Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

ON THE MONOTONIC PERFORMANCE OF STOCHASTIC KRIGING PREDICTORS

Bing Wang
Jiaqiao Hu

Department of Applied Mathematics and Statistics
State University of New York at Stony Brook
Stony Brook, NY 11794, USA

ABSTRACT

Stochastic kriging (SK) has been recognized as a useful and effective technique for approximating the response surface of a simulation model. In this paper, we analyze the performance of SK metamodels in a fully sequential setting when design points are selected one at a time. We consider both cases when the trend term in the model is either known or estimated and show that the prediction performance of the corresponding optimal SK predictor is monotonically improving as the number of design points increases. Numerical examples are also provided to illustrate our findings.

1 INTRODUCTION

Complex systems arising in engineering design and manufacturing have necessitated the need to use simulation models to evaluate their performance. When the cost of simulation is high and the number of design alternatives to be simulated is large, it is often desirable to use metamodels (surrogate models) to represent simulation input-output relations. A variety of metamodeling techniques have been developed, ranging from simple polynomial regressions to sophisticated models based on neural networks and radial-basis function approximations; see, e.g., Barton (2009), Barton and Meckesheimer (2006), Jin et al. (2002), Wang and Shan (2006) and references therein for a review. Well-built metamodels are easier to handle and computationally less demanding than simulation models and can be used to provide good approximations to the underlying simulation response surfaces.

Kriging, originally proposed in geostatistics, is an interpolation-based metamodeling technique that aims to provide a global approximation to the response surface of an unknown (deterministic) function. It models the response surface as a Gaussian process governed by a predefined spatial covariance model and predicts the response values at new locations from observations collected at sampled design points. Ever since its introduction, kriging has been the subject of extensive research and widely used in the design and analysis of deterministic computer experiments (Kleijnen 2008, Kleijnen 2009, Sacks et al. 1989, Santner et al. 2003, Wang and Shan 2006). Recently, an extension of kriging to stochastic simulations has been proposed in Ankenman, Nelson, and Staum (2010). The idea is to augment the original kriging model to include an additional noise term to represent the uncertainty inherent in evaluating the output of a stochastic simulation model. This leads to a general approach called stochastic kriging (SK) that is capable of capturing both the extrinsic and intrinsic uncertainty in approximating the response surfaces of stochastic simulation models.

Obtaining an accurate kriging or SK predictor requires careful selection of design points. A simple but widely accepted method in practice is to generate design points all at once according to a predefined space-filling criterion. When computational resource is limited, it has been argued in e.g., Jin, Chen, and Sudjianto (2002) and Sacks et al. (1989), that selecting design points one at a time in a sequential manner has the benefit of allowing the metamodel to be updated sequentially as data accumulate, so that new design

points can be adaptively determined based on the locations of existing design points and the available information contained in the updated metamodel. A number of sequential strategies have been tested and compared with space-filling designs and showed great promise in deterministic simulations (Sacks et al. 1989, Jin et al. 2002). However, it is not clear whether these strategies can be modified and carried over to the SK framework because response values in stochastic simulations are corrupted by sampling noise. Although sequential experimental design strategies have been proposed for SK models (e.g., Chen and Zhou 2014), a key question from a theoretical standpoint is whether the added information implied by a new design point, which might be misleading due to statistical variance, will actually result in an increase in the prediction performance of an SK predictor.

Motivated by the above question, in this paper we examine the performance of SK metamodels in a fully sequential setting when design points are selected one at a time. Our main result is to show that when all model parameters are fixed (either known or estimated), the prediction performance of SK predictors, as measured by MSE (or equivalently IMSE), is monotonically non-increasing as the number of design points increases. This leads to the interesting finding that the prediction performance of SK models can always be improved by including additional design points, regardless of how these points and simulation replications are allocated. We consider both cases when the trend term in the model is known or estimated and provide explicit formulas showing the reduction in MSE when a new design point is added to the model. The results indicate that the degree of reduction in MSE depends on a number of factors, including the assumed correlation model, the MSE of the old predictor (based on old design points) and the intrinsic variance at the new design point. This suggests that proper selection of design points in SK is also important and may have a significant impact on prediction performance. Note that since SK reduces to kriging when the extra intrinsic noise term vanishes, the same monotonicity property should hold for deterministic kriging models as well and may have direct applications in theoretically justifying the validity of existing sequential strategies based on optimizing MSE or IMSE.

