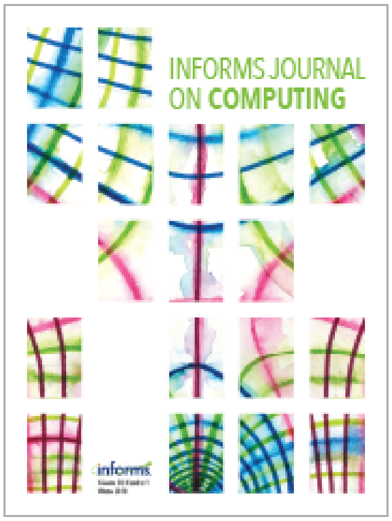


INFORMS Journal on Computing

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Convergence Analysis of Stochastic Kriging-Assisted Simulation with Random Covariates

Cheng Li, Siyang Gao, Jianzhong Du

To cite this article:
Cheng Li, Siyang Gao, Jianzhong Du (2023) Convergence Analysis of Stochastic Kriging-Assisted Simulation with Random Covariates. INFORMS Journal on Computing 35(2):386-402. https://doi.org/10.1287/ijoc.2022.1263

INFORMS JOURNAL ON COMPUTING
Vol. 35, No. 2, March–April 2023, pp. 386–402
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

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

Convergence Analysis of Stochastic Kriging-Assisted Simulation with Random Covariates

Cheng Li,[a] Siyang Gao,[b,*] Jianzhong Du[c]

[a] Department of Statistics and Data Science, National University of Singapore, Singapore 117546, Singapore; [b] Department of Advanced Design and Systems Engineering and School of Data Science, City University of Hong Kong, Hong Kong; [c] School of Management, Fudan University, Shanghai 200433, China

*Corresponding author

Contact: stalic@nus.edu.sg, https://orcid.org/0000-0001-7522-7072 (CL); siyangao@cityu.edu.hk, https://orcid.org/0000-0002-3574-6393 (SG); jianzhodu2-c@my.cityu.edu.hk, https://orcid.org/0000-0002-5355-5902 (JD)

Received: November 23, 2021
Revised: July 20, 2022; November 17, 2022
Accepted: November 29, 2022
Published Online in Articles in Advance: January 10, 2023

https://doi.org/10.1287/ijoc.2022.1263

Abstract. We consider performing simulation experiments in the presence of covariates. Here, covariates refer to some input information other than system designs to the simulation model that can also affect the system performance. To make decisions, decision makers need to know the covariate values of the problem. Traditionally in simulation-based decision making, simulation samples are collected after the covariate values are known; in contrast, as a new framework, simulation with covariates starts the simulation before the covariate values are revealed and collects samples on covariate values that might appear later. Then, when the covariate values are revealed, the collected simulation samples are directly used to predict the desired results. This framework significantly reduces the decision time compared with the traditional way of simulation. In this paper, we follow this framework and suppose there are a finite number of system designs. We adopt the metamodel of stochastic kriging (SK) and use it to predict the system performance of each design and the best design. The goal is to study how fast the prediction errors diminish with the number of covariate points sampled. This is a fundamental problem in simulation with covariates and helps quantify the relationship between the offline simulation efforts and the online prediction accuracy. Particularly, we adopt measures of the maximal integrated mean squared error (IMSE) and integrated probability of false selection (IPFS) for assessing errors of the system performance and the best design predictions. Then, we establish convergence rates for the two measures under mild conditions. Last, these convergence behaviors are illustrated numerically using test examples.

History: Accepted by Bruno Tuffin, area editor for simulation.

Funding: This work was supported in part by Singapore Ministry of Education Academic Research Funds [Tier 1 Grants R-155-000-201-114 and A-0004822-00-00], the City University of Hong Kong [Grants 7005269 and 7005568], and the National Natural Science Foundation of China [Grant 72091211].

Supplemental Material: The software that supports the findings of this study is available within the paper and its Supplemental Information (https://pubsonline.informs.org/doi/suppl/10.1287/ijoc.2022.1263) as well as from the IJOC GitHub software repository (https://github.com/INFORMSJoC/2021.0329) at (http://dx.doi.org/10.5281/zenodo.7344997).

Keywords: simulation with covariates • convergence rate • stochastic kriging • ranking and selection

1. Introduction

Stochastic simulation is a powerful tool for analyzing large-scale complex systems. In most real situations, systems are highly complex, precluding the possibility of applying analytical solutions; in contrast, simulation makes it possible to accurately describe a system using logically complex and often nonmathematical models. Consequently, detailed dynamics of the system can be faithfully modeled, the system performance can be studied, and the best system design can be selected (Chen and Lee 2011). Simulation has been a widely used operations research and management science technique, for example, in the management of power systems (Benini et al. 1998), production planning (Kleijnen 1993), supply chain network (Ding et al. 2005), emergency department (Ahmed and Alkhamis 2009), and so on.

In these applications, the standard process for analyzing the system is to first establish estimators for measures of interest based on the simulation output and then develop optimization methods to find the best design of the system. This process highlights the two main purposes of a constructed simulation model for estimating the system performance and optimizing it over a set of system designs. Throughout the paper, we will refer to these two purposes of simulation as the *estimation problem* and the *optimization problem*.

When conducting simulation experiments, a common practice is to first reveal and fix the covariate values for the problem under consideration and then repeat experiments on the simulation model with various system designs. Here, covariates refer to some input information other than system designs to the simulation model that will also affect the system performance. In the literature, covariates are also known as the side information or context. For example, in queueing network design, covariates can be the arrival rate of the customers, which influences the queue length and the mean waiting time of the network. In disease treatment, covariates can be the biometric characteristics of the patients, which influence the efficacy of the treatment methods.

However, given the computational expense of simulation experiments, a notable issue with this practice, for both the purposes of estimation and optimization, is that the time for obtaining the desired simulation results can be very long for some real systems. In addition to the huge monetary cost it incurs, it significantly limits the use of simulation for online problems in which system performance and the best system design are expected soon after the covariate values are revealed. This is also one of the key concerns for simulation-related research (Law 2015).

To address this issue, Hong and Jiang (2019) and Shen et al. (2021) recently proposed a new framework of using simulation. Instead of running simulation after the covariate values are revealed, the new framework does it before that with randomly sampled covariate values that might possibly appear in future problem instances. It establishes an offline simulation data set that is useful in describing the system. More importantly, this data set serves for the purpose of prediction. When the covariate values of a certain problem are known, machine learning and data mining tools can be adopted to build predictive models and predict the performance of each design (the estimation problem) and the best design (the optimization problem) in real time.[1] For example, a doctor can learn the efficacy of the potential treatment methods and recommend a personalized treatment for a diabetic patient immediately on his/her arrival by checking the simulation results under the same biometric characteristics (covariate values) of this patient (Bertsimas et al. 2017). By doing so, the time for obtaining performance estimation and the best decision can be substantially reduced. It enables simulation to be used in a much broader range of applications for which simulation was hardly a feasible technique before. We call this framework *simulation with covariates*.

The framework of simulation with covariates is quite general and new. A lot of key questions remain largely unexplored. In this research, we focus on the use of this framework in prediction and consider a fundamental problem in it, the quantification of the relationship between the offline simulation efforts and the online prediction accuracy. This quantification provides a good assessment on the quality of the estimated system performance and the best design that can be achieved using the offline data set. We consider a continuous covariate space and a finite number of system designs. We sample the covariate space using a fixed distribution, conduct the same number of simulation replications on all the designs and sampled covariate points, and construct a predictive model for each design for predicting its performance and selecting the best design. Our main research question is to study the convergence rates of the prediction errors with the number of covariate points ever collected and to facilitate further decision making.

We use the stochastic kriging (SK) model as the predictive model. SK has is one of the most extensively studied models for simulation output (Ankenman et al. 2010, Chen et al. 2013, Qu and Fu 2014, Wang and Hu 2018). It is a general purpose model with less structural assumptions than linear and some nonlinear models and tends to be more resistant to overfitting than general interpolators (Sabuncuoglu and Touhami 2002).

To evaluate the prediction errors of the estimation and optimization problems, we will use the maximal *integrated mean squared error* (IMSE) and *integrated probability of false selection* (IPFS), respectively. IMSE is the integral of the mean squared error of the SK model over the covariate space. An IMSE is associated to a system design and describes the average MSE of the estimated system performance of this design over all the possible covariate values. The maximal IMSE corresponds to the largest IMSE from the designs. It serves as a measure for the worst-case error of the estimation problem, whose convergence rate governs the prediction errors for the performance of each design under consideration. IPFS is the integral of the probability of false selection, that is, the probability of falsely selecting the best design using the SK predictions. It serves as a measure for the error of the optimization problem.

In this study, we use a fixed distribution to sample the covariate space for three reasons. First, for real systems, covariates usually follow a fixed population distribution that can be estimated from historical data. Therefore, the offline data set generated from this distribution can faithfully describe the distributional characteristics of the system and lead to more accurate estimation over the covariate space. Second, from the experiment design perspective, although more sophisticated sequential designs may have the benefit of using fewer design points in the covariate space, they may not be able to incorporate the distributional information due to the high computational cost in each iteration and may incur higher simulation cost for certain types of response surfaces. In comparison, sampling from a fixed distribution has the advantage of being simple with a fixed prespecified offline simulation cost. The distributional information also helps achieve sufficiently good performance when the number of covariate points sampled is large, and this advantage becomes more obvious when the covariate space has a higher dimension. Third, the setting of fixed-distribution sampling enables us to theoretically derive concrete convergence rates for the two target measures. These convergence rates serve as a good benchmark, against which improvement from future design methods with possibly faster convergence rates might be measured (theoretically or numerically).

1.1. Contributions

Our work makes three main contributions.

First, we establish a formulation for characterizing the performance of simulation with covariates in both the estimation and optimization problems. As one of the first simulation-based real-time decision-making frameworks, simulation with covariates resolves the long-standing issue of efficiency for simulation experiments but has rendered itself unclear about the effectiveness of the decision that is made. Our research builds an SK prediction model for each system design under study and proposes measures for the estimation and optimization problems that evaluate the quality of the prediction over all the possible problem instances that might be encountered. It lays the ground for theoretical analysis of simulation with covariates and other possible simulation frameworks of this kind.

Second, we derive the convergence rates of the two target measures (the maximal IMSE and IPFS) with the number of sampled covariate points m for three common types of SK covariance kernels: finite-rank kernels, exponentially decaying kernels, and polynomially decaying kernels. Derivation for the rates of the two measures is based on the upper bounds of the IMSE of a single SK model and contains additional analysis on the structures of the target measures. Specifically, we show that convergence rates of the two measures are both at the magnitudes of $1/m$, $(\log m)^{\frac{d}{\kappa_*}}/m$ and $m^{-\frac{2\nu_*}{2\nu_*+d}}$ for the three types of kernels, respectively. In these rates, κ_* and ν_* are some kernel parameters, and d is the dimension of covariates. We also show that the convergence rate of IPFS can be improved to exponential with additional mild assumptions on the tail of MSE of each SK model. They provide good insight into the practical performance simulation with covariates can achieve.

Third, based on the polynomial convergence rates of the maximal IMSE, we further propose a simple regression-based procedure to determine the number of distinct covariate points needed to achieve a target precision of the maximal IMSE in Section 5.3 of the online supplement. In addition, we numerically illustrate the convergence behaviors of the maximal IMSE and IPFS via several test examples and show the impact of several factors on their convergence rates, including the problem structure, dimension of the covariate space, number of simulation replications, and sampling distribution.

1.2. Literature Review

There are two streams of literature related to this study.

The first stream is kriging, or Gaussian process regression, which is a popular interpolation method for building metamodels (Stein 1999, Kleijnen 2009). It interpolates the response surface of an unknown function using the realization of a Gaussian random field and has proven to be a highly effective tool for global metamodeling. In Ankenman et al. (2010), kriging was extended to simulation modeling, in which the observations of the unknown function are no longer deterministic but are corrupted by random noises. It is known as the SK. Chen et al. (2013) and Qu and Fu (2014) further enhanced SK by using the gradient information when it is available, called stochastic kriging with gradient estimators (SKG). Wang and Hu (2018) proved the monotonicity of MSE in a sequential setting for both SK and SKG. Theoretical properties of Gaussian process regression and the related kernel ridge regression have been previously studied in van der Vaart and van Zanten (2011), Steinwart et al. (2009), and so on. Instead of a single SK model studied in those papers, in this research, we are interested in measures from multiple SK models that are caused by multiple designs.

The second stream is ranking and selection (R&S), in particular, the fixed-budget R&S. Fixed-budget R&S is a basic problem in simulation-based optimization, seeking to determine the allocation of a fixed simulation budget to correctly select the best simulated system design among a finite set of alternatives. Popular methods in this field include the optimal computing budget allocation (OCBA; Chen et al. 2000, 2008; Gao and Chen 2017; Gao et al. 2017) and value of information procedure (VIP; Frazier et al. 2008, Ryzhov 2016). In particular, Gao et al. (2019a) used the OCBA approach to solve the R&S problem with discrete covariates and derived the asymptotic optimal sampling rule. Similar to fixed-budget R&S, this research is also set up with a finite number of designs and samples them with a fixed simulation budget to make decisions. However, this research is different in objective. It aims to analyze the convergence rates of the target measures based on an existing sampling scheme instead of developing a new sampling scheme as in fixed-budget R&S.

The rest of the paper is organized as follows. Section 2 presents the formulation of the problem. Sections 3 and 4 provide the main convergence rate results on the maximal IMSE and IPFS. Numerical examples are presented in Section 5, followed by conclusions and discussion in Section 6. A preliminary study of this research appeared in Gao et al. (2019b). That paper only focused on the exponentially decaying kernels and presented the convergence rates of the maximal IMSE and IPFS without proof.

2. Problem Formulation

In this section, we provide some preliminaries on the SK model and the definitions of the two target measures. For a summary of the key notation we use, please refer to Table 1 of the online supplement. Throughout the paper, the subscript i is exclusively used to index the system design, and we will fold it for circumstances with no ambiguity.

2.1. SK

We consider a finite number of k system designs. The performance of each design depends on $\mathbf{X} = (X_1,\ldots,X_d)^\top$, a vector of random covariates with support $\mathcal{X} \subseteq \mathbb{R}^d$. For each $i = 1,2,\ldots,k$, let $Y_{il}(\mathbf{X})$ be the lth simulation sample from design i under covariate $\mathbf{X}$, and $y_i(\mathbf{X})$ be the mean of design i, where the mean is taken with respect to the simulation noise. We assume that for any $\mathbf{X} = \mathbf{x}$, $Y_{il}(\mathbf{x}) = y_i(\mathbf{x}) + \epsilon_{il}(\mathbf{x})$, where $\epsilon_{il}(\mathbf{x})$s are mean-zero simulation noises and are independent across different i, l and $\mathbf{x}$.

