Convergence Analysis of Stochastic Kriging-Assisted Simulation with Random Covariates

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

[a] *Department of Statistics and Data Science, National University of Singapore, Singapore 117546, stalic@nus.edu.sg*

[b] *Department of Advanced Design and Systems Engineering and School of Data Science, City University of Hong Kong, Hong Kong, China, siyangao@cityu.edu.hk*

[c] *School of Management, Fudan University, China, jianzhodu2-c@my.cityu.edu.hk*

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 to 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.

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 of the real situations, systems are highly complex, precluding the possibility of applying analytical solutions; in contrast, simulation makes it possible to accurately describe a system through the use of logically complex, and often non-mathematical 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). Now simulation has been a widely-used operations-research and management-science technique, e.g., 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), etc.

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 which 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 dataset that is useful in describing the system. More importantly, this dataset 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 upon 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 dataset. 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

[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.

of covariate points ever collected and to facilitate further decision making.

We employ the stochastic kriging (SK) model as the predictive model. SK has is one of the most extensively studied models for simulation output, e.g., in 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, i.e., 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 dataset 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 stochastic kriging (SK). Chen et al. (2013) and Qu and Fu (2014) further enhanced SK by utilizing 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), etc. 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 in order 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 et al. (2017), Gao and Chen (2017)) and value of information procedure (VIP, Frazier et al. (2008), Ryzhov (2016)). In particular, Gao et al. (2019a) utilized 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 Stochastic Kriging

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 l-th 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 $\boldsymbol{\Sigma}_{M,i}(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}\left[M_i(\mathbf{x}), M_i(\mathbf{x}')\right]$ 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 $\mathcal{F}_i = (\mathbf{f}_i(\mathbf{x}_1), \ldots, \mathbf{f}_i(\mathbf{x}_m))^\top$. Let $\mathbf{\Sigma}_{M,i}(\mathbf{x}^m, \mathbf{x}^m)$ be the $m \times m$ covariance matrix across all covariate points $\mathbf{x}_1, \ldots, \mathbf{x}_m$, i.e., for $s, t \in \{1, \ldots, m\}$, the (s,t) entry of $\mathbf{\Sigma}_{M,i}(\mathbf{x}^m, \mathbf{x}^m)$ is $[\mathbf{\Sigma}_{M,i}(\mathbf{x}^m, \mathbf{x}^m)]_{st} = \mathrm{Cov}\left[y_i(\mathbf{x}_s), y_i(\mathbf{x}_t)\right]$. For any $\mathbf{x} \in \mathcal{X}$, let

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

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

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 $\overline{\mathbf{Y}}_i$, i.e., the simulation noise. We call the predictor that minimizes $\mathrm{MSE}_i(\mathbf{x}_0)$ MSE-optimal linear predictor. Stein (1999) (and also Ankenman et al. 2010, Chen et al. 2013) has shown that the MSE-optimal linear predictor has the form

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

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

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

$$\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) = \mathbf{\Sigma}_{M,i}(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{\Sigma}_{M,i}^\top(\mathbf{x}^m, \mathbf{x}_0) \left[\mathbf{\Sigma}_{M,i}(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_{\epsilon,i}(\mathbf{x}^m)\right]^{-1} \mathbf{\Sigma}_{M,i}(\mathbf{x}^m, \mathbf{x}_0)$$

$$+ \eta_i(\mathbf{x}_0)^\top \left[\mathcal{F}_i^\top \left(\mathbf{\Sigma}_{M,i}(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_{\epsilon,i}(\mathbf{x}^m)\right)^{-1} \mathcal{F}_i\right]^{-1} \eta_i(\mathbf{x}_0), \tag{3}$$

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

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 $[T_{\mathbf{\Sigma}_M} f](\mathbf{x}) = \int_{\mathcal{X}} \mathbf{\Sigma}_M(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') \mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x}')$ for any $\mathbf{x} \in \mathcal{X}$. Since $\mathbf{\Sigma}_M(\cdot,\cdot)$ is a continuous symmetric non-negative 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 $T_{\mathbf{\Sigma}_M}$, i.e., $\int_{\mathcal{X}} \phi_l^2(\mathbf{x}) d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) = 1$, $\int_{\mathcal{X}} \phi_l(\mathbf{x})\phi_{l'}(\mathbf{x}) d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) = 0$ for $l \neq l'$, and $[T_{\mathbf{\Sigma}_M}\phi_l](\mathbf{x}) = \mu_l \phi_l(\mathbf{x})$ for some eigenvalue $\mu_l \geq 0$, all $l = 1, 2, \ldots$ and $\mathbf{x} \in \mathcal{X}$. According to Mercer's theorem (e.g. Theorem 4.2 of Rasmussen and Williams 2006), the kernel $\mathbf{\Sigma}_M$ (which can be taken as any $\mathbf{\Sigma}_{M,i}$ for $i = 1, \ldots, k$) has the series expansion $\mathbf{\Sigma}_M(\mathbf{x}, \mathbf{x}') = \sum_{l=1}^{\infty} \mu_l \phi_l(\mathbf{x})\phi_l(\mathbf{x}')$ with respect to $\mathbb{P}_{\mathbf{X}}$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, where we assume that the eigenvalues of $\mathbf{\Sigma}_M$ are sorted into the decreasing order $\mu_1 \geq \mu_2 \geq \ldots \geq 0$. The trace of the kernel $\mathbf{\Sigma}_M$ is defined as $\mathrm{tr}(\mathbf{\Sigma}_M) = \sum_{l=1}^{\infty} \mu_l$. 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 reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ attached to the kernel $\mathbf{\Sigma}_M$ 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 below: 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 $\mu_1 \geq \ldots \geq \mu_{l_*} > 0$ and $\mu_{l_*+1} = \mu_{l_*+2} = \ldots = 0$ for some finite integer $l_* \in \mathbb{N}$. One example of finite-rank kernels is $\mathbf{\Sigma}_M(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^\top \mathbf{x}')^D$ 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 $\mathbf{\Sigma}_M(\mathbf{x}, \cdot)$ 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 $\mathbf{\Sigma}_M(\mathbf{x}, \mathbf{x}') = \exp\left\{-\varphi \|\mathbf{x} - \mathbf{x}'\|^2\right\}$ 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. So $\mu_l \asymp \exp(-cl^\kappa)$ 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 $\mathrm{tr}(\mathbf{\Sigma}_M) < \infty$). One example is the kernel $\mathbf{\Sigma}_M(\mathbf{x}, \mathbf{x}') = \min\{\mathbf{x}, \mathbf{x}'\}$ 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 $\mathbf{\Sigma}_{M,i}(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\varphi\|\mathbf{x} - \mathbf{x}'\|\right)^\nu K_\nu(\sqrt{2\nu}\varphi\|\mathbf{x} - \mathbf{x}'\|)$, where K_ν is the modifed 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 $\text{MSE}_{i,\text{opt}}(\mathbf{X}_0)$, 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 i-th design is the integral of $\text{MSE}_{i,\text{opt}}(\mathbf{X}_0)$ with respect to the sampling distribution of $\mathbf{X}_0$