The rest of this paper is structured as follows. Section 2 introduces the notations used in this paper and outlines the basic mathematical framework of SK. In Section 3, we establish the monotonic performance of SK predictors. We illustrate our finding through several numerical examples in Section 4 and conclude the paper in Section 5.

2 STOCHASTIC KRIGING

Consider the problem of describing the response surface of an unknown function $f(\mathbf{x})$, $\mathbf{x} \in \mathscr{X}$, where $\mathbf{x}$ is a vector of design variables and $\mathscr{X}$ is a compact full-dimensional subset of $\mathbb{R}^d$. At each point $\mathbf{x}$, we assume that the true function value $f(\mathbf{x})$ cannot be evaluated exactly but can be estimated in a path-wise manner through stochastic simulation.

Given a set of design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, after we replicate n_i simulations at each point $\mathbf{x}_i$, $i = 1, \ldots, k$, the performance measures at these k design points can be estimated by the vector $\bar{\mathbf{y}} = (\bar{y}(\mathbf{x}_1), \bar{y}(\mathbf{x}_2), \ldots, \bar{y}(\mathbf{x}_k))^\intercal$, where $\bar{y}(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} y_j(\mathbf{x}_i)$ and $y_j(\mathbf{x}_i)$ is the simulation output at $\mathbf{x}_i$ obtained on the jth replication run. In stochastic kriging, $y_j(\mathbf{x}_i)$ is assumed to take the following form:

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

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

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

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

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

and the corresponding optimal MSE is given by

$$MSE(\hat{y}(\mathbf{x}_0)) = \Sigma_M(\mathbf{x}_0, \mathbf{x}_0) - \Sigma_M(\mathbf{x}_0, \cdot)^\intercal (\Sigma_M + \Sigma_\varepsilon)^{-1} \Sigma_M(\mathbf{x}_0, \cdot). \tag{3}$$

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

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

where

$$\hat{\beta} = (\mathbf{F}^\intercal (\Sigma_M + \Sigma_\varepsilon)^{-1} \mathbf{F})^{-1} \mathbf{F}^\intercal (\Sigma_M + \Sigma_\varepsilon)^{-1} \bar{\mathbf{y}}, \tag{5}$$

and the optimal MSE is

$$MSE(\hat{y}(\mathbf{x}_0)) = \Sigma_M(\mathbf{x}_0, \mathbf{x}_0) - \Sigma_M(\mathbf{x}_0, \cdot)^\intercal (\Sigma_M + \Sigma_\varepsilon)^{-1} \Sigma_M(\mathbf{x}_0, \cdot) + \eta^\intercal \left(\mathbf{F}^\intercal (\Sigma_M + \Sigma_\varepsilon)^{-1} \mathbf{F}\right)^{-1} \eta, \tag{6}$$

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

3 MONOTONIC PERFORMANCE OF STOCHASTIC KRIGING PREDICTORS

In this section, we analyze the performance of SK metamodels under the setting where design points are selected one at a time, e.g., via a sequential sampling strategy. We show that under the SK framework, and when either 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. The following condition, due to (Ankenman, Nelson, and Staum 2010), is assumed throughout our analysis:

Assumption 1: *The random field M is a zero mean second-order stationary Gaussian random field, and the intrinsic simulation noises $\varepsilon_1(\mathbf{x}_i), \varepsilon_2(\mathbf{x}_i), \ldots$ are i.i.d. $N(0, V(\mathbf{x}_i))$, independent of $\varepsilon_j(\mathbf{x}_h)$ for all j and $h \neq i$, and independent of M.*

The condition on the random field M implies that the covariance between $M(\mathbf{x}_i)$ and $M(\mathbf{x}_j)$ can be expressed in the form $\mathrm{Cov}[M(\mathbf{x}_i), M(\mathbf{x}_j)] = \tau^2 R_M(d(\mathbf{x}_i, \mathbf{x}_j); \theta)$, 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 $R_M(d, \theta)$ is continuous in its first argument d and satisfies $R_M(0, \theta) = 1$ and $\lim_{d \to \infty} R_M(d, \theta) = 0$. In addition, we also assume that the variance function $V(\mathbf{x})$ is uniformly bounded for all $\mathbf{x} \in \mathscr{X}$.

