The why, what, and how of stochastic kriging. In *Simulation Conference (WSC), Proceedings of the 2009 Winter*, pages 119–133. IEEE, 2009.

[27] Jun Yin, Szu Hui Ng, and Kien Ming Ng. A study on the effects of parameter estimation on kriging model's prediction error in stochastic simulations. In *Simulation Conference (WSC), Proceedings of the 2009 Winter*, pages 674–685. IEEE, 2009.

[28] Margaret A Oliver and Richard Webster. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System*, 4(3):313–332, 1990.

[29] Xi Chen, Barry L Nelson, and Kyoung-Kuk Kim. Stochastic kriging for conditional value-at-risk and its sensitivities. In *Simulation Conference (WSC), Proceedings of the 2012 Winter*, pages 1–12. IEEE, 2012.

[30] Ming Liu and Jeremy Staum. Stochastic kriging for efficient nested simulation of expected shortfall. *Journal of Risk*, 12(3):3, 2010.

[31] R Evren Baysal, Barry L Nelson, and Jeremy Staum. Response surface methodology for simulating hedging and trading strategies. In *Simulation Conference, 2008. WSC 2008. Winter*, pages 629–637. IEEE, 2008.

[32] Ming Liu, Barry L Nelson, and Jeremy Staum. Simulation on demand for pricing many securities. In *Simulation Conference (WSC), Proceedings of the 2010 Winter*, pages 2782–2789. IEEE, 2010.

[33] Russell R Barton and Martin Meckesheimer. Metamodel-based simulation optimization. *Handbooks in operations research and management science*, 13:535–574, 2006.

[34] Gary G Wang and S. Shan. Review of metamodeling techniques in support of engineering design optimization. *Journal of mechanical design*, 129:370–380, 2006.

[35] Deng Huang, Theodore T Allen, William I Notz, and Ning Zeng. Global optimization of stochastic black-box systems via sequential kriging metamodels. *Journal of global optimization*, 34(3):441–466, 2006.

[36] Ruichen Jin, Wei Chen, and Agus Sudjianto. On sequential sampling for global metamodeling in engineering design. In *ASME 2002 International Design Engineering Technical Conferences and Computers and*

Information in Engineering Conference, pages 539–548. American Society of Mechanical Engineers, 2002.

[37] Liang Zhao, KK Choi, Ikjin Lee, and D Gorsich. A metamodeling method using dynamic kriging and sequential sampling. In *The 13th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Fort Worth, TX, Sept*, pages 13–15, 2010.

[38] Mustafa H Tongarlak, Bruce E Ankenman, and Barry L Nelson. Relative error stochastic kriging. In *Proceedings of the Winter Simulation Conference*, pages 3628–3640. Winter Simulation Conference, 2011.

[39] Ali Ajdari and Hashem Mahlooji. An adaptive exploration-exploitation algorithm for constructing metamodels in random simulation using a novel sequential experimental design. *Communications in Statistics-Simulation and Computation*, 43(5):947–968, 2014.

[40] Averill M Law. *Simulation Modelling and Analysis (4th ed.).* McGraw-Hill, NY, 2007.

[41] Jack P C Kleijnen. Simulation-optimization via kriging and bootstrapping: a survey. *Journal of Simulation*, 8(4):241–250, 2014.

[42] Huashuai Qu and Michael C Fu. Gradient extrapolated stochastic kriging. *ACM Transactions on Modeling and Computer Simulation*, 24(4):Article No. 23, 2014.

[43] Momin Jamil and Xin-She Yang. A literature survey of benchmark functions for global optimisation problems. *International Journal of Mathematical Modelling and Numerical Optimisation*, 4(2):150–194, 2013.

[44] Jiaqiao Hu and Ping Hu. Annealing adaptive search, cross-entropy, and stochastic approximation in global optimization. *Naval Research Logistics (NRL)*, 58(5):457–477, 2011.