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

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

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 $n_1 = \ldots = n_m = n$, i.e., 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 $y_i(\mathbf{x})$, $i = 1, \ldots, k$. Since 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 $\mathbf{f}_i(\cdot)$. Note that 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 $\widehat{i^\circ}(\mathbf{x}_0)$ at test point $\mathbf{X}_0 = \mathbf{x}_0$ are

$$y^\circ(\mathbf{x}_0) = \min_{i\in\{1,\ldots,k\}} y_i(\mathbf{x}_0), \qquad i^\circ(\mathbf{x}_0) \in \arg\min_{i\in\{1,\ldots,k\}} y_i(\mathbf{x}_0),$$

$$\widehat{y}^{\circ}(\mathbf{x}_0) = \min_{i\in\{1,\ldots,k\}} \widehat{y}_i(\mathbf{x}_0), \qquad \widehat{i}^{\circ}(\mathbf{x}_0) \in \arg\min_{i\in\{1,\ldots,k\}} \widehat{y}_i(\mathbf{x}_0). \tag{4}$$

Typically in R&S problems, the correct selection for the best design is defined as $\widehat{i}^{\circ}(\mathbf{x}_0) = i^{\circ}(\mathbf{x}_0)$. 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 $y_{\widehat{i}^{\circ}(\mathbf{x}_0)}(\mathbf{x}_0)$ 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 $y_{\widehat{i}^{\circ}(\mathbf{x}_0)}(\mathbf{x}_0)$ 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 $\mathrm{PCS}(\mathbf{x}_0)$ and false selection $\mathrm{PFS}(\mathbf{x}_0)$ among the k alternatives at $\mathbf{x}_0$ are given by

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

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 $\mathrm{PCS}(\mathbf{x}_0)$ and $\mathrm{PFS}(\mathbf{x}_0)$ on the constant IZ parameter δ_0.

Consequently, the integrated PFS is defined as

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

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.

We note two key differences bewteen our setting and existing research in the simulation litera-

ture. 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, same in appearance to the expected measure $\mathrm{PCS}_{\mathrm{E}}$ 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 intrinsical difference in 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 $\mathrm{PCS}_{\mathrm{E}}$ 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 above, 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, the maximal IMSE. We make the following assumptions:

A.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 $\mathbf{x} \in \mathcal{X}$.

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

A.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 $l = 1, 2, \ldots, \infty$.

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

A.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. A.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.

A.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. A.3 imposes a mild moment condition on the orthonormal basis functions. Sometimes A.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 A.3 trivially holds. The quantities ρ_* and r_* do not need to depend on i, because if the ith design satisfies $\mathrm{E}_{\mathbf{X}}\{\phi_{i,l}^{2r_i}(\mathbf{X})\} \leq \rho_i^{2r_i}$ for $r_i \geq 2$, one can let $r_* = \min_{i\in\{1,\ldots,k\}} r_i \geq 2$ and $\rho_* = \max\left(\max_{i\in\{1,\ldots,k\}} \rho_i, 1\right)$. By Jensen's inequality, $\mathrm{E}_{\mathbf{X}}\{\phi_{i,l}^{2r_*}(\mathbf{X})\} \leq \left[\mathrm{E}_{\mathbf{X}}\{\phi_{i,l}^{2r_i}(\mathbf{X})\}\right]^{r_*/r_i} \leq \rho_i^{2r_i \cdot r_*/r_i} \leq \rho_*^{2r_*}$ and A.3 holds.

A.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 $\boldsymbol{\beta}_i$, since 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. Since 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:

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

where $\eta(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0) - \mathcal{F}^{\top} \left(\boldsymbol{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \boldsymbol{\Sigma}_\epsilon\right)^{-1} \boldsymbol{\Sigma}_M(\mathbf{x}^m, \mathbf{x}_0)$. 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 A.1-A.3, the following relation holds*

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

where

$$\begin{aligned}
b(m, \zeta, r_*) &= \max\left(\sqrt{\max(r_*, \log\zeta)}, \ \frac{\max(r_*, \log\zeta)}{m^{1/2 - 1/r_*}}\right), \\
\gamma(a) &= \sum_{l=1}^{\infty} \frac{\mu_l}{\mu_l + a} \ \textit{for any } a > 0, \quad \mathrm{tr}\left(\boldsymbol{\Sigma}_M^{(\zeta)}\right) = \sum_{l=\zeta+1}^{\infty} \mu_l \ \textit{for any } \zeta \in \mathbb{N}.
\end{aligned}$$

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 of all, the first term in the upper bound (7) is the

dominant term, while 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 $\gamma(\frac{\overline{\sigma}_0^2}{mn})$ 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 $\frac{\overline{\sigma}_0^2}{mn}\gamma(\frac{\overline{\sigma}_0^2}{mn})$ 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 A.1-A.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 A.1-A.4, the following relation holds*

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

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 to 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 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 A.1-A.4, the following results hold with* r_* *given in Assumption A.3:*

(i) (Finite-rank kernels) If for every $i = 1,\ldots,k$, $\mathbf{\Sigma}_{M,i}$ *is a finite-rank kernel of rank* l_{*i}, *i.e., its eigenvalues satisfy* $\mu_{i,1} \ge \mu_{i,2} \ge \ldots \ge \mu_{i,l_{*i}} > 0$ *and* $\mu_{i,l_{*i}+1} = \mu_{i,l_{*i}+2} = \ldots = 0$, *then as*

$m \to \infty$,

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

(ii) (Exponentially decaying kernels) If for every $i = 1,\ldots,k$, $\mathbf{\Sigma}_{M,i}$ *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* $\kappa_* = \min_{i\in\{1,\ldots,k\}} \kappa_i$. *Then, as* $m \to \infty$,

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

(iii) (Polynomially decaying kernels) If for every $i = 1,\ldots,k$, $\mathbf{\Sigma}_{M,i}$ *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* $\nu_* = \min_{i\in\{1,\ldots,k\}} \nu_i$. *Then, as* $m \to \infty$,