Unless otherwise specified, we use the subscript k to signify the quantities obtained based on a given set of k design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. Similarly, if a new design point $\mathbf{x}_{k+1}$ is added to an SK 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 1 If Assumption 1 holds, then the matrix $(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}$ is positive definite for all k.

Proof. 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, $\Sigma_{M_k}+\Sigma_{\varepsilon_k}$ is positive definite. This shows that $(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}$ 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 (2) 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 1 Suppose that $\mathbf{x}_{k+1}\notin\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}$. For any prediction point $\mathbf{x}_0\in\mathscr{X}$, let $MSE(\hat{y}_k(\mathbf{x}_0))$ 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 (2). If Assumption 1 holds, then $MSE(\hat{y}_k(\mathbf{x}_0))\geq MSE(\hat{y}_{k+1}(\mathbf{x}_0))$.

Proof. We shall prove the result when the optimal MSE is given by (3). In particular, the MSE of $\hat{y}_k$ can be written as $MSE(\hat{y}_k(\mathbf{x}_0))=\tau^2-\Sigma_{M_k}(\mathbf{x}_0,\cdot)^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0,\cdot)$. 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 $\Sigma_{M_{k+1}}+\Sigma_{\varepsilon_{k+1}}$ can be written in the form $\Sigma_{M_{k+1}}+\Sigma_{\varepsilon_{k+1}}=\begin{pmatrix}\Sigma_{M_k}+\Sigma_{\varepsilon_k} & \Sigma_{k\times 1}\\ \Sigma_{k\times 1}^\intercal & \tau^2+\Sigma_\varepsilon(\mathbf{x}_{k+1},\mathbf{x}_{k+1})\end{pmatrix}$, 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 $\Sigma_\varepsilon(\mathbf{x}_{k+1},\mathbf{x}_{k+1})=\mathrm{Cov}[\bar{\varepsilon}(\mathbf{x}_{k+1}),\bar{\varepsilon}(\mathbf{x}_{k+1})]$. By Lemma 1, $\Sigma_{M_{k+1}}+\Sigma_{\varepsilon_{k+1}}$ is positive definite and thus invertible. Its inverse, denoted by A, can be calculated using the block matrix inversion formula as follows:

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

where

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

Thus, it follows that

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

$$
\begin{aligned}
&= \tau^2 - \Sigma_{M_k}(\mathbf{x}_0,\cdot)^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0,\cdot) - \Big(\Sigma_{M_k}(\mathbf{x}_0,\cdot)^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\Sigma_{k\times 1} \\
&\qquad\qquad - \Sigma_M(\mathbf{x}_0,\mathbf{x}_{k+1})^\intercal\Big)\Phi\Big(\Sigma_{k\times 1}^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0,\cdot) - \Sigma_M(\mathbf{x}_0,\mathbf{x}_{k+1})\Big) \\
&= MSE(\hat{y}_k(\mathbf{x}_0)) - \phi(\mathbf{x}_0)^2\Phi,
\end{aligned}
$$

where we have defined $\phi(\mathbf{x}_0) = \Sigma_{k\times 1}^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0,\cdot) - \Sigma_M(\mathbf{x}_0,\mathbf{x}_{k+1})$. 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 1 still holds true when the predictors are constructed using Equation (4).