The relationship between the performance $y_i(\mathbf{x})$ of design i and $\mathbf{x}$ is generally unknown and can only be estimated via stochastic simulations. In this paper, we use the SK model to describe $y_i(\mathbf{x})$:

$$y_i(\mathbf{x}) = \mathbf{f}_i(\mathbf{x})^\top \boldsymbol{\beta}_i + M_i(\mathbf{x}), \quad i = 1,\ldots,k, \tag{1}$$

where $\mathbf{f}_i(\mathbf{x}) = (\mathrm{f}_{i1}(\mathbf{x}),\ldots,\mathrm{f}_{iq}(\mathbf{x}))^\top$ and $\boldsymbol{\beta}_i = (\beta_{i1},\ldots,\beta_{iq})^\top$ are a $q \times 1$ vector of known functions of $\mathbf{x}$ and a $q \times 1$ vector of unknown parameters; $M_i(\mathbf{x})$ is a realization (or sample path) of a mean zero stationary Gaussian process, with the covariance function $\boldsymbol{\Sigma}_{M,i}(\mathbf{x},\mathbf{x}') = \mathrm{Cov}[M_i(\mathbf{x}), M_i(\mathbf{x}')]$ quantifying the covariance between $M_i(\mathbf{x})$ and $M_i(\mathbf{x}')$ for any $\mathbf{x},\mathbf{x}' \in \mathcal{X}$. Model (1) with regressor functions $\mathbf{f}_i(\cdot)$ is sometimes called *universal kriging* (Stein 1999).

In our model setting, we assume that we randomly draw m covariate (design) points $\mathbf{X}^m = \{\mathbf{X}_1,\ldots,\mathbf{X}_m\}$ of $\mathbf{X}$ from a sampling distribution $\mathbb{P}_{\mathbf{X}}$. For a given covariate point sample $\mathbf{x}^m = \{\mathbf{x}_1,\ldots,\mathbf{x}_m\}$, we perform n_j replications at covariate $\mathbf{x}_j$ for each of the k designs. We denote the sample mean for design i and covariate $\mathbf{x}_j$ by $\overline{Y}_i(\mathbf{x}_j) = n_j^{-1}\sum_{l=1}^{n_j} Y_{il}(\mathbf{x}_j)$, and correspondingly the averaged simulation errors by $\overline{\epsilon}_i(\mathbf{x}_j) = n_j^{-1}\sum_{l=1}^{n_j} \epsilon_{il}(\mathbf{x}_j)$. For $i = 1,\ldots,k$ and $j = 1,\ldots,m$, we let $\mathbf{Y}_{ij} = (Y_{i1}(\mathbf{x}_j),\ldots,Y_{in_j}(\mathbf{x}_j))^\top$, and let $\overline{\mathbf{Y}}_i = \left(\overline{Y}_i(\mathbf{x}_1),\ldots,\overline{Y}_i(\mathbf{x}_m)\right)^\top$. For design i, let the $m \times q$ design matrix be $\mathbf{F}_i = (\mathbf{f}_i(\mathbf{x}_1),\ldots,\mathbf{f}_i(\mathbf{x}_m))^\top$. Let $\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^m,\mathbf{x}^m)$ be the $m \times m$ covariance matrix across all covariate points $\mathbf{x}_1,\ldots,\mathbf{x}_m$; that is, for $s,t \in \{1,\ldots,m\}$, the $(s,\ t)$ entry of $\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^m,\mathbf{x}^m)$ is $[\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^m,\mathbf{x}^m)]_{st} = \mathrm{Cov}\left[y_i(\mathbf{x}_s), y_i(\mathbf{x}_t)\right]$. For any $\mathbf{x} \in \mathcal{X}$, let

$$\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^m,\mathbf{x}) = \left(\mathrm{Cov}\left[y_i(\mathbf{x}), y_i(\mathbf{x}_1)\right],\ldots,\mathrm{Cov}\left[y_i(\mathbf{x}), y_i(\mathbf{x}_m)\right]\right)^\top.$$

Let $\boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^m)$ be the $m \times m$ covariance matrix of the averaged simulation errors across m covariate points in the design i; that is, for $s,t \in \{1,\ldots,m\}$, the (s,t) entry of $\boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^m)$ is $\{\boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^m)\}_{st} = \mathrm{Cov}[\overline{\epsilon}_i(\mathbf{x}_s), \overline{\epsilon}_i(\mathbf{x}_t)]$. Let $\boldsymbol{\Sigma}_{y,i} = \boldsymbol{\Sigma}_{M,i}(\mathbf{x}^m,\mathbf{x}^m) + \boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^m)$.

To estimate $y_i(\mathbf{x})$ in (1), we consider linear predictors in the form of $\alpha_{i,0}(\mathbf{x}_0) + \boldsymbol{\alpha}_i(\mathbf{x}_0)\overline{\mathbf{Y}}_i$, where $\alpha_{i,0}(\mathbf{x}_0)$ and $\boldsymbol{\alpha}_i(\mathbf{x}_0)$ are weights that depend on the test covariate point $\mathbf{x}_0 \in \mathcal{X}$. The mean squared error MSE of the predictors at $\mathbf{x}_0$ is given by $\mathrm{MSE}_i(\mathbf{x}_0) = \mathrm{E}[(y_i(\mathbf{x}_0) - \alpha_{i,0}(\mathbf{x}_0) - \boldsymbol{\alpha}_i(\mathbf{x}_0)\overline{\mathbf{Y}}_i)^2]$, where the expectation is with respect to the randomness in $\overline{\mathbf{Y}}_i$, that is, the simulation noise. We call the predictor that minimizes $\mathrm{MSE}_i(\mathbf{x}_0)$ the MSE-optimal linear predictor. Stein (1999) (also Ankenman et al. 2010 and Chen et al. 2013) has shown that the MSE-optimal linear predictor has the form

$$\hat{y}_i(\mathbf{x}_0) = \mathbf{f}_i(\mathbf{x}_0)^\top \widehat{\boldsymbol{\beta}}_i + \boldsymbol{\Sigma}_{M,i}(\mathbf{x}^m,\mathbf{x}_0)^\top \boldsymbol{\Sigma}_{y,i}^{-1}\left(\overline{\mathbf{Y}}_i - \mathbf{F}_i\widehat{\boldsymbol{\beta}}_i\right), \tag{2}$$

where $\widehat{\boldsymbol{\beta}}_i = (\mathbf{F}_i^\top \boldsymbol{\Sigma}_{y,i}^{-1}\mathbf{F}_i)^{-1}\mathbf{F}_i^\top \boldsymbol{\Sigma}_{y,i}^{-1}\overline{\mathbf{Y}}_i$.

In addition, Ankenman et al. (2010) has shown that the optimal MSE from Equation (2) at $\mathbf{x}_0 \in \mathcal{X}$ is

$$\begin{aligned}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) = {}& \boldsymbol{\Sigma}_{M,i}(\mathbf{x}_0,\mathbf{x}_0) - \boldsymbol{\Sigma}_{M,i}^\top(\mathbf{x}^m,\mathbf{x}_0) \\ & [\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^m,\mathbf{x}^m) + \boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^m)]^{-1}\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^m,\mathbf{x}_0) \\ & + \eta_i(\mathbf{x}_0)^\top[\mathbf{F}_i^\top(\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^m,\mathbf{x}^m) + \boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^m))^{-1}\mathbf{F}_i]^{-1}\eta_i(\mathbf{x}_0),\end{aligned} \tag{3}$$

where $\eta_i(\mathbf{x}_0) = \mathbf{f}_i(\mathbf{x}_0) - \mathbf{F}_i^\top(\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^m,\mathbf{x}^m) + \boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^m))^{-1}\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^m,\mathbf{x}_0)$.

In the following, we define some useful notation. For any finite dimensional vector $\mathbf{v}$, we let $\|\mathbf{v}\|$ be its Euclidean norm. For any generic matrix A, we use A_{ab} to denote its (a,b)-entry, cA to denote the matrix whose (a,b)-entry is cA_{ab} for any constant $c \in \mathbb{R}$. For any positive definite matrix A, let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ be its largest and smallest eigenvalues. For two sequences of positive numbers $\{a_l\}_{l\geq 1}$ and $\{b_l\}_{l\geq 1}$, $a_l \lesssim b_l$ means that $\limsup_{l\to\infty} a_l/b_l < \infty$, and $a_l \asymp b_l$ means that both $a_l \lesssim b_l$ and $b_l \lesssim a_l$ hold true.

We introduce some concepts from the reproducing kernel Hilbert space (RKHS) theory that will be used in our theorems. Let $\mathbb{P}_{\mathbf{X}}$ be a probability distribution over $\mathcal{X}$, $L_2(\mathbb{P}_{\mathbf{X}})$ be the L_2 space under $\mathbb{P}_{\mathbf{X}}$. The inner product in $L_2(\mathbb{P}_{\mathbf{X}})$ is defined as $\langle f,g\rangle_{L_2(\mathbb{P}_{\mathbf{X}})} = \mathrm{E}_{\mathbf{X}}[f(\mathbf{X})g(\mathbf{X})]$ for any $f,g \in L_2(\mathbb{P}_{\mathbf{X}})$. For any $f \in L_2(\mathbb{P}_{\mathbf{X}})$, define the linear operator $[T_{\boldsymbol{\Sigma}_M} f](\mathbf{x}) = \int_{\mathcal{X}} \boldsymbol{\Sigma}_M(\mathbf{x},\mathbf{x}') f(\mathbf{x}')\mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x}')$ for any $\mathbf{x} \in \mathcal{X}$. Because $\boldsymbol{\Sigma}_M(\cdot,\cdot)$ is a continuous symmetric nonnegative definite kernel on $\mathcal{X} \times \mathcal{X}$, there exists an orthonormal basis $\{\phi_l(\mathbf{x}) : l = 1,2,\ldots\}$ with respect to $\mathbb{P}_{\mathbf{X}}$ consisting of eigenfunctions of the linear operator $T_{\boldsymbol{\Sigma}_M}$, that is, $\int_{\mathcal{X}} \phi_l^2(\mathbf{x})d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) = 1$, $\int_{\mathcal{X}} \phi_l(\mathbf{x})\phi_{l'}(\mathbf{x})$

$d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) = 0$ for $l \neq l'$, and $[T_{\boldsymbol{\Sigma}_M}\phi_l](\mathbf{x}) = \mu_l\phi_l(\mathbf{x})$ for some eigenvalue $\mu_l \geq 0$, all $l = 1, 2, \ldots$ and $\mathbf{x} \in \mathcal{X}$. According to Mercer's theorem (theorem 4.2 of Rasmussen and Williams 2006), the kernel $\boldsymbol{\Sigma}_M$ (which can be taken as any $\boldsymbol{\Sigma}_{M,i}$ for $i = 1, \ldots, k$) has the series expansion $\boldsymbol{\Sigma}_M(\mathbf{x}, \mathbf{x}') = \sum_{l=1}^{\infty} \mu_l\phi_l(\mathbf{x})\phi_l(\mathbf{x}')$ with respect to $\mathbb{P}_{\mathbf{X}}$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, where we assume that the eigenvalues of $\boldsymbol{\Sigma}_M$ are sorted into the decreasing order $\mu_1 \geq \mu_2 \geq \cdots \geq 0$. The trace of the kernel $\boldsymbol{\Sigma}_M$ is defined as $\mathrm{tr}(\boldsymbol{\Sigma}_M) = \sum_{l=1}^{\infty} \mu_l$. Any function $f \in L_2(\mathbb{P}_{\mathbf{X}})$ has the series expansion $f(\mathbf{x}) = \sum_{l=1}^{\infty} \theta_l\phi_l(\mathbf{x})$, where $\theta_l = \langle f, \phi_l \rangle_{L_2(\mathbb{P}_{\mathbf{X}})}$. The RKHS $\mathbb{H}$ attached to the kernel $\boldsymbol{\Sigma}_M$ is the space of all functions $f \in L_2(\mathbb{P}_{\mathbf{X}})$ such that its $\mathbb{H}$-norm $\|f\|_{\mathbb{H}}^2 = \sum_{l=1}^{\infty} \theta_l^2/\mu_l < \infty$. We refer the readers to Gu (2002) and Hsing and Eubank (2015) for a complete treatment of the RKHS theory.

Based on the decaying rates of eigenvalues, most commonly used covariance functions (kernels) can be categorized into the three types described here: the finite-rank kernels, exponentially decaying kernels, and polynomially decaying kernels. For a comprehensive review of covariance functions, see chapter 4 of Rasmussen and Williams (2006).

1. *Finite-rank kernels* satisfy $\mu_1 \geq \cdots \geq \mu_{l_*} > 0$ and $\mu_{l_*+1} = \mu_{l_*+2} = \cdots = 0$ for some finite integer $l_* \in \mathbb{N}$. One example of finite-rank kernels is $\boldsymbol{\Sigma}_M(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^\top\mathbf{x}')^D$ for some fixed positive integer D and any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. The sample paths generated from this kernel are the class of all polynomial functions up to the degree D and has the finite rank at most equal to $D + 1$ (Rasmussen and Williams 2006). If $D = 1$, then $\boldsymbol{\Sigma}_M(\mathbf{x}, \cdot)$ generates the class of linear functions in $\mathbf{x}$.

2. *Exponentially decaying kernels* satisfy $\mu_l \asymp \exp(-cl^{\kappa/d})$ for some constants $c > 0, \kappa > 0$, with d being the dimension of covariate $\mathbf{x}$. The most important example is the squared exponential kernel $\boldsymbol{\Sigma}_M(\mathbf{x}, \mathbf{x}') = \exp\{-\varphi\|\mathbf{x} - \mathbf{x}'\|^2\}$ for $\varphi > 0$ and $\mathbf{x}, \mathbf{x}' \in \mathcal{X} \subseteq \mathbb{R}^d$. If $d = 1$, $\mathbb{P}_{\mathbf{X}} = N(0, (4a_1)^{-1})$ for some $a_1 > 0$, then it is known (Rasmussen and Williams 2006, section 4.3.1) that for $l = 0,\ 1, 2, \ldots,$ the eigenfunctions can be taken as $\phi_l(\mathbf{x}) = (a_2/a_1)^{1/4}\exp\{-(a_2 - a_1)\mathbf{x}^2\}H_l(\sqrt{2a_2}\mathbf{x})/\sqrt{2^l l!}$, and the corresponding eigenvalues are $\mu_l = \sqrt{2a_1/(a_1 + a_2 + \varphi)}\exp\{-l\log(1/a_3)\}$, where $a_2 = \sqrt{a_1^2 + 2a_1\varphi}$, $a_3 = \varphi/(a_1 + a_2 + \varphi) \in (0, 1)$, and $H_l(z) = (-1)^l\exp(x^2)\frac{d^l}{dx^l}\exp(-x^2)$ is the lth order Hermite polynomial. Therefore, $\mu_l \asymp \exp(-cl^{\kappa})$ holds with $c = \log(1/a_3)$ and $\kappa = 1$. In general, $\mu_l \asymp \exp(-cl^{\kappa/d})$ holds for infinitely smooth stationary kernels on a bounded domain $\mathcal{X} \subseteq \mathbb{R}^d$ (Santin and Schaback 2016).