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

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, since 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 A.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 $\mathrm{E}_{\mathbf{X}^m}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}^{(M)}_{\mathrm{opt}}(\mathbf{X}_0)\right]$ and $\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}^{(\boldsymbol{\beta})}_{\mathrm{opt}}(\mathbf{X}_0)\right]$ in Theorems 1 and 2. These rates are generally tight and cannot be improved. In Remark 2 below, 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* $l = 1, 2, ..., n$*,* $i = 1, \ldots, k$*, 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* $j = 1, \ldots, m$ *and* $i = 1, \ldots, k$*. Let* $\mathcal{X} \subseteq \mathbb{R}^d$*, and let the* i*th covariance kernel be* $\mathbf{\Sigma}_{M,i}(\mathbf{x}, \mathbf{x}') = a_i(\mathbf{x}^\top \mathbf{x}' + b_i)$ *for some known constants* $a_i > 0$ *and* $b_i > 0$*,* $i = 1, \ldots, k$*. 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* $\mathbf{\Sigma}_{M,i}(\mathbf{x}, \mathbf{x}') = a_i\big(\mathbf{x}^\top \mathbf{x}' + b_i\big)$ *for* $\mathbf{x}, \mathbf{x}' \in \mathcal{X} \subseteq \mathbb{R}^d$*, known constants* $a_i > 0, b_i > 0$ *and* $i = 1, \ldots, k$*. Under Assumptions A.1-A.4 and the model setup described above, the MSE-optimal linear predictor in* (2) *and the optimal MSE in* (3) *are given by*

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

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*

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

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* $m \to \infty$*,*

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

Theorem 4 shows that the maximal IMSE of the covariance kernel $\mathbf{\Sigma}_{i,M}(\mathbf{x}, \mathbf{x}') = a_i\big(\mathbf{x}^\top \mathbf{x}' + b_i\big)$

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 $\text{PFS}(\mathbf{X}_0)$ converges to 0 (or equivalently, how fast $\text{PCS}(\mathbf{X}_0)$ converges to 1). Similar to the analysis of the maximal IMSE before, the convergence rate here is again in the average sense, by taking expectations of $\text{PFS}(\mathbf{X}_0)$ 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.

A.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 $\mathbf{x} \in \mathcal{X}$.

A.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 $t > 0$,

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

A.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 $m \geq m_0$,

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

Although A.5 is stronger than A.1 by assuming normal observation noises, it is a common assumption in simulation-based optimization problems. We emphasize that the normality assumption

in A.5 is only needed for deriving tighter and exponentially small bounds for IPFS in Theorem 5 below. Without A.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 A.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. A.7 is an alternative condition stronger than A.6, requiring that the supremum of MSE over $\mathcal{X}$ to be bounded with a high probability. Both A.6 and A.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. A.5 together with either A.6 or A.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 A.1-A.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 A.1-A.6 hold, then as $m \to \infty$,

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

where w_2 is given in Assumption A.6;

(iii) If Assumptions A.1-A.5 and A.7 hold, then as $m \to \infty$,

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

where w_3 is given in Assumption A.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 A.1-A.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 A.5 and Assumption A.6 provide sharpened convergence rates of IPFS than in 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 A.5 and A.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. Note that 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 A.1 that the simulation samples are independent across different designs* i*. This assumption is naturally the case when the designs are categorical, e.g., 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, i.e., 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_i*'s are no longer identical to each other. It falls in the well-established problem of ranking and selection (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* $n_c = mn$*, while the second term depends on* m *and* n *separately. In order 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 in order 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_i's 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_i's, 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, since 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.

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

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

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

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

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

Similar to Ankenman et al. (2010), the covariance matrices $\mathbf{\Sigma}_{\epsilon,i}(\mathbf{x}^m)$'s are estimated by diag $\left\{\widetilde{\sigma}_i^2(\mathbf{x}_1)/n_1, \ldots, \widetilde{\sigma}_i^2(\mathbf{x}_m)/n_m\right\}$, where $\widetilde{\sigma}_i^2(\mathbf{x}_j)$ $(j = 1, \ldots, m)$ 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 $\mathbf{\Sigma}_{\epsilon,i}(\mathbf{x}^m)$'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 experiments below, we compute the estimated MSE at a single point $\mathbf{x}_0$ by the formula $\widehat{\mathrm{MSE}}(\mathbf{x}_0) = [\widehat{y}(\mathbf{x}_0) - y(\mathbf{x}_0)]^2$, where $y(\mathbf{x}_0)$ is the true function value at $\mathbf{x}_0$ and $\widehat{y}(\mathbf{x}_0)$ 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 $\widehat{y}^\circ(\mathbf{x}_0)$ 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 $x_l = z_l$, $l = 1, 2, ..., d$ with $M(\mathbf{x}^*) = 0$.

We consider 10 discrete designs with the i-th design $\mathbf{z}^i = (\underbrace{i, ..., i}_{d})$, $i = 1, 2, ..., 10$.

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 $x_l = z_l$, $l = 1, 2, ..., d$ with $M(\mathbf{x}^*) = 0$. We consider 10 discrete designs with the i-th design $\mathbf{z}^i = (\underbrace{i, ..., i}_{d})$, $i = 1, 2, ..., 10$.

Note that 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 $m = 1000$ 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.

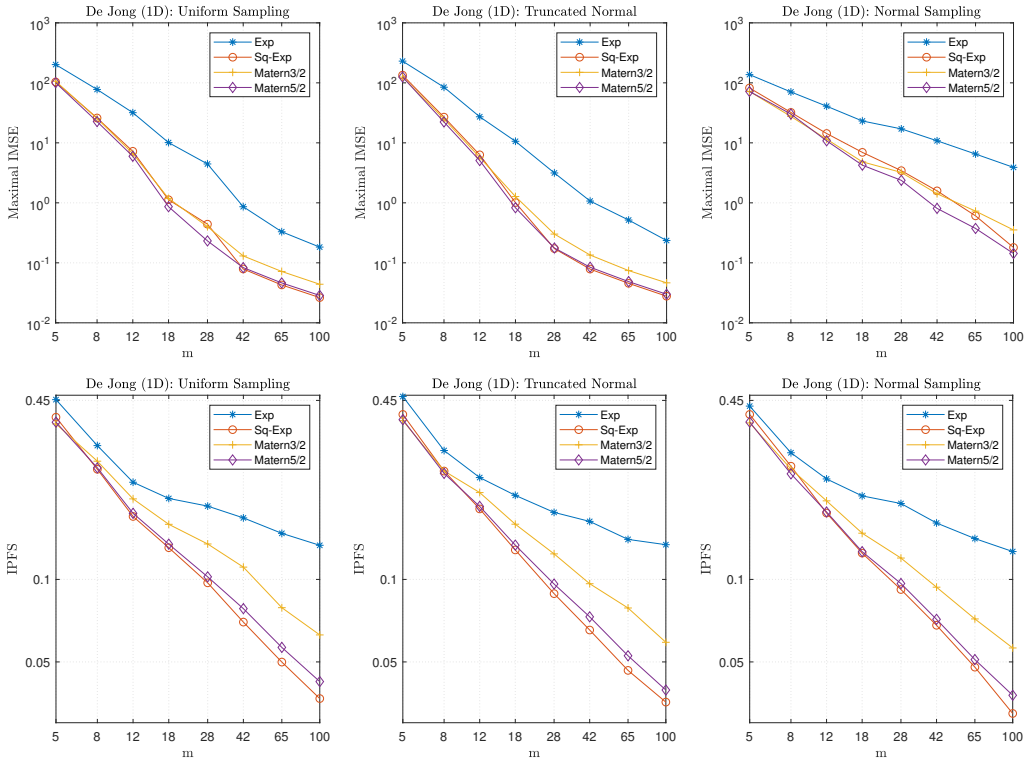

Figure 1: 1-d De Jong's functions: maximal IMSE and IPFS under different covariance kernels and sampling distributions.