Theorem 2 Suppose that $\mathbf{x}_{k+1} \notin \{\mathbf{x}_1,\ldots,\mathbf{x}_k\}$. For any prediction point $\mathbf{x}_0 \in \mathscr{X}$, let $MSE(\hat{y}_k(\mathbf{x}_0))$ 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 (4). 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. We use the same shorthand notation A, Φ, and $\phi(\mathbf{x}_0)$ as in the proof of Theorem 1. When $\hat{y}_{k+1}(\mathbf{x}_0)$ is constructed using (4), its associated MSE becomes

$$
\begin{aligned}
MSE(\hat{y}_{k+1}(\mathbf{x}_0)) = \tau^2 &- \Sigma_{M_{k+1}}(\mathbf{x}_0,\cdot)^\intercal(\Sigma_{M_{k+1}}+\Sigma_{\varepsilon_{k+1}})^{-1}\Sigma_{M_{k+1}}(\mathbf{x}_0,\cdot) \\
&+ \eta_{k+1}(\mathbf{x}_0)^\intercal(\mathbf{F}_{k+1}^\intercal(\Sigma_{M_{k+1}}+\Sigma_{\varepsilon_{k+1}})^{-1}\mathbf{F}_{k+1})^{-1}\eta_{k+1}(\mathbf{x}_0), \qquad (7)
\end{aligned}
$$

where $\eta_{k+1}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0) - \mathbf{F}_{k+1}^\intercal(\Sigma_{M_{k+1}}+\Sigma_{\varepsilon_{k+1}})^{-1}\Sigma_{M_{k+1}}(\mathbf{x}_0,\cdot)$. From the proof of Theorem 1, it is easy to see that $\Sigma_{M_{k+1}}(\mathbf{x}_0,\cdot)^\intercal(\Sigma_{M_{k+1}}+\Sigma_{\varepsilon_{k+1}})^{-1}\Sigma_{M_{k+1}}(\mathbf{x}_0,\cdot) = \Sigma_{M_k}(\mathbf{x}_0,\cdot)^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0,\cdot) + \phi(\mathbf{x}_0)^2\Phi$. Thus, the right-hand-side of (7) can be written in terms of $MSE(\hat{y}_k(\mathbf{x}_0))$ as

$$
\begin{aligned}
MSE(\hat{y}_{k+1}(\mathbf{x}_0)) = MSE(\hat{y}_k(\mathbf{x}_0)) &- \phi(\mathbf{x}_0)^2\Phi - \eta_k(\mathbf{x}_0)^\intercal(\mathbf{F}_k^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\mathbf{F}_k)^{-1}\eta_k(\mathbf{x}_0) \\
&+ \eta_{k+1}(\mathbf{x}_0)^\intercal(\mathbf{F}_{k+1}^\intercal(\Sigma_{M_{k+1}}+\Sigma_{\varepsilon_{k+1}})^{-1}\mathbf{F}_{k+1})^{-1}\eta_{k+1}(\mathbf{x}_0). \qquad (8)
\end{aligned}
$$

Regarding the last term in (8), we have

$$
\begin{aligned}
&\mathbf{F}_{k+1}^\intercal(\Sigma_{M_{k+1}}+\Sigma_{\varepsilon_{k+1}})^{-1}\mathbf{F}_{k+1} \\
&= (\mathbf{F}_k^\intercal, \mathbf{f}(\mathbf{x}_{k+1}))A\begin{pmatrix}\mathbf{F}_k \\ \mathbf{f}(\mathbf{x}_{k+1})^\intercal\end{pmatrix} \\
&= \mathbf{F}_k^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\mathbf{F}_k + \big(\mathbf{F}_k^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\Sigma_{k\times 1} - \mathbf{f}(\mathbf{x}_{k+1})\big)\Phi\big(\Sigma_{k\times 1}^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\mathbf{F}_k - \mathbf{f}(\mathbf{x}_{k+1})^\intercal\big) \\
&= \mathbf{F}_k^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\mathbf{F}_k + \psi^\intercal\Phi\psi, \qquad (9)
\end{aligned}
$$