3. *Polynomially decaying kernels* satisfy $\mu_l \asymp l^{-2\nu/d-1}$ for some constant $\nu > 0$ (such that $\mathrm{tr}(\boldsymbol{\Sigma}_M) < \infty$). One example is the kernel $\boldsymbol{\Sigma}_M(\mathbf{x}, \mathbf{x}') = \min\{\mathbf{x}, \mathbf{x}'\}$ for $\mathbf{x}, \mathbf{x}' \in \mathcal{X} = [0, 1]$. This kernel generates the first-order Sobolev class that contains all Lipschitz functions on $[0, 1]$. If $\mathbb{P}_{\mathbf{X}}$ is the uniform distribution on $[0, 1]$, then it is known that $\mu_l \asymp 1/l^4$ (Gu 2002). Another very important example is the Matérn kernel $\boldsymbol{\Sigma}_{M,i}(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)}(\sqrt{2\nu}\varphi\|\mathbf{x} - \mathbf{x}'\|)^{\nu}K_{\nu}(\sqrt{2\nu}\varphi\|\mathbf{x} - \mathbf{x}'\|)$, where K_ν is the modified Bessel function, and the smoothness parameter ν satisfies $\nu > 0$. The Matérn kernel is widely used for fitting spatial surfaces with varying roughness from ν. A smaller ν generates rougher sample paths. If $\mathcal{X} \subseteq \mathbb{R}^d$ is a bounded set, then the Matérn kernel has eigenvalues decaying as $\mu_l \leq Cl^{-2\nu/d-1}$ for some constant $C > 0$ (Santin and Schaback 2016).

2.2. Target Measures

For the estimation problem and a given covariate point sample $\mathbf{x}^m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, the optimal MSE of the linear predictor (2) for design i is $\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)$, where the test point $\mathbf{X}_0$ is randomly drawn from the same distribution $\mathbb{P}_{\mathbf{X}}$ as for $\mathbf{X}^m$. The IMSE for the ith design is the integral of $\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)$ with respect to the sampling distribution of $\mathbf{X}_0$

$$\mathrm{IMSE}_i = \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right],$$

and the maximal IMSE is defined as $\max_{i \in \{1, \ldots, k\}} \mathrm{IMSE}_i$.

Under our consideration, the maximal IMSE can be viewed as a measurement of the prediction error with the worst MSE-optimal linear predictor among the k designs over all possible locations in $\mathcal{X}$. Our goal for the estimation problem is to prove that as the simulation budget increases to infinity, the maximal IMSE decreases at a certain rate to zero, under the correct specification of Model (1) and other necessary mild technical assumptions. In particular, for the ease of presentation, we assume that all points in $\mathbf{x}^m$ receive the same number of simulation runs $n_1 = \cdots = n_m = n$; that is, we do not need to decide the number of simulation replications among different designs and covariate points. We will show that for any given n, the maximal IMSE converges to zero at some decreasing rate of m, which is the number of distinct points in $\mathbf{x}^m$. Intuitively, this goal is reasonable, because an SK model allows us to interpolate the unknown surface of $y_i(\mathbf{x})$ at a new location with higher accuracy if m becomes larger. How fast the maximal IMSE converges to zero in terms of m depends mainly on the smoothness of all the unknown true surfaces $y_i(\mathbf{x})$, $i = 1, \ldots, k$. Because we assume that the true surface $y_i(\mathbf{x})$ is correctly specified as in Model (1), then equivalently, the convergence rate of the maximal IMSE depends on the properties of the covariance kernel $\boldsymbol{\Sigma}_{M,i}(\cdot, \cdot)$ and the functions $\mathbf{f}_i(\cdot)$. The maximal IMSE is still random with respect to the covariate point sample $\mathbf{X}^m$, and our rate result for the maximal IMSE will be obtained in $\mathbb{P}_{\mathbf{X}^m}-$ probability.

For the optimization problem, given configuration of designs $M_i(\cdot)$'s and a covariate point sample $\mathbf{x}^m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, the real best design $i^\circ(\mathbf{x}_0)$ and the estimated best design $\hat{i}^\circ(\mathbf{x}_0)$ at test point $\mathbf{X}_0 = \mathbf{x}_0$ are

$$\begin{aligned} y^\circ(\mathbf{x}_0) &= \min_{i\in\{1,\ldots,k\}} y_i(\mathbf{x}_0), \qquad i^\circ(\mathbf{x}_0) \in \arg\min_{i\in\{1,\ldots,k\}} y_i(\mathbf{x}_0), \\ \hat{y}^\circ(\mathbf{x}_0) &= \min_{i\in\{1,\ldots,k\}} \hat{y}_i(\mathbf{x}_0), \qquad \hat{i}^\circ(\mathbf{x}_0) \in \arg\min_{i\in\{1,\ldots,k\}} \hat{y}_i(\mathbf{x}_0). \end{aligned} \tag{4}$$

Typically in R&S problems, the correct selection for the best design is defined as $\hat{i}^\circ(\mathbf{x}_0) = i^\circ(\mathbf{x}_0)$. However, due to the continuous nature of $\mathbf{x}_0$ in the framework of simulation with covariates, the best design $i^\circ(\mathbf{x}_0)$ might not be unique for certain values of $\mathbf{x}_0$, causing ambiguity in this definition. To solve this issue, in this research, we will focus the event of good selection (Ni et al. 2017). Similarly as in the indifference-zone (IZ) formulation for R&S problems (Kim and Nelson 2006), suppose there is an IZ parameter $\delta_0 > 0$ showing the minimal difference for the means of designs that we believe is worth detecting. A good selection for $i^\circ(\mathbf{x}_0)$ happens when the mean of the estimated best design $y_{\hat{i}^\circ(\mathbf{x}_0)}(\mathbf{x}_0)$ is better than $y^\circ(\mathbf{x}_0) + \delta_0$ for the test point $\mathbf{x}_0 \in \mathcal{X}$; equivalently, a false (not good) selection happens when $y_{\hat{i}^\circ(\mathbf{x}_0)}(\mathbf{x}_0)$ is no better than $y^\circ(\mathbf{x}_0) + \delta_0$. This definition allows some flexibility for determining the best design when the means of the top two designs are very close or exactly the same under some covariate value. Consequently, probabilities of good selection $\text{PGS}(\mathbf{x}_0)$ and false selection $\text{PFS}(\mathbf{x}_0)$ among the k alternatives at $\mathbf{x}_0$ are given by

$$\begin{aligned} \text{PGS}(\mathbf{x}_0) &= \mathbb{P}_\epsilon\left(y_{\hat{i}^\circ(\mathbf{x}_0)}(\mathbf{x}_0) - y^\circ(\mathbf{x}_0) < \delta_0\right), \\ \text{PFS}(\mathbf{x}_0) &= \mathbb{P}_\epsilon\left(y_{\hat{i}^\circ(\mathbf{x}_0)}(\mathbf{x}_0) - y^\circ(\mathbf{x}_0) \geq \delta_0\right), \end{aligned} \tag{5}$$

where $\mathbb{P}_\epsilon$ is the joint probability measure of all simulation error terms $\epsilon_{il}(\mathbf{x}_j)$ for $i = 1, \ldots, k$, $j = 1, \ldots, m$ and $l = 1, \ldots, n$. To ease the burden of notation, we hide the dependence of $\text{PGS}(\mathbf{x}_0)$ and $\text{PFS}(\mathbf{x}_0)$ on the constant IZ parameter δ_0.

Consequently, the integrated PFS is defined as

$$\text{IPFS} = \mathrm{E}_M\, \mathrm{E}_{\mathbf{X}_0}[\text{PFS}(\mathbf{X}_0)],$$

where M contains the randomness from all $M_i(\cdot)$ s, $i = 1, \ldots, k$, measuring the *extrinsic uncertainty* (Ankenman et al. 2010). Our goal for the optimization problem is to identify the convergence rate of IPFS with the number of covariate points m. Similarly as for the maximal IMSE, IPFS is still random with respect to $\mathbf{X}^m$, and our rate result for IPFS will be obtained in $\mathbb{P}_{\mathbf{X}^m}-$ probability.

There are two key differences between our setting and existing research in the simulation literature. First, we assume that $\mathbf{X}_0$ is randomly drawn from $\mathbb{P}_\mathbf{X}$, independently of the random sample $\mathbf{X}^m$. Our treatment of both $\mathbf{X}^m$ and $\mathbf{X}_0$ is different from most SK studies (Ankenman et al. 2010, Chen et al. 2013, Wang and Hu 2018), which usually treat $\mathbf{X}^m$ as fixed covariate points and $\mathbf{X}_0$ as uniformly sampled from $\mathcal{X}$. The randomness in $\mathbf{X}^m$ allows us to derive the asymptotic convergence rates of the two target measures for various types of covariance kernels.

Second, although the maximal IMSE and IPFS are expected (integrated) measures and the same in appearance to the expected measure PCS_E in the research of ranking and selection with covariates (Shen et al. 2021), the expectations in these two papers are caused by different types of randomness, leading to an intrinsic difference in the meaning and structure of these measures and the approaches used to analyze them. Shen et al. (2021) considered a fixed number of m covariate points, and the expectation in PCS_E is with respect to the random covariate points, which seeks to assess the average of selection quality over all the possible covariate values (problem instances). In this paper, expectation is with respect to the random test point, which seeks to assess the average of prediction quality over all the possible covariate values (problem instances). This research also faces the randomness of the covariate point sample $\mathbf{X}^m$, and as discussed previously, it is handled with the development of convergence rates in $\mathbb{P}_{\mathbf{X}^m}-$ probability.

3. Convergence Rates of the Maximal IMSE

In this section, we study the convergence rate of the first target measure: maximal IMSE. We make the following assumptions:

Assumption 1. *For $i = 1, \ldots, k$, Model (1) is correctly specified with $M_i(\cdot)$ being a sample path from a known covariance function $\Sigma_{M,i}(\cdot,\cdot)$. For $i = 1, \ldots, k$, $j = 1, \ldots, m$, $l = 1, \ldots, n$, $\epsilon_{il}(\mathbf{x}_j)$s are random variables with mean zero and variance $\sigma_i^2(\mathbf{x}_j)$, and they are independent across different i, j, and l. The simulation errors $\epsilon_{il}(\mathbf{x}_j)$s are independent of the Gaussian process $M_i(\mathbf{x})$ for all i, j, l, and $\mathbf{x} \in \mathcal{X}$. There exist finite constants $\underline{\sigma}_0^2$ and $\overline{\sigma}_0^2$ such that $0 < \underline{\sigma}_0^2 \leq \sigma_i^2(\mathbf{x}) \leq \overline{\sigma}_0^2$ for all i and $\mathbf{x} \in \mathcal{X}$.*

Assumption 2. *Trace class kernel: The kernel $\mathbf{\Sigma}_{M,i}$ satisfies $\mathrm{tr}\left(\mathbf{\Sigma}_{M,i}\right) < \infty$ for $i = 1, \ldots, k$.*

Assumption 3. *Basis functions: Let $\{\phi_{i,l}(\mathbf{x}) : l = 1, 2, \ldots\}$ be an orthonormal basis with respect to $\mathbb{P}_\mathbf{X}$ consisting of eigenfunctions of the linear operator $T_{\mathbf{\Sigma}_{M,i}}$. There are positive constants ρ_* and $r_* \geq 2$ common for all $i = 1, \ldots, k$ such that $\mathrm{E}_\mathbf{X}\{\phi_{i,l}^{2r_*}(\mathbf{X})\} \leq \rho_*^{2r_*}$ for every $l = 1, 2, \ldots, \infty$.*

Assumption 4. *Regressors: The regression functions satisfy $\mathrm{f}_{is} \in \mathbb{H}_i$ for all $i = 1, \ldots, k$ and $s = 1, \ldots, q$, where $\mathbb{H}_i$ the RKHS attached to kernel $\mathbf{\Sigma}_{M,i}$. Furthermore, $\lambda_{\min}(\mathrm{E}_\mathbf{X}[\mathbf{f}_i$*

$(\mathbf{X})\mathbf{f}_i(\mathbf{X})^\top]$) *is lower bounded by a positive constant for all* $i = 1, \ldots, k$ *if* $\mathbf{X}$ *follows the distribution* $\mathbb{P}_{\mathbf{X}}$.

Assumption 1 assumes independence of the simulation noise $\epsilon_{il}(\mathbf{x}_j)$ between different designs, covariate points, and replications, so we do not consider the common random number technique in the simulation experiments. An implication of this setting is that learning the performance of a design does not enable learning the performance of another design. Assumption 1 also makes a mild assumption on the second moment of the error distribution. For all derivations related to IMSE in this paper, we do not require $\epsilon_{il}(\mathbf{x}_j)$ to be normally distributed. The lower and upper bounds for the error variance are technical, which is trivially satisfied if the errors are homogeneous with a constant variance.