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 $d = 10$. Since 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

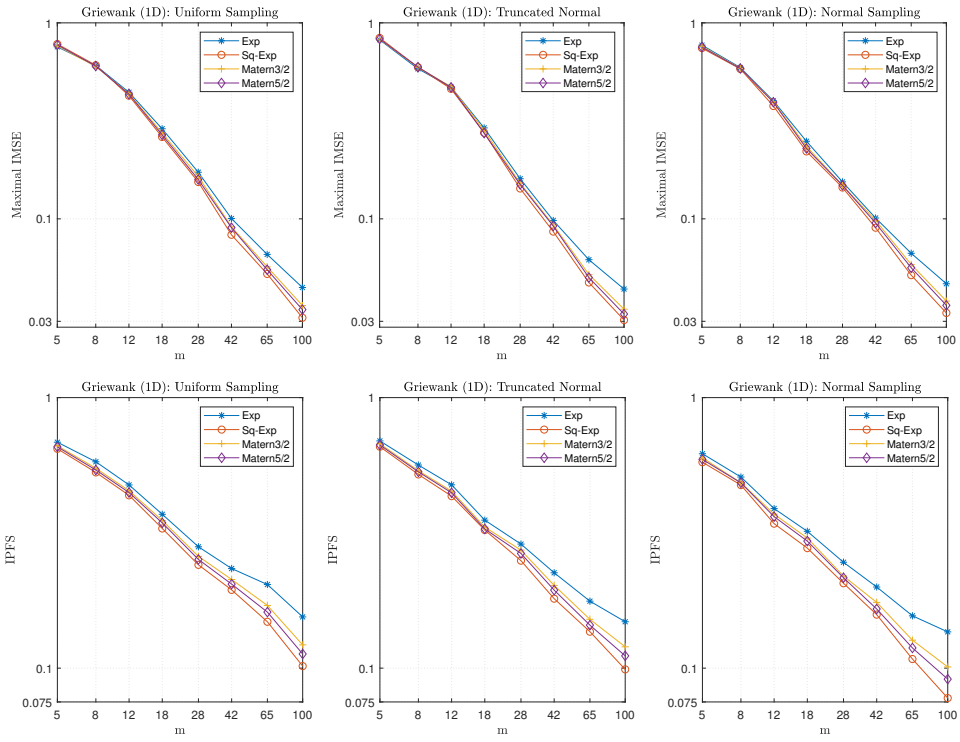

Figure 2: 1-d Griewank's functions: maximal IMSE and IPFS under different covariance kernels and sampling distributions.

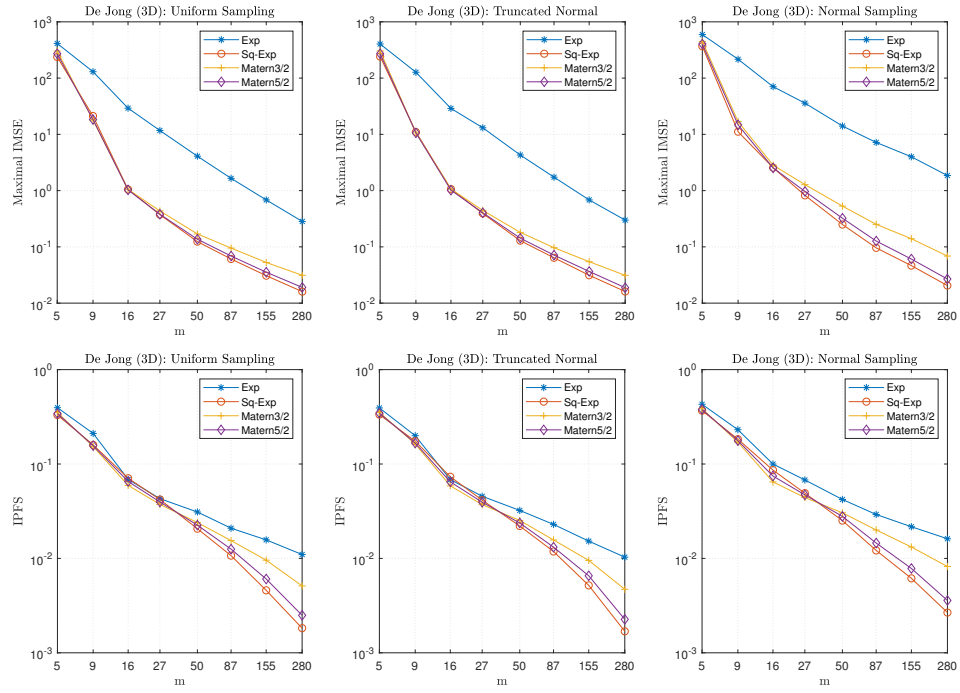

Figure 3: 3-d De Jong's functions: maximal IMSE and IPFS under different covariance kernels and sampling distributions.