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

$$
\begin{aligned}
\eta_{k+1}(\mathbf{x}_0) &= \mathbf{f}(\mathbf{x}_0) - \mathbf{F}_{k+1}^\intercal(\Sigma_{M_{k+1}}+\Sigma_{\varepsilon_{k+1}})^{-1}\Sigma_{M_{k+1}}(\mathbf{x}_0,\cdot) \\
&= \mathbf{f}(\mathbf{x}_0) - (\mathbf{F}_k^\intercal, \mathbf{f}(\mathbf{x}_{k+1}))A\begin{pmatrix}\Sigma_{M_k}(\mathbf{x}_0,\cdot) \\ \Sigma_M(\mathbf{x}_0,\mathbf{x}_{k+1})\end{pmatrix} \\
&= \mathbf{f}(\mathbf{x}_0) - \mathbf{F}_k^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0,\cdot) \\
&\quad - \big(\mathbf{F}_k^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\Sigma_{k\times 1} - \mathbf{f}(\mathbf{x}_{k+1})\big)\Phi\big(\Sigma_{k\times 1}^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\Sigma_{M_k}(\mathbf{x}_0,\cdot) - \Sigma_M(\mathbf{x}_0,\mathbf{x}_{k+1})\big) \\
&= \eta_k(\mathbf{x}_0) - \psi^\intercal\Phi\phi(\mathbf{x}_0) \qquad (10)
\end{aligned}
$$

Substituting (9) and (10) into (8), we obtain

$$
\begin{aligned}
MSE(\hat{y}_{k+1}(\mathbf{x}_0)) = MSE(\hat{y}_k(\mathbf{x}_0)) &- \phi(\mathbf{x}_0)^2\Phi - \eta_k(\mathbf{x}_0)^\intercal(\mathbf{F}_k^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\mathbf{F}_k)^{-1}\eta_k(\mathbf{x}_0) \\
&+ (\eta_k(\mathbf{x}_0) - \psi^\intercal\Phi\phi(\mathbf{x}_0))^\intercal(\mathbf{F}_k^\intercal(\Sigma_{M_k}+\Sigma_{\varepsilon_k})^{-1}\mathbf{F}_k + \psi^\intercal\Phi\psi)^{-1}(\eta_k(\mathbf{x}_0) - \psi^\intercal\Phi\phi(\mathbf{x}_0)). \qquad (11)
\end{aligned}
$$

Next, define $W = \mathbf{F}_k^\intercal(\Sigma_{M_k} + \Sigma_{\varepsilon_k})^{-1}\mathbf{F}_k$. Since $\mathbf{F}_k$ is assumed to have full column rank and $(\Sigma_{M_k} + \Sigma_{\varepsilon_k})^{-1}$ 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

$$(\mathbf{F}_k^\intercal(\Sigma_{M_k} + \Sigma_{\varepsilon_k})^{-1}\mathbf{F}_k + \psi^\intercal\Phi\psi)^{-1} = (W + \psi^\intercal\Phi\psi)^{-1} = W^{-1} - \frac{W^{-1}\psi^\intercal\psi W^{-1}}{\Phi^{-1} + \psi W\psi^\intercal}. \tag{12}$$

Finally, substituting (12) into (11), we get

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

This completes the proof of the Theorem 2. □

4 NUMERICAL EXAMPLES

In this section, we perform some computational experiments to illustrate the monotonic performance of SK predictors. We consider two sets of examples: an M/M/1 queue and two deterministic functions with added noises. In all experiments, we construct SK models using a constant trend term (estimated via (5)) and the Gaussian correlation function $R_M(d(\mathbf{x}_i, \mathbf{x}_j), \theta) = \exp(-\theta(\mathbf{x}_i - \mathbf{x}_j)^2)$. To quantify the overall quality of an SK predictor over the entire decision domain, we use the IMSE as a measure of performance:

$$IMSE(k) \triangleq \int_{\mathbf{x}\in\mathscr{X}} MSE(\hat{y}_k(\mathbf{x}))d\mathbf{x}. \tag{13}$$

To test our theoretical results, we consider the following simple sequential version of a space-filling scheme based on quasi-Monte Carlo sampling:

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 $\mathscr{X}$ $(k < N)$, and the number of simulation replications n_0 at each design point.
Step 1: Collect output performance measures at each $\mathbf{x}_i$. Fit an initial SK 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. Compute the IMSE of the SK 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**.

In all testing examples, we set $N = 15$, the number of initial points to 5, and the number of simulation replications $n_0 = 30$. For each example, we repeat the above procedure 30 times and then plot the mean IMSE versus the number of added design points.