Assumption 2 assumes that the operator associated to the kernel $\mathbf{\Sigma}_{M,i}$ is a trace class operator (Hsing and Eubank 2015). This will be verified later for all the three types of kernels described before, in which their eigenvalues typically decrease at least polynomially and are usually summable. Assumption 3 imposes a mild moment condition on the orthonormal basis functions. Sometimes Assumption 3 can be strengthened to the assumption that the L_∞ norms of $\phi_{i,l}(\mathbf{x})$ s are uniformly bounded for all $l = 1, 2, \ldots$ and all $\mathbf{x} \in \mathcal{X}$. For example, if $\mathcal{X} = [0, 1]$ and $\mathbb{P}_{\mathbf{X}}$ is the uniform distribution on $\mathcal{X}$, then the eigenfunctions of the Matérn covariance kernel with $\nu = 1/2$ are the sine functions (section 3.4.1 of Van Trees 2001), whose L_∞ norms are naturally bounded from above by constant, so that Assumption 3 trivially holds. The quantities ρ_* and r_* do not need to depend on i, because if the ith design satisfies $\mathrm{E}_{\mathbf{X}}\{\phi_{i,l}^{2r_i}(\mathbf{X})\} \le \rho_i^{2r_i}$ for $r_i \ge 2$, one can let $r_* = \min_{i\in\{1,\ldots,k\}} r_i \ge 2$ and $\rho_* = \max\left(\max_{i\in\{1,\ldots,k\}} \rho_i, 1\right)$. By Jensen's inequality, $\mathrm{E}_{\mathbf{X}}\{\phi_{i,l}^{2r_*}(\mathbf{X})\} \le [\,\mathrm{E}_{\mathbf{X}}\{\phi_{i,l}^{2r_i}(\mathbf{X})\}]^{r_*/r_i} \le \rho_i^{2r_i \cdot r_*/r_i} \le \rho_*^{2r_*}$ and Assumption 3 holds.

Assumption 4 requires that the matrix $\mathrm{E}_{\mathbf{X}}[\mathbf{f}_i(\mathbf{X})\mathbf{f}_i(\mathbf{X})^\top]$ is nonsingular. This is a necessary condition for the identifiability of $\boldsymbol{\beta}_i$ because a singular $\mathrm{E}_{\mathbf{X}}[\mathbf{f}_i(\mathbf{X})\mathbf{f}_i(\mathbf{X})^\top]$ implies that some functions in $\{\mathrm{f}_{i1}(\mathbf{x}), \ldots, \mathrm{f}_{iq}(\mathbf{x})\}$ can be written as a linear combination of others, making it impossible to estimate $\boldsymbol{\beta}_i$. In most real applications, f_{is} s are highly smooth functions such as monomials (see p. 12 of Stein (1999) for a cogent argument). In such cases, $\mathrm{f}_{is} \in \mathbb{H}_i$ is satisfied in general. For example, if the domain $\mathcal{X}$ is a bounded set and the covariance kernel is a Matérn kernel, then $\mathbb{H}$ is norm equivalent to a Sobolev space of functions with certain smoothness. Because a monomial f_{is} is infinitely differentiable, f_{is} lies in $\mathbb{H}_i$.

We first restrict our discussion to a single SK model and drop the subscript i. From (3), for a given test point $\mathbf{x}_0$ and an SK model, we can decompose the optimal MSE into two parts:

$$\begin{aligned}
\mathrm{MSE}_{\mathrm{opt}}(\mathbf{x}_0) &= \mathrm{MSE}_{\mathrm{opt}}^{(M)}(\mathbf{x}_0) + \mathrm{MSE}_{\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{x}_0),\\
\mathrm{MSE}_{\mathrm{opt}}^{(M)}(\mathbf{x}_0) &= \mathbf{\Sigma}_M(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{\Sigma}_M^\top(\mathbf{x}^m, \mathbf{x}_0)\\
&\qquad [\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon]^{-1}\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}_0),\\
\mathrm{MSE}_{\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{x}_0) &= \eta(\mathbf{x}_0)^\top \left[\mathbf{F}^\top(\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon)^{-1}\mathbf{F}\right]^{-1} \eta(\mathbf{x}_0),
\end{aligned} \tag{6}$$

where $\eta(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top(\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon)^{-1}\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}_0)$. They are two distinct contributions to the total MSE from estimating $M(\mathbf{x})$ and $\boldsymbol{\beta}$, respectively.

The following two theorems provide upper bounds for the integrated $\mathrm{MSE}_{\mathrm{opt}}^{(M)}(\mathbf{X}_0)$ and $\mathrm{MSE}_{\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{x}_0)$ in (6). Based on them, we can analyze the convergence behavior of the integrated $\mathrm{MSE}_{\mathrm{opt}}(\mathbf{x}_0)$, and consequently the maximal IMSE.

Theorem 1. *Under Assumptions 1–3, the following relation holds*

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}^m}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] \le{}& \frac{2\overline{\sigma}_0^2}{mn}\gamma\left(\frac{\overline{\sigma}_0^2}{mn}\right)\\
&+ \inf_{\zeta\in\mathbb{N}}\left[\left\{\frac{3mn}{\overline{\sigma}_0^2}\mathrm{tr}(\mathbf{\Sigma}_M) + 1\right\}\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right)\right.\\
&\left.+ \mathrm{tr}(\mathbf{\Sigma}_M)\left\{300\rho_*^2\frac{b(m,\zeta,r_*)\gamma(\frac{\overline{\sigma}_0^2}{mn})}{\sqrt{m}}\right\}^{r_*}\right],
\end{aligned} \tag{7}$$

where

$$b(m,\zeta,r_*) = \max\left(\sqrt{\max(r_*, \log\zeta)}, \frac{\max(r_*, \log\zeta)}{m^{1/2-1/r_*}}\right),$$

$$\gamma(a) = \sum_{l=1}^{\infty}\frac{\mu_l}{\mu_l + a} \qquad \textit{for any } a > 0,$$

$$\mathrm{tr}\left(\mathbf{\Sigma}_{\mathrm{M}}^{(\zeta)}\right) = \sum_{\mathrm{l}=\zeta+1}^{\infty}\mu_1 \qquad \textit{for any } \zeta \in \mathbb{N}.$$

Theorem 1 provides an upper bound for the expectation of the IMSE $\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right]$. The reason we have another expectation $\mathrm{E}_{\mathbf{X}^m}$ before this IMSE is that $\mathbf{X}^m$ is a random sample from $\mathbb{P}_{\mathbf{X}}$, and hence, this IMSE is also random in $\mathbf{X}^m$. The upper bound in Theorem 1 takes a complicated form and some discussion is in order. First, the first term in the upper bound (7) is the dominant term, whereas the terms inside the infimum are typically of smaller stochastic orders than the first term, as we will show later in the proof of Theorem 3 for three types of kernels. Second, inside the first term in (7), the term $\gamma\left(\frac{\overline{\sigma}_0^2}{mn}\right)$ is known as the *effective dimensionality* of the kernel $\mathbf{\Sigma}_M$ with respect to $L_2(\mathbb{P}_{\mathbf{X}})$ (Zhang 2005). As we will show later in Theorem 3, the term $\frac{\overline{\sigma}_0^2}{mn}\gamma\left(\frac{\overline{\sigma}_0^2}{mn}\right)$ is the dominant term that determines the convergence rate of IMSE. Third, the terms inside the infimum sign are stochastic errors due to the randomness in $\mathbf{X}^m$, and under Assumptions 1–3, they are of negligible orders by choosing a proper $\zeta \in \mathbb{N}$.

For two random variables U_m and V_m that are measurable with respect to the sigma-algebra generated by $\mathbf{X}^m$, we use $U_m \lesssim_{\mathbb{P}_{\mathbf{X}^m}} V_m$ to denote the relation that $|U_m/V_m|$ is bounded in $\mathbb{P}_{\mathbf{X}^m}-$ probability.

Theorem 2. *Under Assumptions 1–4, the following relation holds:*

$$\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}}$$
$$\frac{8q\,\mathrm{tr}(\boldsymbol{\Sigma}_M)}{\lambda_{\min}(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top])}\left\{8C_{\mathrm{f}}^2\frac{\overline{\sigma}_0^2}{mn}+\inf_{\zeta\in\mathbb{N}}\left[8C_{\mathrm{f}}^2\frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4\,\mathrm{tr}(\boldsymbol{\Sigma}_M)\,\mathrm{tr}\left(\boldsymbol{\Sigma}_M^{(\zeta)}\right)\right.\right.$$
$$\left.\left.+\,C_{\mathrm{f}}^2\,\mathrm{tr}\left(\boldsymbol{\Sigma}_M^{(\zeta)}\right)+C_{\mathrm{f}}^2\,\mathrm{tr}(\boldsymbol{\Sigma}_M)\left\{200\rho_*^2\frac{b(m,\zeta,r_*)\gamma\left(\frac{\overline{\sigma}_0^2}{mn}\right)}{\sqrt{m}}\right\}^{r_*}\right]\right\}, \tag{8}$$

where $C_{\mathrm{f}} = \max_{1\le s\le q}\|\mathrm{f}_s\|_{\mathbb{H}}$, $b(m,\zeta,r_*)$ and $\gamma(\cdot)$ are defined in Theorem 1.

Similar to the upper bound in Theorem 1, the terms inside the infimum can be made negligible compared with the leading term of $\frac{\overline{\sigma}_0^2}{mn}$ by choosing a proper $\zeta \in \mathbb{N}$. The upper bound in Theorem 2 is a bound in probability, which means that as $m \to \infty$, the IMSE in (8) is upper bounded in probability by the right-hand side. It is slightly weaker than the upper bound on the expectation of IMSE in Theorem 1 but suffices for deriving the convergence rate of the maximal IMSE.

The following theorem gives our main rate result on the maximal IMSE.

Theorem 3. *Suppose that all k designs have the sampling distribution $\mathbb{P}_{\mathbf{X}}$ for $\mathbf{X}^m$ and $\mathbf{X}_0$. Under Assumptions 1–4, the following results hold with r_* given in Assumption 3:*

(i) *(Finite-rank kernels) If for every $i = 1,\dots,k$, $\boldsymbol{\Sigma}_{M,i}$ is a finite-rank kernel of rank l_{*i}; that is, its eigenvalues satisfy $\mu_{i,1} \ge \mu_{i,2} \ge \cdots \ge \mu_{i,l_{*i}} > 0$ and $\mu_{i,l_{*i}+1} = \mu_{i,l_{*i}+2} = \cdots = 0$, then as $m \to \infty$,*

$$\max_{i\in\{1,\dots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} R^F(m,n)$$
$$\equiv \max\left(\frac{1}{mn},\frac{1}{m^{\frac{r_*}{2}}}\right). \tag{9}$$

(ii) *(Exponentially decaying kernels) If for every $i = 1,\dots,k$, $\boldsymbol{\Sigma}_{M,i}$ is a kernel with eigenvalues satisfying $\mu_{i,l} \le c_{1i}\exp\left(-c_{2i}l^{\kappa_i/d}\right)$ for some constants $c_{1i} > 0$, $c_{2i} > 0$, $\kappa_i > 0$ and all $l \in \mathbb{N}$. Let $\kappa_* = \min_{i\in\{1,\dots,k\}}\kappa_i$. Then, as $m \to \infty$,*

$$\max_{i\in\{1,\dots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} R^E(m,n)$$
$$\equiv \max\left\{\frac{(\log(mn))^{\frac{d}{\kappa_*}}}{mn},\frac{(\log(mn))^{\frac{r_*(\kappa_*+d)}{\kappa_*}}}{m^{\frac{r_*}{2}}}\right\}. \tag{10}$$

(iii) *(Polynomially decaying kernels) If for every $i = 1,\dots,k$, $\boldsymbol{\Sigma}_{M,i}$ is a kernel with eigenvalues satisfying $\mu_{i,l} \le c_i l^{-2\nu_i/d-1}$ for some constants $\nu_i > d/2$, $c_i > 0$ and all $l \in \mathbb{N}$. Let $\nu_* = \min_{i\in\{1,\dots,k\}}\nu_i$. Then, as $m \to \infty$,*

$$\max_{i\in\{1,\dots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} R^P(m,n)$$
$$\equiv \max\left\{\frac{1}{(mn)^{\frac{2\nu_*}{2\nu_*+d}}},\frac{n^{\frac{dr_*}{2\nu_*+d}}(\log(mn))^{r_*}}{m^{\frac{r_*(2\nu_*-d)}{2\nu_*+d}}}\right\}. \tag{11}$$

Remark 1 (Simplified Convergence Rates for Fixed *n*). The convergence rates of the maximal IMSE for the three types of kernels in Theorem 3 appear somehow complicated. However, because we perform the same number of simulation replications n for each pair of covariate point and design, we can simplify the rate results by considering a *fixed* n and an increasing m (to infinity). If $r_* > 2$ in Assumption 3, then the larger terms in (9) and (10) are the first terms in the brackets; if $r_* > \frac{2\nu_*}{2\nu_*-d}$ in Case (iii), then the larger term in (11) is also the first term. By dropping the fixed constant of n, the convergence rates for the three kernels in Theorem 3 can be simplified to $1/m$ for Case (i), $(\log m)^{\frac{d}{\kappa_*}}/m$ for Case (ii), and $m^{-\frac{2\nu_*}{2\nu_*+d}}$ for Case (iii).

The convergence rates of the maximal IMSE have been derived based on the upper bounds of $\mathrm{E}_{\mathbf{X}^m}\,\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right]$ and $\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right]$ in Theorems 1 and 2. These rates are generally tight and cannot be improved. In Remark 2, we discuss the finite-rank kernels and formally prove in Theorem 4 that the rate function $R^F(m,n)$ is optimal in the sense that it cannot be improved further.

Remark 2 (Example of a Finite-Rank Kernel). To illustrate the tightness of the bounds in Theorem 3, we show that the rate $1/(mn)$ in (9) can be attained for fixed n as $m \to \infty$. For simplicity, we assume that in Model (1), $\mathbf{f}_i(\mathbf{x}) \equiv 0$ and $\epsilon_{il}(\mathbf{x})$ is a homogeneous white noise process with mean 0 and a common constant variance $\sigma^2 > 0$ for $l = 1,2,\dots,n$, $i = 1,\dots,k$, and $\mathbf{x} \in \mathcal{X}$. Thus, the model becomes $\overline{Y}_i(\mathbf{x}_j) = M_i(\mathbf{x}_j) + \overline{\epsilon}_i(\mathbf{x}_j)$ for $j = 1,\dots,m$ and $i = 1,\dots,k$. Let $\mathcal{X} \subseteq \mathbb{R}^d$, and let the ith covariance kernel be $\boldsymbol{\Sigma}_{M,i}(\mathbf{x},\mathbf{x}') = a_i(\mathbf{x}^\top\mathbf{x}' + b_i)$ for some known constants $a_i > 0$ and $b_i > 0$, $i = 1,\dots,k$. We analyze the MSE-optimal linear predictor in (2) and the asymptotic behavior of the optimal MSE in (3).