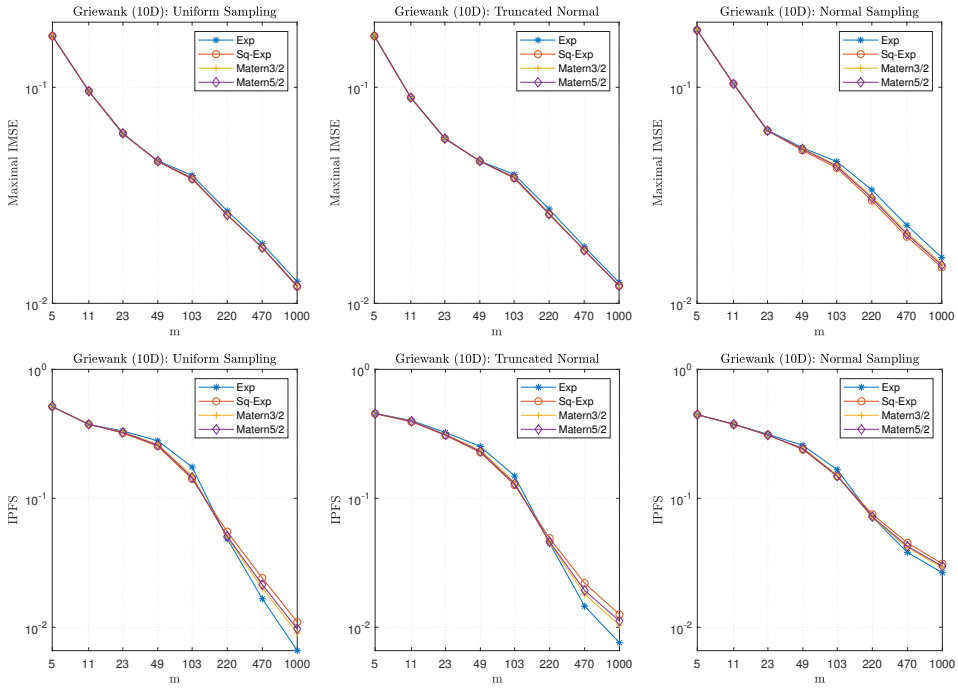

Figure 4: 10-d Griewank's functions: maximal IMSE and IPFS under different covariance kernels and sampling distributions.

faster than polynomial. These observations agree with the rate results in Theorem 5.

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) while 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, i.e., $\mathcal{X} = [1, 10]^d$ 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 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

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

Note that 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 i-th design $\lambda_i = 6 + 0.3i$, $i = 1, 2, ..., 10$, 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

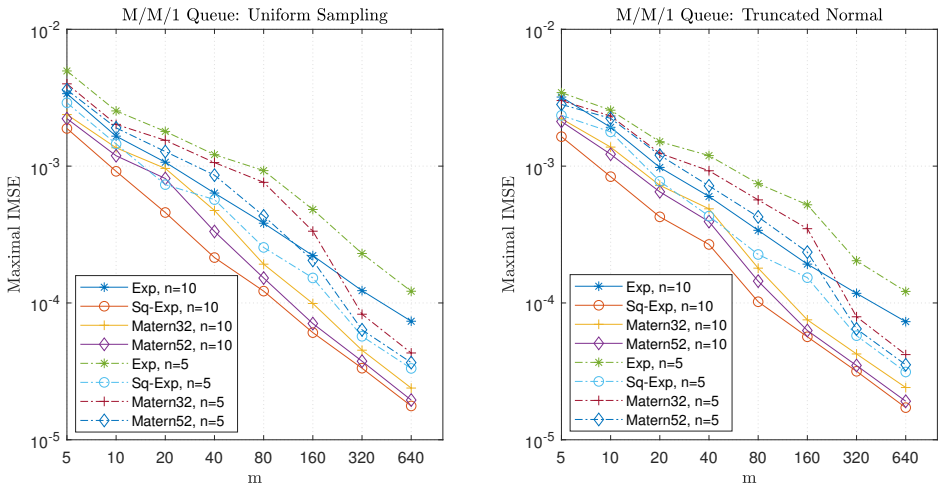

Figure 5: Maximal IMSE under different covariance kernels, sampling distributions and values of n.

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. We note that this linear trend can be utilized 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 the smooth target function, since $TC(x, \lambda)$ 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 convergence rates of the IPFS are faster, demonstrating a superlinear pattern on the logarithmic scale.

REMARK 7. *In this research, we have employed 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*

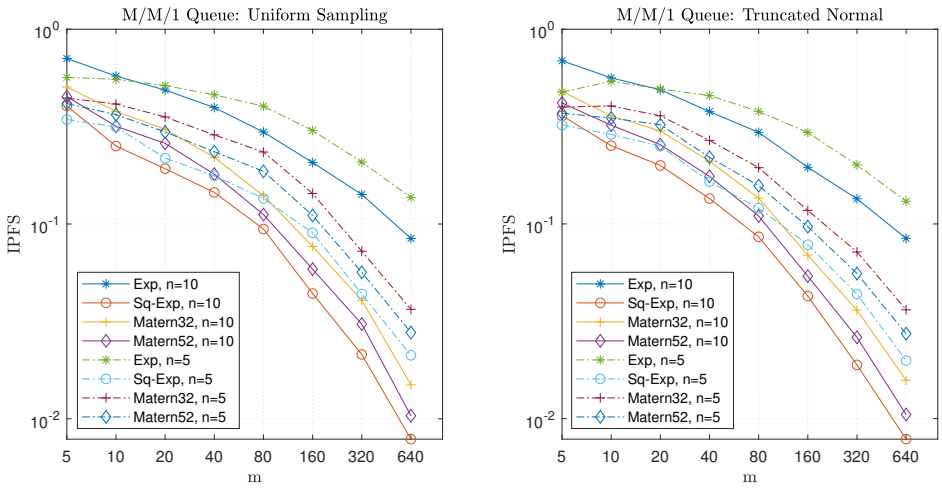

Figure 6: IPFS under different covariance kernels, sampling distributions and values of n.

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 thousand 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 have 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, in order 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.

Since 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, e.g., when adaptive design procedures are used to explore the covariate space.

A Notation

We summarize the key notation used in this paper in the following table.

Table 1: Table of notation.

Symbol	i folded[2]	Meaning
$\|\cdot\|$		Euclidean norm of a vector
$\|\|\|\cdot\|\|\|$		operator norm of a matrix, defined as $\sup_{\|\mathbf{v}\|=1} \|\cdot\mathbf{v}\|$
$\|\cdot\|_2^2$		L_2 norm of a function
$a_l \lesssim b_l$		mean that $\limsup_{l\to\infty} a_l/b_l < \infty$
$a_l \asymp b_l$		mean that $a_l \lesssim b_l$ and $b_l \lesssim a_l$
k		number of system designs
d		dimension of the covariate space
m		number of covariate points
n		number of replications for each pair of covariate point and design
q		dimension of the regressors $\mathbf{f}$ and the regression coefficient $\boldsymbol{\beta}$
$\mathcal{X}$		support of covariate points
$y_i(\cdot)$	$y(\cdot)$	mean of design i
$Y_{il}(\cdot)$		the l-th simulation sample from design i
$\epsilon_{il}(\cdot)$		simulation noise of the l-th sample of design i
$\bar{\epsilon}_i(\cdot)$		averaged simulation errors, defined as $n^{-1}\sum_{l=1}^n \epsilon_{il}(\cdot)$
$\sigma_i^2(\cdot)$		variance of $\epsilon_{il}(\cdot)$
$\overline{Y}_i(\cdot)$	$\overline{Y}(\cdot)$	sample mean of design i
$\overline{\mathbf{Y}}_i$	$\overline{\mathbf{Y}}$	vector of samples means at m covariate points
$\mathbf{x}, \mathbf{X}$		vector of covariates with support $\mathcal{X} \subseteq \mathbb{R}^d$
$\mathbf{x}_0, \mathbf{X}_0$		test covariate point for the SK model
$\mathbf{x}_j, \mathbf{X}_j$		the j-th covariate point
$\mathbf{x}^m, \mathbf{X}^m$		vector of m covariate points
$\mathbf{f}_i(\cdot)$	$\mathbf{f}(\cdot)$	vector of known basis functions
$\boldsymbol{\beta}_i$	$\boldsymbol{\beta}$	vector of unknown parameters for $\mathbf{f}_i(\cdot)$

[2]For simplicity of notation, in this research, we have folded the design index i in circumstances with no ambiguity. This column shows the symbol if its subscript i has been folded in the paper.