4.1 An M/M/1 queue

This example is taken from Ankenman, Nelson, and Staum (2010). Consider an M/M/1 queue with service rate 1 and arrival rate $x \in (0,1)$. Let $f(x)$ be the long run expected number of customers in system. The goal is to approximate the response surface $f(x)$ over the domain $[0.05, 0.95]$ using an SK model with an increasing number of design points and observe how its IMSE changes. 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 (t = 1000 time units), 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 (e.g., Ankenman, Nelson, and Staum (2010)).

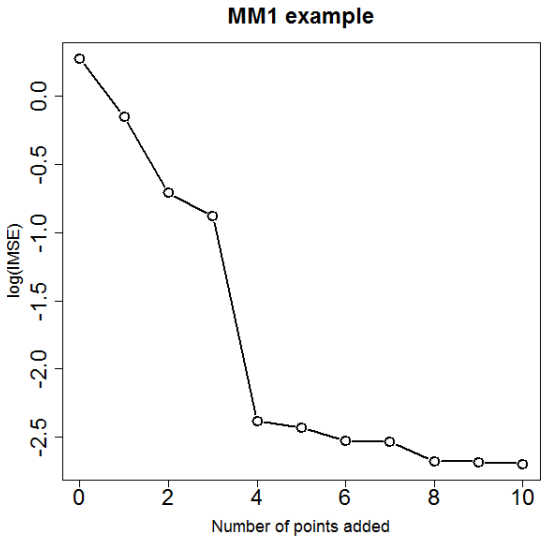


Figure 1: M/M/1 queue example

Figure 1 shows the mean IMSE (averaged over 30 independent runs) in log scale as a function of the number of added design points. The figure clearly indicates the monotonicity of the IMSE of the SK model, which conforms well with our theoretical result. Notice that these results are based on space-filling designs, so the prediction performance of the model can still be further improved via a refined selection of design points, e.g., via optimizing MSE or IMSE.

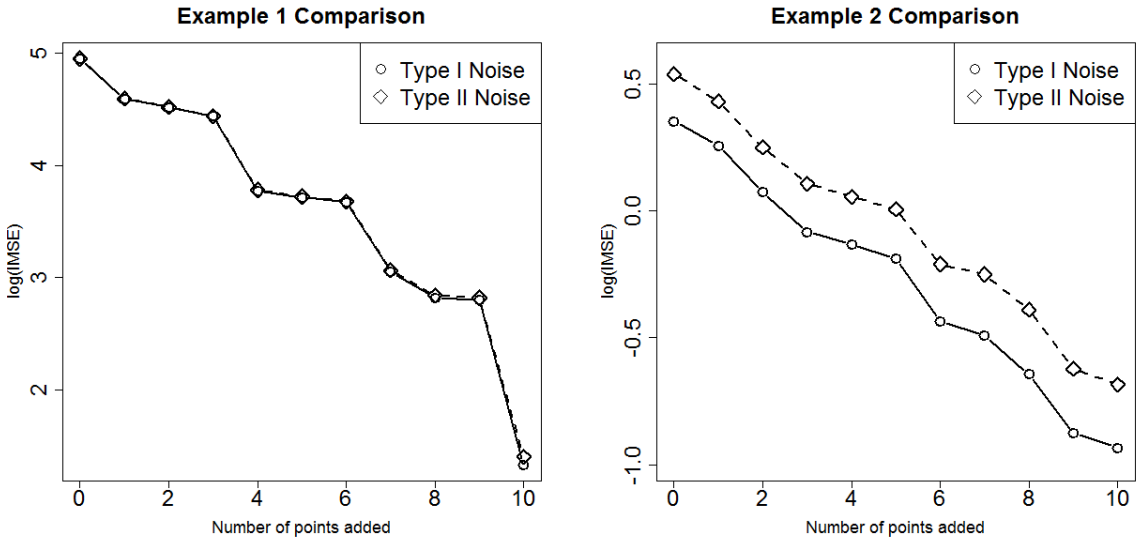
4.2 Deterministic examples with added noise

We consider the following benchmark functions (Qu and Fu 2014, Santner et al. 2003):

(1) $y(x) = Y(x) + \varepsilon(x)$, $x \in [-3,0]$, where $Y(x) = \exp(-1.4x)\cos(\frac{7\pi x}{2}) + 50$ and $\varepsilon(x) \sim \mathcal{N}(0, V(x))$.