Theorem 4. (Exact Rate for a Finite-Rank Kernel). *Suppose that the covariance kernels are $\boldsymbol{\Sigma}_{M,i}(\mathbf{x},\mathbf{x}') = a_i(\mathbf{x}^\top\mathbf{x}' + b_i)$ for $\mathbf{x},\mathbf{x}' \in \mathcal{X} \subseteq \mathbb{R}^d$, known constants $a_i > 0, b_i > 0$ and $i = 1,\dots,k$. Under Assumptions 1–4 and the model setup described previously, the MSE-optimal linear predictor in (2) and the optimal MSE in (3) are given by*

$$\hat{y}_i(\mathbf{x}_0) = a_i\tilde{\mathbf{x}}_{i,0}^\top\mathbf{Z}_i^\top\left(a_i\mathbf{Z}_i\mathbf{Z}_i^\top+\frac{\sigma^2}{n}\mathbf{I}_m\right)^{-1}\overline{\mathbf{Y}}_i,$$
$$\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) = a_i\tilde{\mathbf{x}}_{i,0}^\top\left(\mathbf{I}_{d+1}+\frac{a_in}{\sigma^2}\mathbf{Z}_i^\top\mathbf{Z}_i\right)^{-1}\tilde{\mathbf{x}}_{i,0}, \tag{12}$$

for any $\mathbf{x}_0 \in \mathbb{R}$ *and* $i = 1, \ldots, k$, *where* $\mathbf{I}_l$ *is the* $l \times l$ *identity matrix, and*

$$\overline{\mathbf{Y}}_i = (\overline{Y}_i(\mathbf{x}_1),\ldots,\overline{Y}_i(\mathbf{x}_m))^\top \in \mathbb{R}^m,$$

$$\tilde{\mathbf{x}}_{i,0} = \begin{pmatrix}\sqrt{b_i}\\ \mathbf{x}_0\end{pmatrix} \in \mathbb{R}^{d+1}, \quad \mathbf{Z}_i = \begin{pmatrix}\sqrt{b_i} & \cdots & \sqrt{b_i}\\ \mathbf{x}_1 & \cdots & \mathbf{x}_m\end{pmatrix}^\top \in \mathbb{R}^{m\times(d+1)}.$$

Let $\mathbb{P}_{\mathbf{X}}$ *be any sampling distribution on* $\mathbb{R}^d$ *for* $\mathbf{X}_1, \ldots, \mathbf{X}_m, \mathbf{X}_0$, *and assume that its second moment* $\mathrm{E}_{\mathbf{X}_0}(\mathbf{X}_0\mathbf{X}_0^\top)$ *exists. Then as* $m \to \infty$,

$$mn \cdot \max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \to (d+1)\sigma^2, \quad \textit{almost surely in } \mathbb{P}_{\mathbf{X}^m}. \tag{13}$$

Theorem 4 shows that the maximal IMSE of the covariance kernel $\mathbf{\Sigma}_{i,M}(\mathbf{x},\mathbf{x}') = a_i(\mathbf{x}^\top\mathbf{x}' + b_i)$ decreases asymptotically at the rate $(d+1)\sigma^2/(mn)$. For fixed n, this has shown that the rate $1/m$ given in (9) for finite-rank kernels is tight and cannot be improved.

4. Convergence Rates of IPFS

We next consider the problem of selecting the best design from the k alternatives, with their mean functions given in Model (1), and study how fast $\mathrm{PFS}(\mathbf{X}_0)$ converges to zero (or equivalently, how fast $\mathrm{PGS}(\mathbf{X}_0)$ converges to one). Similar to the analysis of the maximal IMSE before, the convergence rate here is again in the average sense, by taking expectations of $\mathrm{PFS}(\mathbf{X}_0)$ under three probability measures: (i) the joint Gaussian measure on $M_i(\cdot)$ $(i = 1, \ldots, k)$, denoted by $\mathbb{P}_M$ (with the expectation denoted by E_M), induced by the k independent Gaussian processes with mean zero and covariance function $\mathbf{\Sigma}_{M,i}(\cdot,\cdot)$ for $i = 1, \ldots, k$; (ii) the probability measure of the testing point $\mathbb{P}_{\mathbf{X}_0}$; and (iii) the probability measure of the sample $\mathbb{P}_{\mathbf{X}^m}$.

In the following, $R(m,n)$ refers to the rate function of the maximal IMSE, which becomes $R^F(m,n)$, $R^E\,(m,n)$ or $R^P(m,n)$ under the corresponding kernels in Theorem 3. The following additional assumptions will lead to faster convergence rates of PFS in some particular scenarios.

Assumption 5. *The simulation errors* $\epsilon_{il}(\mathbf{x})$*'s are independent normal random variables following* $N(0,\sigma_i^2(\mathbf{x}))$ *for all* $i = 1, \ldots, k$, $l = 1, \ldots, n$ *and* $\mathbf{x} \in \mathcal{X}$.

Assumption 6. *For any given* $\xi \in (0, 1/2)$, *there exist constants* $w_1 > 0, w_2 > 0, m_0 \geq 1$ *that depend on* ξ, *such that for* $m \geq m_0$, *for any* $t > 0$,

$$\mathbb{P}_{\mathbf{X}^m}\left\{\mathbb{P}_{\mathbf{X}_0}\left(\frac{\max_{i\in\{1,\ldots,k\}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)}{R(m,n)} \geq t\right) \leq w_1 \exp\left(-w_2 t\right)\right\} \geq 1-\xi. \tag{14}$$

Assumption 7. *For any given* $\xi \in (0, 1/2)$, *there exist constants* $w_3 > 0, m_0 \geq 1$ *that depend on* ξ, *such that for* $m \geq m_0$,

$$\mathbb{P}_{\mathbf{X}^m}\left\{\frac{\max_{i\in\{1,\ldots,k\}}\sup_{\mathbf{x}_0\in\mathcal{X}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)}{R(m,n)} \leq w_3\right\} \geq 1-\xi. \tag{15}$$

Although Assumption 5 is stronger than Assumption 1 by assuming normal observation noises, it is a common assumption in simulation-based optimization problems. We emphasize that the normality assumption in Assumption 5 is only needed for deriving tighter and exponentially small bounds for IPFS in Theorem 5. Without Assumption 5, we can still establish convergence rates of IPFS directly from the convergence rates of IMSE in Theorem 3 (see Theorem 5, Part (i)). Assumption 6 requires that the maximum of the k MSE's decays at an exponential rate with a high probability. This is often the case when the MSE is distributed like chi-square with an exponentially decaying right tail. Assumption 7 is an alternative condition stronger than Assumption 6, requiring that the supremum of MSE over $\mathcal{X}$ to be bounded with a high probability. Both Assumptions 6 and 7 can be rigorously verified for the finite-rank kernel in Remark 2 and Theorem 4 (see Theorem 6 and its proof in the online supplement). Assumption 5 together with either Assumption 6 or Assumption 7 will allow tighter bounds for the tail probability of PFS and hence sharpened convergence rates of IPFS, as shown in the next theorem.

Theorem 5. *Suppose that all the k designs have the sampling distribution* $\mathbb{P}_{\mathbf{X}}$ *for* $\mathbf{X}^m$ *and* $\mathbf{X}_0$. *Let* δ_0 *be the IZ parameter in the definition of* $\mathrm{PFS}(\mathbf{X}_0)$.

(i) *If Assumptions 1–4 hold, then as* $m \to \infty$, $\mathrm{E}_M\,\mathrm{E}_{\mathbf{X}_0}[\mathrm{PFS}(\mathbf{X}_0)] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} R(m,n)$;

(ii) *If Assumptions 1–6 hold, then as* $m \to \infty$,

$$\mathrm{E}_M\,\mathrm{E}_{\mathbf{X}_0}[\mathrm{PFS}(\mathbf{X}_0)] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \exp\left\{-\frac{1}{2}w_2^{1/2}\delta_0[R(m,n)]^{-1/2}\right\},$$

where w_2 *is given in Assumption 6;*

(iii) *If Assumptions 1–5 and Assumption 7 hold, then as* $m \to \infty$,

$$\mathrm{E}_M\,\mathrm{E}_{\mathbf{X}_0}[\mathrm{PFS}(\mathbf{X}_0)] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \exp\left\{-\frac{1}{4}w_3^{-1}\delta_0^2[R(m,n)]^{-1}\right\},$$

where w_3 *is given in Assumption 7.*

The convergence rates of IPFS in Theorem 5 include the measure $\mathbb{P}_M$ and its expectation E_M, mainly for the convenience of technical treatment, so that our result is general and does not depend on the particular shapes of the $M_i(\cdot)$ functions.

Theorem 5 provides three convergence rates, from slower to faster, under sequentially stronger sets of assumptions. In Part (i), if we only assume Assumption 1–4 without the normality assumption on error terms, then by a direct application of Markov's inequality, the convergence rate of IPFS is at least as fast as that of the maximal IMSE given in Theorem 3. If the covariance kernels of the k

designs belong to one of the three types of kernels described before, then when n is fixed, we know from Theorem 3 and Remark 1 that $R(m, n)$ converges to zero at the rate of $1/m$, $(\log m)^{\frac{d}{\kappa_*}}/m$ and $m^{-\frac{2\nu_*}{2\nu_*+d}}$ for the three types of kernels, respectively. As a result, Part (i) of Theorem 5 implies that these polynomial rates for IMSE also hold for IPFS (and IPGS): When n is fixed, IPFS converges to zero (and the IPGS converges to one) at least polynomially fast in m, at least at the rate of $1/m$, $(\log m)^{\frac{d}{\kappa_*}}/m$ and $m^{-\frac{2\nu_*}{2\nu_*+d}}$ for the three types of kernels, respectively.

In Part (ii) of Theorem 5, the additional normality assumption of Assumption 5 and Assumption 6 provide sharpened convergence rates of IPFS compared with Part (i), from the polynomial rate in Part (i) to an exponential rate. In particular, following Theorem 3 and Remark 1, if n is fixed and $R(m, n)$ converges to zero at the rate of $1/m$, $(\log m)^{\frac{d}{\kappa_*}}/m$ and $m^{-\frac{2\nu_*}{2\nu_*+d}}$ for the three types of kernels, respectively, then Part (ii) of Theorem 5 implies that the IPFS converges to zero (and the IPGS converges to one) at least exponentially fast in m, at least at the rate of $\exp(-c\sqrt{m})$, $\exp(-c\sqrt{m}(\log m)^{-\frac{d}{2\kappa_*}})$, and $\exp(-cm^{\frac{\nu_*}{2\nu_*+d}})$ for the three types of kernels, respectively, where the constant $c = w_2^{1/2}\delta_0/2$.

In Part (iii) of Theorem 5, the additional Assumptions 5 and 7 provide even more sharpened convergence rates of IPFS than in Part (ii). Following Theorem 3 and Remark 1, if n is fixed and $R(m, n)$ converges to zero at the rate of $1/m$, $(\log m)^{\frac{d}{\kappa_*}}/m$ and $m^{-\frac{2\nu_*}{2\nu_*+d}}$ for the three types of kernels, respectively, then Part (iii) of Theorem 5 implies that the IPFS converges to zero (and the IPGS converges to one) at least exponentially fast in m, at least at the rate of $\exp(-cm)$, $\exp(-cm(\log m)^{-\frac{d}{\kappa_*}})$ and $\exp(-cm^{\frac{2\nu_*}{2\nu_*+d}})$ for the three types of kernels, respectively, where the constant $c = w_3^{-1}\delta_0^2/4$. Each of these exponential rates converges to zero faster than the corresponding exponential rate from Part (ii).

Remark 3. Parts (ii) and (iii) of Theorem 5 show that under additional assumptions on the distribution of simulation noises and tails of $\max_{i\in\{1,\ldots,k\}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)$ and $\max_{i\in\{1,\ldots,k\}}\sup_{\mathbf{x}_0\in\mathcal{X}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)$, the convergence rate of IPFS can be exponentially fast. This is distinguished from the well-established exponential convergence rate of the PFS in R&S by comparing sample means of different designs (Dai 1996, Glynn and Juneja 2004). In those studies, PFS is reduced by increasing the number of simulation replications for each design instead of increasing the number of covariate points, and its exponential convergence rate takes the form of $\exp(-\varrho n_{tot})$, where n_{tot} is the total number of simulation samples and ϱ is related to some large-deviations rate function.

Remark 4 (On the Independence Across Different Designs). In the development of convergence rates of the two target measures, we have assumed in Assumption 1 that the simulation samples are independent across different designs i. This assumption is naturally the case when the designs are categorical, for example, when the designs are the treatment methods for a certain disease. However, when the designs are represented as vectors in a metric space, they usually demonstrate spatial correlation, that is, designs that are close to each other tend to have similar performance. For this case, our method and analysis can still be applied, but if the model can capture this spatial correlation between designs, it might lead to higher convergence rates for the maximal IMSE and IPFS. A possible way to do it is to build one SK that includes both the covariates and designs as inputs for predicting the system performance. That model is substantially different from ours, and further investigation along this direction is beyond the scope of this paper.

Remark 5 (On the Choices of m and n). In Theorems 1–5, we have assumed that the number of replications n_i for covariate points of design i remains the same across different designs. In practice, it is possible that the decision maker wants to unevenly allocate the simulation samples among the designs to optimize some target measures. In this case, n_is are no longer identical to each other. It falls in the well-established problem of R&S in simulation. For this purpose, our analysis can still be applied. We will discuss this direction in Section 4 of the online supplement.

When all the covariate points receive the same number of replications n, we can see that in all three cases of Theorem 3, the first term inside the maximum function in the rate expression is always a function of $n_c = mn$, whereas the second term depends on m and n separately. To make the maximal IMSE and IPFS decrease as fast as possible, we need to make the second term as small as possible, which means that for all three cases of Theorem 3, the best choice is to set $n = O(1)$, such that m increases in the same order as n_c. Intuitively, this is because the maximal IMSE involves averaging MSE over all potential location $\mathbf{x}_0 \in \mathcal{X}$, and we should use as many distinct covariate points as possible to cover more locations in $\mathcal{X}$. We emphasize that this analysis on the orders of m and n is only in the asymptotic sense based on our theoretical upper bounds.

Remark 6 (Determining the Value of m). When n_is are of a constant order, Theorems 3 and 5 imply that the maximal IMSE and IPFS decrease no slower than a polynomial order of m. This theory supports a natural procedure to determine the number of covariate points m. First, for given m and n_is, the maximal IMSE and IPFS can be either calculated by numerical integration or approximated by simple Monte Carlo estimators (see Section 3 of the online supplement). Second, after we fit a sequence of SK models with different sample sizes m, we can further fit a linear regression model with the logarithm of the maximal IMSE or IPFS as the response variable and $\log m$ as the predictor. Third, based on this fitted linear model, we reversely solve for the sample size m^* such that the maximal IMSE or IPFS hits a small prespecified target precision. This simple procedure for determining m is often accurate with IMSE and can be slightly conservative with IPFS because sometimes IPFS can decay exponentially fast in m as shown in Theorem 5. We will illustrate the practical implementation of this procedure in Section 5.3 of the online supplement.