$M_i(\cdot)$	$M(\cdot)$	realization of a mean zero stationary Gaussian process for design i
$\mathbf{\Sigma}_{M,i}(\mathbf{x},\mathbf{x}')$	$\mathbf{\Sigma}_M(\mathbf{x},\mathbf{x}')$	covariance function, defined as $\mathrm{Cov}\left[M_i(\mathbf{x}), M_i(\mathbf{x}')\right]$
$\mathbf{\Sigma}_{\epsilon,i}(\mathbf{x}^m)$	$\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)$	covariance matrix of the averaged simulation errors in design i
$\widehat{y}_i(\cdot)$	$\widehat{y}(\cdot)$	MSE-optimal linear predictor of the i-th SK model
$\mathrm{MSE}_{i,\mathrm{opt}}(\cdot)$	$\mathrm{MSE}_{\mathrm{opt}}(\cdot)$	the MSE of predictor $\widehat{y}_i(\cdot)$
$\lambda_{\max}(\cdot), \lambda_{\min}(\cdot)$		the largest and smallest eigenvalues of a matrix
$A_1 \prec A_2, A_2 \succ A_1$		mean that $A_2 - A_1$ is positive definite
$A_1 \preceq A_2, A_2 \succeq A_1$		mean that $A_2 - A_1$ is positive semi-definite
$\mathbb{1}(\cdot)$		indicator function
$\mathbb{P}_{\mathbf{X}}, \mathrm{E}_{\mathbf{X}}$		a probability distribution/expectation over $\mathcal{X}$
$L_2(\mathbb{P}_{\mathbf{X}})$		L_2 space under $\mathbb{P}_{\mathbf{X}}$
$\langle\cdot,\cdot\rangle_{L_2(\mathbb{P}_{\mathbf{X}})}$		inner product in $L_2(\mathbb{P}_{\mathbf{X}})$, defined as $\mathrm{E}_{\mathbf{X}}(\cdot)$
$[T_{\mathbf{\Sigma}_M} f](\mathbf{x})$		a linear operator of $\mathbf{x} \in \mathcal{X}$, defined as $\int_{\mathcal{X}} \mathbf{\Sigma}_M(\mathbf{x},\mathbf{x}') f(\mathbf{x}') \mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x}')$
$\left\{\phi_{i,l}(\mathbf{x}) : l = 1, \dots\right\}$	$\phi_l(\mathbf{x})$	the orthonormal basis for $\mathbf{\Sigma}_{M,i}$ (from Mercer's theorem)
$\mathrm{tr}(\cdot)$		trace of a kernel (matrix)
$\{\mu_{i,l} : l = 1, \dots\}$	μ_l	eigenvalues of $\mathbf{\Sigma}_{M,i}$
$\mathbb{H}_i$	$\mathbb{H}$	reproducing kernel Hilbert space attached to $\mathbf{\Sigma}_{M,i}$
$\langle\cdot,\cdot\rangle_{\mathbb{H}}$		$\mathbb{H}$-inner product
ρ_*, r_*		parameters made for $\phi_{i,l}(\mathbf{x})$ in Assumption A.3
$\kappa_i, \nu_i, \tau_i, \varphi_i$	$\kappa, \nu, \tau, \varphi$	kernel parameters of the i-th SK model
κ_*, ν_*		parameters in rate functions, defined as $\kappa_* = \min_{i\in\{1,2,\dots,k\}} \kappa_i$ and $\nu_* = \min_{i\in\{1,2,\dots,k\}} \nu_i$
$\lesssim_{\mathbb{P}_{\mathbf{X}^m}}$		mean bounding in $\mathbb{P}_{\mathbf{X}^m}-$ probability
δ_0		indifference-zone parameter
$\underline{\sigma}_0^2, \overline{\sigma}_0^2$		lower and upper bounds for $\sigma_i^2(\mathbf{x})$ for all i and all $\mathbf{x} \in \mathcal{X}$
$i^\circ(\cdot), \widehat{i^\circ}(\cdot)$		the real and estimated optimal designs
$R^F(m,n)$		rate function of the maximal IMSE for finite-rank kernels

$R_i^E(m,n)$	$R^E(m,n)$	rate function of the maximal IMSE for exponentially decaying kernels and design i
$R_i^P(m,n)$	$R^P(m,n)$	rate function of the maximal IMSE for polynomially decaying kernels and design i
$R(m,n)$		rate function of IMSE
$\mu_i^{\mathcal{X}}$		expected mean performance of design i over the covariate space
$q_{i,\alpha}^{\mathcal{X}}$		α-quantile of the performance of design i over the covariate space
$w_i^{\mathcal{X}}$		proportion of design i being the best over the covariate space

B Technical Proofs and Additional Theoretical Results

In this section, we first prove Theorems 1 to 5 in the main text. Next, we present a new theorem (Theorem 6) about the restrictiveness of Assumptions A.6 and A.7.

We reinstate some useful notation and relations. 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 $c \in \mathbb{R}$, and $|||A||| = \sup_{\|\mathbf{v}\|=1} \|A\mathbf{v}\|$ to denote its matrix operator norm. For any positive definite matrix A, let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ be its largest and smallest eigenvalues. For two positive definite matrices A_1, A_2, $A_1 \prec A_2$ and $A_2 \succ A_1$ mean that $A_2 - A_1$ is positive definite; $A_1 \preceq A_2$ and $A_2 \succeq A_1$ mean that $A_2 - A_1$ is positive semi-definite. 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. Let $\mathbb{1}(\cdot)$ be the indicator function and $\mathbf{I}_k$ be the $k \times k$ identity matrix.