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

In both examples, we consider two types of variance functions: (i) $V(\mathbf{x}) = 1$, and (ii) $V(\mathbf{x}) = \mathbf{x}^\intercal\mathbf{x} + 1$.



(a) (b)

Figure 2: Deterministic examples with added noise

In each of the respective test cases, Figure 2 shows that the mean IMSE curve is monotonically decreasing as the number of new design points increases. It is interesting to see from Figure 2 (a) that on test function (1), the two IMSE curves, which are calculated based on different types of noises, are almost identical, whereas on test function (2) (Figure 2(b)) the two mean IMSE curves have roughly the same shape but the curve obtained from type II noise consistently dominates that obtained from type I noise. This suggests that when design points are fixed, the intrinsic variance may have a significant impact on the prediction performance of SK models.

5 CONCLUSION

In this paper, we have investigated the performance of SK metamodels in a fully sequential setting. Our main result is the theoretical finding that the MSE of an SK predictor is monotonically non-increasing in the number of design points. We have illustrated the monotonic performance of SK models through several examples. This result not only has utility in analyzing the performance of existing sequential strategies in constructing deterministic kriging models, but also has potential in developing new sequential sampling procedures under the SK framework.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grant CMMI-1130761.

REFERENCES

Ankenman, B., B. L. Nelson, and J. Staum. 2010. "Stochastic kriging for simulation metamodeling". *Operations research* 58 (2): 371–382.

Barton, R. R. 2009. "Simulation optimization using metamodels". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 230–238: Piscataway, NJ: IEEE.

Barton, R. R., and M. Meckesheimer. 2006. "Metamodel-based simulation optimization". *Handbooks in operations research and management science* 13:535–574.

Chen, X., K. Wang, and F. Yang. 2013. "Stochastic kriging with qualitative factors". In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 790–801: Piscataway, NJ: IEEE.

Chen, X., and Q. Zhou. 2014. "Sequential Experimental Designs for Stochastic Kriging". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 3821–3832: Piscataway, NJ: IEEE.

Jin, R., W. Chen, and A. Sudjianto. 2002. "On sequential sampling for global metamodeling in engineering design". In *ASME 2002 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 539–548. American Society of Mechanical Engineers.

Kleijnen, J. P. 2008. *Design and analysis of simulation experiments*, Volume 111. Springer US.

Kleijnen, J. P. 2009. "Kriging metamodeling in simulation: A review". *European Journal of Operational Research* 192 (3): 707–716.

Qu, H., and M. C. Fu. 2014. "Gradient Extrapolated Stochastic Kriging". *ACM Transactions on Modeling and Computer Simulation* 24 (4): Article No. 23.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn. 1989. "Design and analysis of computer experiments". *Statistical science*:409–423.

Santner, T. J., B. J. Williams, and W. Notz. 2003. *The design and analysis of computer experiments*. Springer.

Wang, G. G., and S. Shan. 2006. "Review of Metamodeling Techniques in Support of Engineering Design Optimization". *Journal of mechanical design* 129:370–380.

AUTHOR BIOGRAPHIES

BING WANG is a PhD candidate in the Department of Applied Mathematics and Statistics at Stony Brook University. His research interests include simulation optimization, design and analysis of simulation experiments and applied predictive modeling. His email address is bing.wang.2@stonybrook.edu.

JIAQIAO HU is an Associate Professor in the Department of Applied Mathematics and Statistics at the State University of New York, Stony Brook. He received the B.S. degree in automation from Shanghai Jiao Tong University, the M.S. degree in applied mathematics from the University of Maryland, Baltimore County, and the Ph.D. degree in electrical engineering from the University of Maryland, College Park. His research interests include Markov decision processes, applied probability, and simulation optimization. His email address is jqhu@ams.sunysb.edu.