5. Numerical Experiments

In this section, we adopt two benchmark functions and an M/M/1 queue example for numerical testing. These experiments can provide concrete presentation for the rates of the maximal IMSE and IPFS and show the impact of the factors such as the problem structure, covariance kernel, dimension of the covariate space, number of simulation replications, and sampling distribution on the convergence rates. The code is available for downloading in Li et al. (2022).

For all the experiments, we implement four types of covariance kernels ($\|\cdot\|$ denotes the Euclidean distance):

(i) Squared exponential kernel: $\boldsymbol{\Sigma}_M(\mathbf{x},\mathbf{x}') = \tau^2\exp\{-\varphi\|\mathbf{x}-\mathbf{x}'\|^2\}$, for $\mathbf{x},\mathbf{x}' \in \mathcal{X}$, $\tau^2 > 0$, and $\varphi > 0$.

(ii) Matérn kernel with smoothness $\nu = 5/2$: $\boldsymbol{\Sigma}_M(\mathbf{x},\mathbf{x}') = \tau^2(1+\sqrt{5}\varphi\|\mathbf{x}-\mathbf{x}'\| + \frac{5}{3}\varphi^2\|\mathbf{x}-\mathbf{x}'\|^2)\cdot\exp\{-\sqrt{5}\varphi\|\mathbf{x}-\mathbf{x}'\|\}$, for $\mathbf{x},\mathbf{x}' \in \mathcal{X}$, $\tau^2 > 0$, and $\varphi > 0$.

(iii) Matérn kernel with smoothness $\nu = 3/2$: $\boldsymbol{\Sigma}_M(\mathbf{x},\mathbf{x}') = \tau^2(1+\sqrt{3}\varphi\|\mathbf{x}-\mathbf{x}'\|)\cdot\exp\{-\sqrt{3}\varphi\|\mathbf{x}-\mathbf{x}'\|\}$, for $\mathbf{x}$, $\mathbf{x}' \in \mathcal{X}$, $\tau^2 > 0$, and $\varphi > 0$.

(iv) Exponential kernel (Matérn kernel with smoothness $\nu = 1/2$): $\boldsymbol{\Sigma}_M(\mathbf{x},\mathbf{x}') = \tau^2\exp\{-\varphi\|\mathbf{x}-\mathbf{x}'\|\}$, for $\mathbf{x},\mathbf{x}' \in \mathcal{X}$, $\tau^2 > 0$, and $\varphi > 0$.

Similar to Ankenman et al. (2010), the covariance matrices $\boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^m)$ s are estimated by $\mathrm{diag}\{\tilde{\sigma}_i^2(\mathbf{x}_1)/n_1, \ldots, \tilde{\sigma}_i^2(\mathbf{x}_m)/n_m\}$, where $\tilde{\sigma}_i^2(\mathbf{x}_j)$ $(j = 1,\ldots,m)$ are estimated by the least-squares method based on the sample variances $\widehat{\sigma}_i^2(\mathbf{x}_j) = (n_j-1)^{-1}\sum_{l=1}^{n_j}[Y_{il}(\mathbf{x}_j)-\overline{Y}_i(\mathbf{x}_j)]^2$ $(j = 1,\ldots,m)$. Then given the estimated $\boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^m)$ s, for each of the four kernels, we estimate the parameters φ and τ^2 by the maximum likelihood estimation. The squared exponential kernel (i) belongs to the exponentially decaying kernels and the other three kernels (ii)–(iv) belong to the polynomially decaying kernels. The smoothness of sample paths decreases from kernel (i) to kernel (iv), with (i) giving the smoothest sample paths and (iv) giving the roughest sample paths.

In all later experiments, we compute the estimated MSE at a single point $\mathbf{x}_0$ by the formula $\widehat{\mathrm{MSE}}(\mathbf{x}_0) = [\hat{y}(\mathbf{x}_0) - y(\mathbf{x}_0)]^2$, where $y(\mathbf{x}_0)$ is the true function value at $\mathbf{x}_0$ and $\hat{y}(\mathbf{x}_0)$ is the fitted mean function. To evaluate the IMSE $\mathrm{E}_{\mathbf{X}_0}[\mathrm{MSE}_{\mathrm{opt}}(\mathbf{X}_0)]$ over the domain $\mathcal{X}$, we sample T points of $\mathbf{x}_0$ from $\mathcal{X}$ according to the distribution $\mathbb{P}_{\mathbf{X}}$ and average their estimated MSEs $\widehat{\mathrm{MSE}}(\mathbf{x}_0)$. In our experiments, T is chosen as 10^3, 10^4, or 10^5, depending on the dimension of $\mathbf{x}$. Monte Carlo estimates based on this setting of T are in general accurate enough. Similarly, for each of the T testing locations $\mathbf{x}_0$, we compute the true minimum mean performance $y^\circ(\mathbf{x}_0)$ and the estimated minimum mean performance $\hat{y}^\circ(\mathbf{x}_0)$ according to (4). Then the IPFS $\mathrm{E}_{\mathbf{X}_0}[\mathrm{PFS}(\mathbf{X}_0)]$ is computed by averaging over the T points drawn from $\mathbb{P}_{\mathbf{X}}$.

5.1. Benchmark Functions

We consider the following common benchmark functions. In all cases, $\mathbf{x} = (x_1,\ldots,x_d)^\top \in \mathbb{R}^d$ is the covariate, $\mathbf{z}_i \in \mathbb{R}^d$ s are the "solutions" that index the different designs, and $\epsilon(\mathbf{x})$ is an independent noise normally distributed as $N(0,(\sqrt{2})^2)$.

1. De Jong's function:

$$Y(\mathbf{x}) = M(\mathbf{x}) + \epsilon(\mathbf{x}) = \sum_{l=1}^{d}(x_l - z_l)^2 + \epsilon(\mathbf{x}). \tag{16}$$

For function $M(\mathbf{x})$, the global minimum $\mathbf{x}^*$ is obtained at $x_l = z_l$, $l = 1,2,\ldots,d$ with $M(\mathbf{x}^*) = 0$. We consider 10 discrete designs with the ith design $\mathbf{z}^i = (\underbrace{i,\ldots,i}_{d})$, $i = 1,2,\ldots,10$.

2. Griewank's function:

$$Y(\mathbf{x}) = M(\mathbf{x}) + \epsilon(\mathbf{x})$$
$$= \frac{1}{4000}\sum_{l=1}^{d}(x_l - z_l)^2 - \prod_{l=1}^{d}\cos\left(\frac{x_l - z_l}{\sqrt{l}}\right) + 1 + \epsilon(\mathbf{x}). \tag{17}$$

For function $M(\mathbf{x})$, the global minimum $\mathbf{x}^*$ is obtained at $x_l = z_l$, $l = 1,2,\ldots,d$ with $M(\mathbf{x}^*) = 0$. We consider 10 discrete designs with the ith design $\mathbf{z}^i = (\underbrace{i,\ldots,i}_{d})$, $i = 1,2,\ldots,10$.

The performance of these functions depends on both the covariate $\mathbf{x}$ and design (solution) $\mathbf{z}$. We denote $y(\mathbf{x})$ to highlight the input $\mathbf{x}$ to the SK model.

In this numerical test, we consider the De Jong's functions with d = 1 and 3 and the Griewank's functions with d = 1 and 10. To better understand the two test functions, we have provided plots of them in Section 5.1 of the online supplement. The De Jong's functions are relatively smooth. The Griewank's functions are highly nonlinear with many oscillations, which brings difficulty to SK modeling when the number of covariate points m is small.

We consider three sampling distributions for $\mathbf{X}^m$: uniform, truncated normal and normal distributions. The covariate space is $\mathcal{X} = [1,10]^d$ when d = 1, is $\mathcal{X} = [1,4]^d$ when d = 3, 10 for the uniform and truncated normal sampling, and is $\mathcal{X} = \mathbb{R}^d$ for the normal sampling. For the truncated normal distribution, the mean and variance on each dimension are $(5.5, 7^2)$ when d = 1 and $(2.5, 3^2)$ when d = 3, 10. The normal distribution on each dimension is $N(5.5, (\sqrt{3})^2)$ when d = 1 and $N(2.5, 1^2)$ when d = 3, 10.

We let the number of covariate points m increase geometrically from m = 5 to m = 100 in the set $\{5, 8, 12, 18, 28, 42, 65, 100\}$, roughly with the common ratio of 1.53 when d = 1. When d = 3, m increases from m = 5 to m = 280 in the set $\{5, 9, 16, 27, 50, 87, 155, 280\}$, roughly with the common ratio of 1.77; when d = 10, m increases from m = 5 to m = 1,000 in the set $\{5, 11, 23, 49, 103, 220, 470, 1000\}$, roughly with the common ratio of 2.13. We fix the number of replications at each $\mathbf{x}$ for all designs at n = 10. For the indifference-zone parameter δ_0, we set $\delta_0 = 0.05$ for the one-dimensional De Jong's functions, $\delta_0 = 0.1$ for the one-dimensional Griewank's functions and three-dimensional De Jong's functions, and $\delta_0 = 0.2$ for the ten-dimensional Griewank's functions. The maximal IMSE and IPFS in all cases are estimated by the average of 100 macro Monte Carlo replications. The convergence rates of the two measures under different sampling distributions, test functions, and covariance kernels are illustrated in Figures 1–4. In the legends, `SqExp` means the squared exponential kernel, `Matern` 5/2 means the Matérn kernel with $\nu = 5/2$, `Matern` 3/2 means the Matérn kernel with $\nu = 3/2$, and `Exp` means the exponential kernel.

In terms of convergence patterns, the maximal IMSE decreases as m increases in all cases, and the decreasing trends are very close to linear when m exceeds 28 with d = 1, 3, and 103 with d = 10. Because the maximal IMSE and m are plotted on logarithmic scales, it implies that when m is large enough, the maximal IMSE decreases polynomially with m. This observation agrees with our rate results in Theorem 3. The IPFS also decreases as m increases in all cases, and the convergence rates are no slower than those of the maximal IMSE. In some cases, such as the uniform and truncated normal sampling on the 10-dimensional Griewank's function, the decreasing trends of the logarithmic IPFS are superlinear, suggesting that the IPFS might enjoy convergence rates faster than polynomial. These observations agree with the rate results in Theorem 5.

Figure 1. (Color online) One-Dimensional De Jong's Functions: Maximal IMSE and IPFS Under Different Covariance Kernels and Sampling Distributions

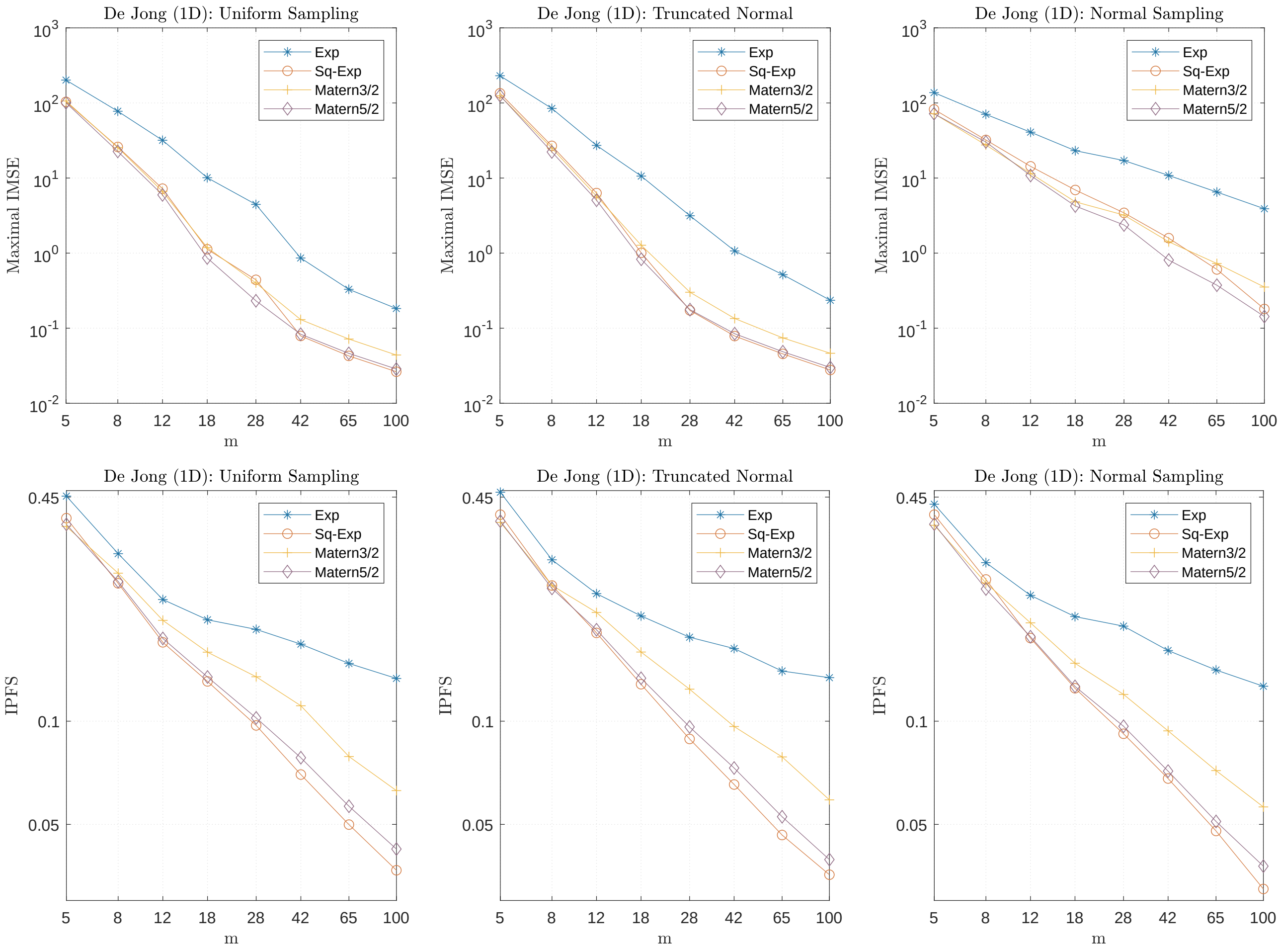
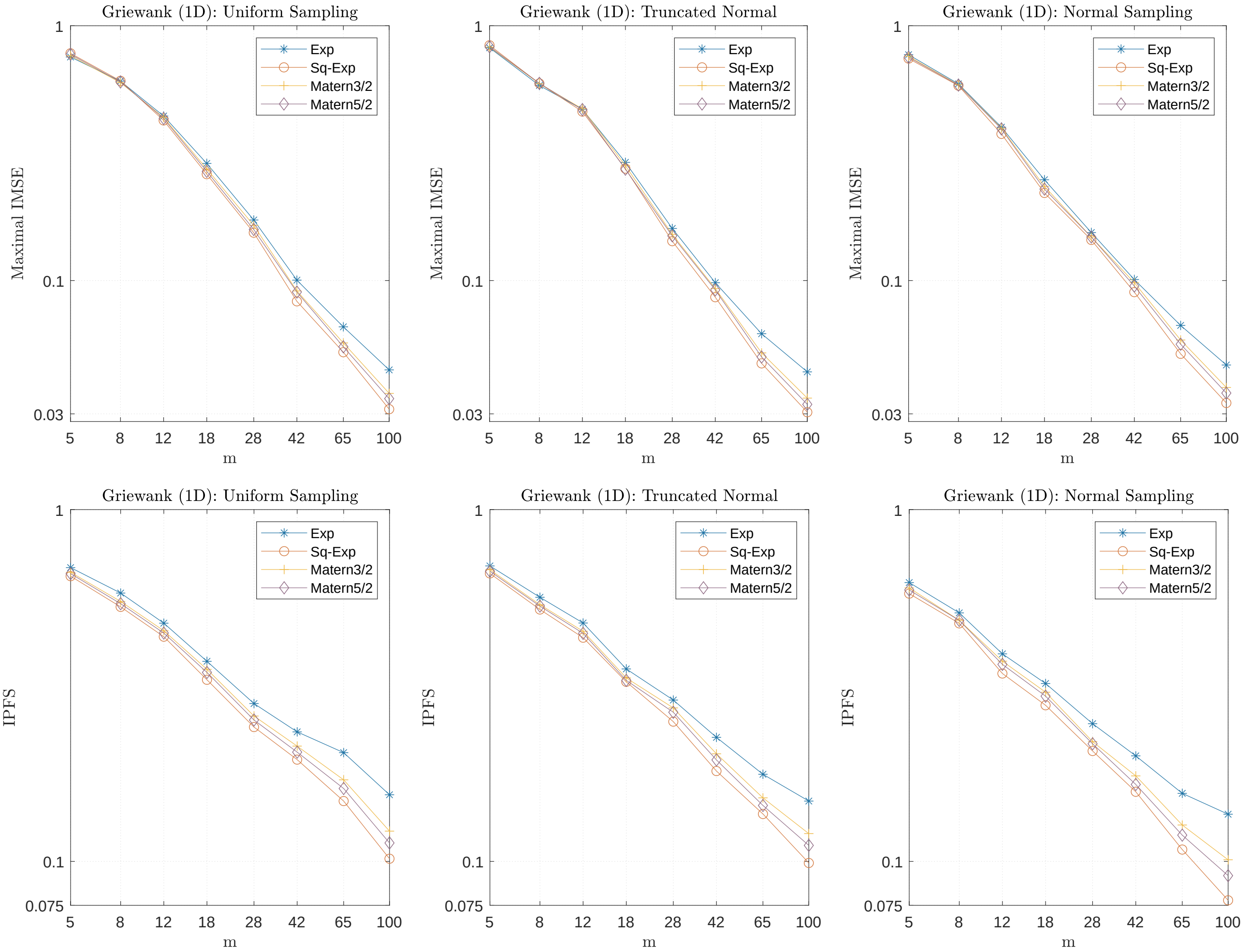