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 L_2 norm of f is given by $\|f\|_2^2 = \sum_{l=1}^{\infty} \theta_l^2$. The reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ attached to $\mathbf{\Sigma}_M$ is the space of all functions $f \in L_2(\mathbb{P}_{\mathbf{X}})$ such that its $\mathbb{H}$-norm $\|f\|_{\mathbb{H}}^2 = \sum_{l=1}^{\infty} \theta_l^2/\mu_l < \infty$. For any two generic functions $\mathrm{h}_1, \mathrm{h}_2 \in \mathbb{H}$, let their $L_2(\mathbb{P}_{\mathbf{X}})$ expansions be $\mathrm{h}_s(\mathbf{x}) = \sum_{l=1}^{\infty} h_{sl}\phi_l(\mathbf{x})$ for $s = 1, 2$. Their $\mathbb{H}$-inner product is given by $\langle \mathrm{h}_1, \mathrm{h}_2 \rangle_{\mathbb{H}} = \sum_{l=1}^{\infty} h_{1l}h_{2l}/\mu_l$. For any $\mathrm{h} \in \mathbb{H}$, the

reproducing property of $\mathbb{H}$ says that for any $\mathbf{x} \in \mathcal{X}$, $\langle \mathbf{\Sigma}_M(\mathbf{x}, \cdot), \mathrm{h}(\cdot) \rangle_{\mathbb{H}} = \mathrm{h}(\mathbf{x})$.

Proof of Theorem 1:

According to Mercer's theorem (e.g. Theorem 4.2 of Rasmussen and Williams 2006), the series expansion of the kernel function $\mathbf{\Sigma}_M(\mathbf{x}, \mathbf{x}') = \sum_{l=1}^{\infty} \mu_l \phi_l(\mathbf{x}) \phi_l(\mathbf{x}')$ holds almost surely for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, and hence

$$\begin{aligned}
&\mathbf{\Sigma}_M(\mathbf{x}_0, \mathbf{x}_0) = \sum_{a=1}^{\infty} \mu_a \phi_a^2(\mathbf{x}_0), \quad \mathbf{\Sigma}_M(\mathbf{x}_j, \mathbf{x}_0) = \sum_{a=1}^{\infty} \mu_a \phi_a(\mathbf{x}_j) \phi_a(\mathbf{x}_0), \text{ for } j = 1, \ldots, m, \\
&\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}_0) = \left[\mathbf{\Sigma}_M(\mathbf{x}_1, \mathbf{x}_0), \ldots, \mathbf{\Sigma}_M(\mathbf{x}_m, \mathbf{x}_0) \right]^\top .
\end{aligned} \tag{18}$$

Under the orthonormal property, if $\mathbf{X} \sim \mathbb{P}_{\mathbf{X}}$, then $\mathrm{E}_{\mathbf{X}}[\phi_a^2(\mathbf{X})] = 1$ and $\mathrm{E}_{\mathbf{X}}[\phi_a(\mathbf{X}) \phi_b(\mathbf{X})] = 0$ for $a \neq b$. Therefore,

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}^m} \mathrm{E}_{\mathbf{X}_0} \left[\mathrm{MSE}_{\mathrm{opt}}^{(M)}(\mathbf{X}_0) \right] \\
&= \mathrm{E}_{\mathbf{X}_0} \left[\mathbf{\Sigma}_M(\mathbf{X}_0, \mathbf{X}_0) \right] - \mathrm{E}_{\mathbf{X}^m} \mathrm{E}_{\mathbf{X}_0} \left\{ \mathbf{\Sigma}_M^\top(\mathbf{X}^m, \mathbf{X}_0) \left[\mathbf{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{X}^m) \right]^{-1} \mathbf{\Sigma}_M(\mathbf{X}^m, \mathbf{X}_0) \right\} \\
&\overset{(i)}{=} \sum_{a=1}^{\infty} \mu_a \mathrm{E}_{\mathbf{X}_0} \left[\phi_a^2(\mathbf{X}_0) \right] - \\
&\mathrm{E}_{\mathbf{X}^m} \mathrm{E}_{\mathbf{X}_0} \sum_{j=1}^{m} \sum_{j'=1}^{m} \sum_{a=1}^{\infty} \sum_{b=1}^{\infty} \mu_a \mu_b \left\{ \left[\mathbf{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{X}^m) \right]^{-1} \right\}_{jj'} \phi_a(\mathbf{X}_j) \phi_a(\mathbf{X}_0) \phi_b(\mathbf{X}_j') \phi_b(\mathbf{X}_0) \\
&\overset{(ii)}{=} \sum_{a=1}^{\infty} \mu_a \mathrm{E}_{\mathbf{X}_0} \left[\phi_a^2(\mathbf{X}_0) \right] - \mathrm{E}_{\mathbf{X}^m} \sum_{j=1}^{m} \sum_{j'=1}^{m} \sum_{a=1}^{\infty} \sum_{b=1}^{\infty} \mu_a \mu_b \left\{ \left[\mathbf{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{X}^m) \right]^{-1} \right\}_{jj'} \\
&\quad \cdot \phi_a(\mathbf{X}_j) \phi_b(\mathbf{X}_{j'}) \mathrm{E}_{\mathbf{X}_0} \left[\phi_a(\mathbf{X}_0) \phi_b(\mathbf{X}_0) \right] \\
&= \sum_{a=1}^{\infty} \mu_a - \mathrm{E}_{\mathbf{X}^m} \sum_{j=1}^{m} \sum_{j'=1}^{m} \sum_{a=1}^{\infty} \mu_a^2 \left\{ \left[\mathbf{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{X}^m) \right]^{-1} \right\}_{jj'} \phi_a(\mathbf{X}_j) \phi_b(\mathbf{X}_{j'}) \\
&= \sum_{a=1}^{\zeta} \mu_a - \mathrm{E}_{\mathbf{X}^m} \sum_{a=1}^{\zeta} \sum_{j=1}^{m} \sum_{j'=1}^{m} \mu_a^2 \left\{ \left[\mathbf{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{X}^m) \right]^{-1} \right\}_{jj'} \phi_a(\mathbf{X}_j) \phi_a(\mathbf{X}_{j'}) \\
&\quad + \mathrm{tr} \left(\mathbf{\Sigma}_M^{(\zeta)} \right) - \mathrm{E}_{\mathbf{X}^m} \sum_{a=\zeta+1}^{\infty} \sum_{j=1}^{m} \sum_{j'=1}^{m} \mu_a^2 \left\{ \left[\mathbf{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{X}^m) \right]^{-1} \right\}_{jj'} \phi_a(\mathbf{X}_j) \phi_a(\mathbf{X}_{j'}) \\
&\overset{(iii)}{\leq} \sum_{a=1}^{\zeta} \left\{ \mu_a - \mathrm{E}_{\mathbf{X}^m} \sum_{j=1}^{m} \sum_{j'=1}^{m} \mu_a^2 \left\{ \left[\mathbf{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{X}^m) \right]^{-1} \right\}_{jj'} \phi_a(\mathbf{X}_j) \phi_a(\mathbf{X}_{j'}) \right\} + \mathrm{tr} \left(\mathbf{\Sigma}_M^{(\zeta)} \right).
\end{aligned} \tag{19}$$

In the derivation above, we exchange the expectation and the summation in several steps.