Figure 2. (Color online) One-Dimensional Griewank's Functions: Maximal IMSE and IPFS Under Different Covariance Kernels and Sampling Distributions



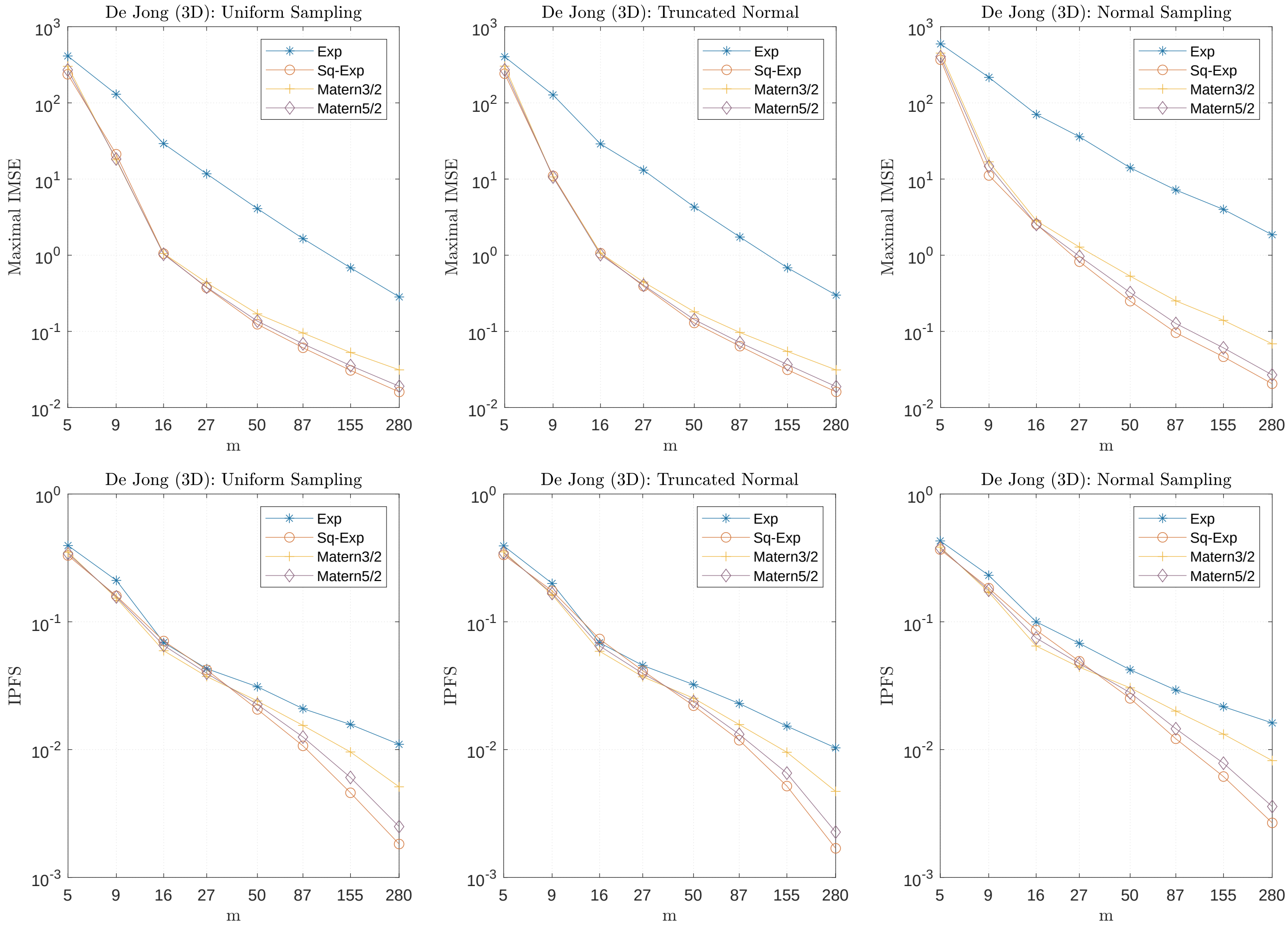
Comparing the performances of the four covariance kernels, we can observe that the exponential kernel performs the worst with the largest maximal IMSE and IPFS in all tested cases, and its disadvantage is more obvious on the De Jong's function. This is mainly because the sample paths from the exponential kernel are rough (continuous but not differentiable), whereas the De Jong's function is very smooth. This mismatch creates bad fitting and predictions, and thus large values of the two target measures. This disadvantage becomes minor on the Griewank's function because the rough sample paths generated from the exponential kernel become appropriate for modeling the oscillations in the Griewank's function. Among the other three kernels, the Matérn kernel with $\nu = 5/2$ and the squared exponential kernel often have better performance because their sample paths are smoother.

Among the three sampling distributions, the uniform and truncated sampling have very similar performance. These two distributions are defined on the same supports, that is, $\mathcal{X} = [1,10]^d$ when $d = 1$ and $\mathcal{X} = [1,4]^d$ when $d = 3, 10$. The truncated normal is set with relatively large variances (7^2 when $d = 1$ and 3^2 when $d = 3,10$), which results in sufficiently spread out covariate points and hence similar performance to the uniform sampling. The performance of the normal sampling is a little different. This is because the normal sampling is defined on an infinite support, so the space that the MSE and PFS are integrated over is different. However, we can see that the normal sampling is effective in reducing the maximal IMSE and IPFS. The values of the two measures under normal sampling are basically on the same order as those under the uniform and truncated normal sampling.

5.2. M/M/1 Queue

The M/M/1 queue is analytical and thus provides convenience for estimating PFS. In this test, our example is taken from Zhou and Xie (2015). Customers arrive at a system according to a Poisson process with rate x, and the service time of the server follows an exponential distribution with mean $1/\lambda$. We consider two types of

Figure 3. (Color online) Three-Dimensional De Jong's Functions: Maximal IMSE and IPFS Under Different Covariance Kernels and Sampling Distributions



cost: the service cost $c_u\lambda$ with c_u being the per unit cost of the service rate, and the waiting cost, determined by the customers' mean waiting time $\mathrm{E}[\mathcal{W}(\lambda)]$ in the system. In addition, there is an upper bound $\mathcal{U}$ on the total cost. When the system is unstable (i.e., $x/\lambda \geq 1$), it will incur the cost $\mathcal{U}$. Therefore, the total cost TC of this system is

$$TC(x,\lambda) = \begin{cases} \min\{\mathrm{E}[\mathcal{W}(\lambda)] + c_u\lambda, \mathcal{U}\}, & \text{if } x/\lambda < 1; \\ \mathcal{U}, & \text{otherwise.} \end{cases}$$

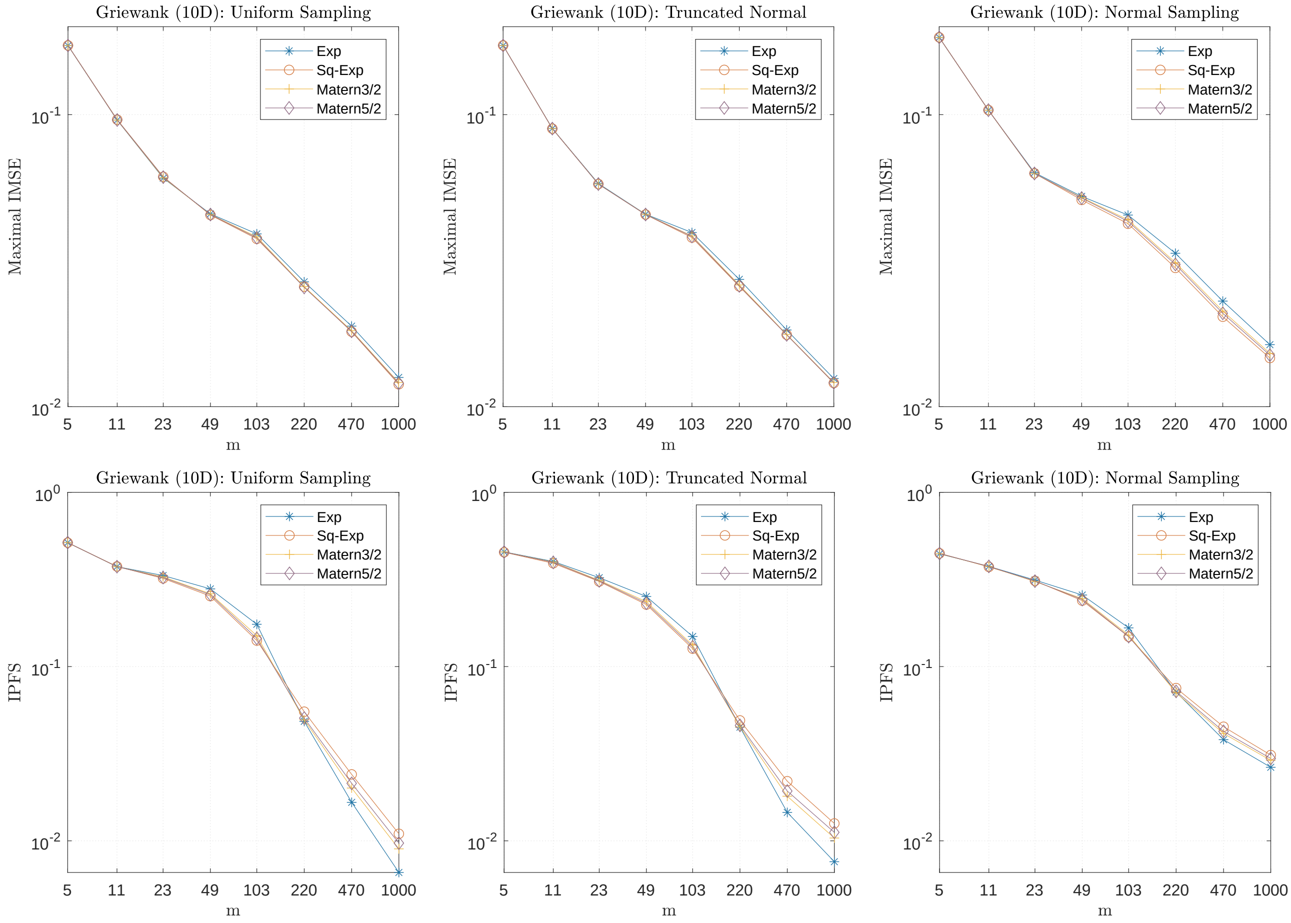
For the M/M/1 queue, the mean waiting time $\mathrm{E}[\mathcal{W}(\lambda)]$ has an analytical form $1/(\lambda - x)$, and the solution that minimizes the total cost is obtained at $\lambda^* = x + 1/\sqrt{c_u}$.

To fit into the framework of simulation with covariates, we consider 10 discrete designs with the ith design $\lambda_i = 6 + 0.3i$, $i = 1, 2, \ldots, 10$, and let $c_u = 0.1$ and $\mathcal{U} = 2.5$. The covariate x is restricted in an open interval $\mathcal{X} = (0.5, 4.5)$. We consider two sampling distributions $\mathbb{P}_{\mathbf{X}}$ for $\mathbf{X}^m$: uniform on $\mathcal{X}$ and truncated normal on $\mathcal{X}$ with mean 2.5 and variance 3^2. We let m take values in $\{5, 10, 20, 40, 80, 160, 320, 640\}$ and n take values in $\{5, 10\}$. The maximal IMSE and IPFS are estimated by the average of 100 macro Monte Carlo replications. The results for the maximal IMSE and the IPFS across the 10 designs are summarized in Figures 5 and 6.

Figure 5 shows that, on the logarithmic scale, the maximal IMSE across the 10 designs decreases almost linearly as the sample size $\log m$ increases, for all the four kernels and numbers of simulation replications tested. This observation agrees with our theory (Theorem 3) that the convergence rates of the maximal IMSE are in the polynomial orders of m for the three types of covariance kernels, including all the four kernels we have implemented here. This linear trend can be used to help an analyst make the design decision for achieving a target precision of the maximal IMSE. More details are available in Section 5.3 of the online supplement.

In Figure 5, increasing n from 5 to 10 does not significantly reduce the maximal IMSE for all kernels. Among the four kernels, the exponential kernel gives larger maximal IMSE than the other three, again due to the mismatch between its rough sample paths and

Figure 4. (Color online) Ten-Dimensional Griewank's Functions: Maximal IMSE and IPFS Under Different Covariance Kernels and Sampling Distributions



the smooth target function, because $TC(x, \lambda)$ is always a smooth function in x (infinitely differentiable) for all values of λ_i. Different sampling distributions on the covariate space do not seem to have a significant impact on the convergence pattern and rate.

Figure 6 shows the convergence of IPFS for $\delta_0 = 0.01$. It can be observed that the relative performance of the IPFS under different kernels, numbers of simulation replications, and sampling distributions basically remains the same as that of the maximal IMSE, but the

Figure 5. (Color online) Maximal IMSE Under Different Covariance Kernels, Sampling Distributions, and Values of n

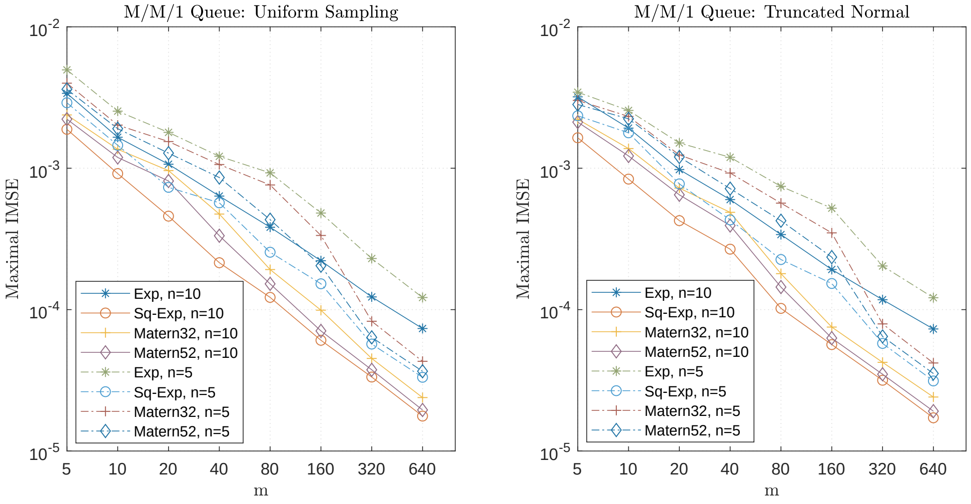
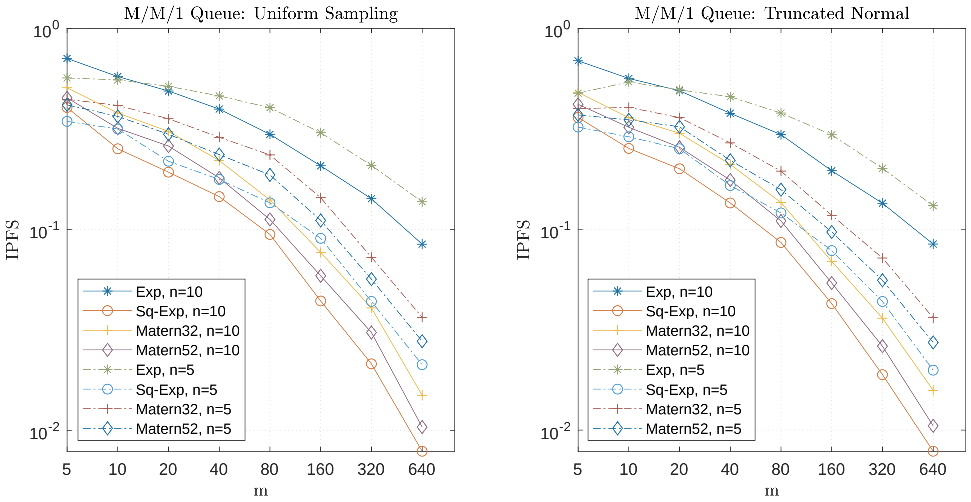


Figure 6. (Color online) IPFS Under Different Covariance Kernels, Sampling Distributions, and Values of n



convergence rates of the IPFS are faster, demonstrating a superlinear pattern on the logarithmic scale.

Remark 7. In this research, we used the SK models for system performance predictions. It is well known that the computational complexity of SK (or Gaussian process models) is $O(m^3)$, where m is the number of covariate points. Although with a fixed sampling distribution for the covariate points, we can collect all the covariate points in advance and build the SK models just once, this complexity only makes the computational time practically acceptable when m is no more than a few thousand, or tens of thousands, when the offline simulation period is long. When m becomes even larger than that, certain techniques in scalable Gaussian processes (Luo and Duraiswami 2013, Hensman et al. 2014, Wilson and Nickisch 2015) might be considered for improving the computational efficiency.

Remark 8. In this research, we adopted a fixed (static) distribution for sampling the covariate space. In the meantime, there has been an increasing interest recently in the development of adaptive design-of-experiment methods (Garud et al. 2017). As an initial investigation for the application potential of adaptive methods for the SK construction in simulation with covariates, we numerically compared our static sampling with an intuitive adaptive design procedure (adaptive MSE procedure). The results are provided in Section 5.2 of the online supplement. We observed that the static sampling considered in this research has similar empirical performance to the adaptive MSE procedure in general and tends to be superior when (i) the dimension of the covariate space is high; (ii) the covariate distribution deviates from uniform; and (iii) the target function has strong oscillation.

6. Conclusions and Discussion

Simulation with covariates is a recently proposed framework for conducting simulation experiments (Hong and Jiang 2019, Shen et al. 2021). It is comprised of the offline simulation and online prediction periods and is able to substantially reduce the decision time. We provide theoretical analysis for the predictive performance of the stochastic kriging model under this framework. We focus on two critical measures for the prediction errors, the maximal IMSE and IPFS, and study their convergence rates to understand the relationship between the offline simulation efforts and the online prediction accuracy.

For the maximal IMSE, we show that the convergence rates are $1/m$, $(\log m)^{\frac{d}{\kappa_*}}/m$, and $m^{-\frac{2\nu_*}{2\nu_*+d}}$ for the finite-rank kernels, exponentially decaying kernels, and polynomially decaying kernels, respectively, where m is the number of sampled covariate points, κ_* and ν_* are some kernel parameters, and d is the dimension of covariates. For the IPFS, we show that the convergence rates are at least as fast as the maximal IMSE and can be enhanced to exponential rates under some conditions.

Because the rates derived for the maximal IMSE and IPFS are simple and concrete and are the first to characterize the convergence rates of the prediction errors in simulation with covariates to the best of our knowledge, they serve as a good benchmark against which improvement in rates might be theoretically or numerically measured from future prediction methods built on possibly different assumptions, prediction models, covariance kernels, and covariate point collection strategies. In addition, the theoretical analysis in this research has the chance to be extended to facilitate new developments in simulation with covariates, for example, when adaptive design procedures are used to explore the covariate space.

Acknowledgments

The authors thank the department editor Bruno Tuffin, the associate editor, and referees for the constructive comments which substantially improved this paper.

Endnote

[1] If certain adaptive methods are used to collect the covariate points, the predictive models need to be built iteratively instead of once after all the covariate points are collected.

References

Ahmed MA, Alkhamis TM (2009) Simulation optimization for an emergency department healthcare unit in Kuwait. *Eur. J. Oper. Res.* 198:936–942.

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

Benini L, Hodgson R, Siegel P (1998) System-level power estimation and optimization. Chandrakasan A, Kiaei S, eds. *Proc. Internat. Sympos. on Low Power Electronics and Design* (ACM Press, New York), 173–178.

Bertsimas D, Kallus N, Weinstein AM, Zhuo YD (2017) Personalized diabetes management using electronic medical records. *Diabetes Care* 40(2):210–217.

Chen CH, Lee LH (2011) *Stochastic Simulation Optimization: An Optimal Computing Budget Allocation* (World Scientific Publishing, Singapore).

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

Chen CH, He D, Fu M, Lee LH (2008) Efficient simulation budget allocation for selecting an optimal subset. *INFORMS J. Comput.* 20(4):579–595.

Chen CH, Lin J, Yücesan E, Chick SE (2000) Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems* 10:251–270.

Dai L (1996) Convergence properties of ordinal comparison in the simulation of discrete event dynamic systems. *J. Optim. Theory Appl.* 91(2):363–388.

Ding H, Benyoucef L, Xie X (2005) A simulation optimization methodology for supplier selection problem. *Internat. J. Comput. Integrated Manufacturing* 18:210–224.

Frazier PI, Powell WB, Dayanik S (2008) A knowledge-gradient policy for sequential information collection. *SIAM J. Control Optim.* 47(5):2410–2439.

Gao S, Chen W (2017) Efficient feasibility determination with multiple performance measure constraints. *IEEE Trans. Automated Control* 62:113–122.

Gao S, Chen W, Shi L (2017) A new budget allocation framework for the expected opportunity cost. *Oper. Res.* 65:787–803.

Gao S, Du J, Chen CH (2019a) Selecting the optimal system design under covariates. Cappelleri D, Dimarogonas D, Dotoli M, Fanti MP, Lutz P, Seatzu C, Xie X, eds. *Proc. IEEE 15th Internat. Conf. on Automation Sci. and Engrg.* (IEEE, New York), 547–552.

Gao S, Li C, Du J (2019b) Rate analysis for offline simulation online application. Mustafee N, Bae K-HG, Lazarova-Molnar S, Rabe M, Szabo C, Haas P, Son Y-J, eds. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 3468–3479.

Garud SS, Karimi IA, Kraft M (2017) Design of computer experiments: A review. *Comput. Chemical Engrg.* 106:71–95.

Glynn P, Juneja S (2004) A large deviations perspective on ordinal optimization. Ingalls RG, Rossetti MD, Smith JS, Peters BA, eds. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 577–585.

Gu C (2002) *Smoothing Spline ANOVA Models* (Springer, New York).

Hensman J, Fusi N, Lawrence N (2014) Gaussian processes for big data. *Proc. 29th Conf. on Uncertainty in Artificial Intelligence* (AUAI Press, Corvallis, OR), 282–290.

Hong LJ, Jiang G (2019) Offline simulation online application: A new framework of simulation-based decision making. *Asia-Pacific J. Oper. Res.* 36(6):1940015.

Hsing T, Eubank R (2015) *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators* (John Wiley & Sons, Hoboken, NJ).

Kim SH, Nelson BL (2006) Selecting the best system. Henderson SG, Nelson BL, eds. *Simulation.* Handbooks in Operations Research and Management Science (Elsevier, Amsterdam), 501–534.

Kleijnen JPC (1993) Simulation and optimization in production planning: A case study. *Decision Support Systems* 9(3):269–280.

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

Law AM (2015) *Simulation Modeling and Analysis*, 5th ed. (McGraw-Hill, New York.)

Li C, Gao S, Du J (2022) Version v2021.0329. https://github.com/INFORMSJoC/2021.0329.

Luo Y, Duraiswami R (2013) Fast near-GRID Gaussian process regression. Carvalho CM, Ravikumar P, eds. *Proc. 16th Internat. Conf. on Artificial Intelligence and Statist.* (PMLR, Cambridge, MA), 424–432.

Ni EC, Ciocan DF, Henderson SG, Hunter SR (2017) Efficient ranking and selection in parallel computing environments. *Oper. Res.* 65(3):821–836.

Qu H, Fu MC (2014) Gradient extrapolated stochastic kriging. *ACM Trans. Modeling Comput. Simulations* 24(4):3.

Rasmussen CE, Williams CK (2006) *Gaussian Process for Machine Learning* (MIT Press, Cambridge, MA).

Ryzhov IO (2016) On the convergence rates of expected improvement methods. *Oper. Res.* 64(6):1515–1528.

Sabuncuoglu I, Touhami S (2002) Simulation metamodeling with neural networks: An experimental investigation. *Internat. J. Production Res.* 40:2483–2505.

Santin G, Schaback R (2016) Approximation of eigenfunctions in kernel-based spaces. *Adv. Comput. Math.* 42(4):973–993.

Shen H, Hong LJ, Zhang X (2021) Ranking and selection with covariates for personalized decision making. *INFORMS J. Comput.* 33(4):1500–1519.

Stein ML (1999) *Interpolation for Spatial Data: Some Theory for Kriging* (Springer, New York).

Steinwart I, Hush D, Scovel C (2009) Optimal rates for regularized least squares regression. Dasgupta S, Klivans A, eds. *Proc. 22nd Annual Conf. on Learn. Theory*, 79–93.

van der Vaart AW, van Zanten JH (2011) Information rates of nonparametric Gaussian process methods. *J. Machine Learn. Res.* 12:2095–2119.

Van Trees HL (2001) *Detection, Estimation, and Modulation Theory* (John Wiley & Sons, Hoboken, NJ).

Wang B, Hu J (2018) Some monotonicity results for stochastic kriging metamodels in sequential settings. *INFORMS J. Comput.* 30(2):278–294.

Wilson A, Nickisch H (2015) Kernel interpolation for scalable structured Gaussian processes (KISS-GP). Bach F, Blei D, eds. *Proc. Internat. Conf. on Machine Learn.* (PMLR, Cambridge, MA), 1775–1784.

Zhang T (2005) Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.* 17:2077–2098.

Zhou E, Xie W (2015) Simulation optimization when facing input uncertainty. Yilmaz L, Chan WKV, Moon I, Roeder TMK, Macal C, Rossetti MD, eds. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 3714–3724.