- For Step (i), because $\left\{\sum_{a=1}^{N}\mu_a\phi_a^2(\mathbf{X}_0),\quad N=1,2,\ldots\right\}$ is a non-decreasing sequence of functions, by the monotone convergence theorem, we have $\mathrm{E}_{\mathbf{X}_0}\left[\mathbf{\Sigma}_M(\mathbf{X}_0,\mathbf{X}_0)\right]=\mathrm{E}_{\mathbf{X}_0}\left[\sum_{a=1}^{\infty}\mu_a\phi_a^2(\mathbf{X}_0)\right]=\sum_{a=1}^{\infty}\mu_a\mathrm{E}_{\mathbf{X}_0}\left[\phi_a^2(\mathbf{X}_0)\right]$.

- For Step (ii), for any $\mathbf{x}^m$, every $j,j'=1,\ldots,m$, and $N_1,N_2=1,2,\ldots$,

$$\begin{aligned}
&\left|\sum_{a=1}^{N_1}\sum_{b=1}^{N_2}\mu_a\mu_b\left\{\left[\mathbf{\Sigma}_M(\mathbf{x}^m,\mathbf{x}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_{j'})\phi_a(\mathbf{x}_0)\phi_b(\mathbf{x}_0)\right| \\
\le&\sum_{a=1}^{N_1}\sum_{b=1}^{N_2}\mu_a\mu_b\left\{\left[\mathbf{\Sigma}_M(\mathbf{x}^m,\mathbf{x}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_{j'})\phi_a(\mathbf{x}_0)\phi_b(\mathbf{x}_0) \\
&\cdot\operatorname{sgn}\left(\left\{\left[\mathbf{\Sigma}_M(\mathbf{x}^m,\mathbf{x}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_{j'})\phi_a(\mathbf{x}_0)\phi_b(\mathbf{x}_0)\right), \qquad (20)
\end{aligned}$$

  where $\operatorname{sgn}(x)=1$ for $x>0$, $\operatorname{sgn}(x)=-1$ for $x<0$, and $\operatorname{sgn}(x)=0$ if $x=0$. By Assumption A.3 and Hölder's inequality,

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}_0}\left\{\phi_a(\mathbf{X}_0)\phi_b(\mathbf{X}_0)\operatorname{sgn}\left(\left\{\left[\mathbf{\Sigma}_M(\mathbf{x}^m,\mathbf{x}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\phi_a(\mathbf{x}_j)\phi_a(\mathbf{X}_0)\phi_b(\mathbf{x}_{j'})\phi_b(\mathbf{X}_0)\right)\right\} \\
&\le\mathrm{E}_{\mathbf{X}_0}\{|\phi_a(\mathbf{X}_0)\phi_b(\mathbf{X}_0)|\}\le\left(\mathrm{E}_{\mathbf{X}_0}\left\{\phi_a^2(\mathbf{X}_0)\right\}\right)^{1/2}\left(\mathrm{E}_{\mathbf{X}_0}\left\{\phi_b^2(\mathbf{X}_0)\right\}\right)^{1/2} \\
&\le\left(\mathrm{E}_{\mathbf{X}_0}\left\{\phi_a^{2r_*}(\mathbf{X}_0)\right\}\right)^{1/(2r_*)}\left(\mathrm{E}_{\mathbf{X}_0}\left\{\phi_b^{2r_*}(\mathbf{X}_0)\right\}\right)^{1/(2r_*)}\le\rho_*^2. \qquad (21)
\end{aligned}$$

  We apply the dominated convergence theorem using (20) and (21) to obtain that

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}_0}\left\{\sum_{j=1}^{m}\sum_{j'=1}^{m}\sum_{a=1}^{\infty}\sum_{b=1}^{\infty}\mu_a\mu_b\left\{\left[\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_{j'})\phi_a(\mathbf{x}_0)\phi_b(\mathbf{x}_0)\right\} \\
=&\sum_{j=1}^{m}\sum_{j'=1}^{m}\mathrm{E}_{\mathbf{X}_0}\left\{\lim_{N_1,N_2\to\infty}\sum_{a=1}^{N_1}\sum_{b=1}^{N_2}\mu_a\mu_b\left\{\left[\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\right. \\
&\left.\cdot\,\phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_{j'})\phi_a(\mathbf{x}_0)\phi_b(\mathbf{x}_0)\right\} \\
=&\sum_{j=1}^{m}\sum_{j'=1}^{m}\lim_{N_1,N_2\to\infty}\sum_{a=1}^{N_1}\sum_{b=1}^{N_2}\mu_a\mu_b\left\{\left[\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'} \\
&\cdot\,\phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_{j'})\mathrm{E}_{\mathbf{X}_0}\left[\phi_a(\mathbf{x}_0)\phi_b(\mathbf{x}_0)\right]
\end{aligned}$$

$$= \sum_{j=1}^{m}\sum_{j'=1}^{m}\sum_{a=1}^{\infty}\sum_{b=1}^{\infty}\mu_a\mu_b \left\{ \left[\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'} \phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_{j'})\mathrm{E}_{\mathbf{X}_0}\left[\phi_a(\mathbf{x}_0)\phi_b(\mathbf{x}_0)\right],$$

which gives the right-hand side of Step (ii).

- For Step (iii), we make the left-hand side larger by dropping the negative quadratic term in the summation $\sum_{a=\zeta+1}^{\infty}\sum_{j=1}^{\infty}\sum_{j'=1}^{\infty}$.

To proceed from (19), we define some useful quantities:

$$\begin{aligned}
&\mathbf{M} = \mathrm{diag}\left(\mu_1,\ldots,\mu_\zeta\right), \quad \mathbf{M}^{\mathrm{rem}} = \mathrm{diag}\left(\mu_{\zeta+1},\mu_{\zeta+2},\ldots\right),\\
&\boldsymbol{\phi}_a = \left[\phi_a(\mathbf{X}_1),\ldots,\phi_a(\mathbf{X}_m)\right]^\top, \text{ for } a=1,2,\ldots,\\
&\boldsymbol{\Phi} = \left[\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_\zeta\right], \quad \boldsymbol{\Phi}^{\mathrm{rem}} = \left[\boldsymbol{\phi}_{\zeta+1},\boldsymbol{\phi}_{\zeta+2},\ldots\right],\\
&\mathbf{B} = \mathbf{M} - \mathbf{M}\boldsymbol{\Phi}^\top\left[\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\boldsymbol{\Phi}\mathbf{M},
\end{aligned}$$

such that $\boldsymbol{\Phi}$ is a $m\times\zeta$ matrix, and $\mathbf{B}$ is a $\zeta\times\zeta$ positive definite matrix. From this definition and (19), we have

$$\begin{aligned}
&\mathrm{tr}(\mathbf{B}) = \sum_{a=1}^{\zeta}\mu_a - \sum_{a=1}^{\zeta}\sum_{j=1}^{m}\sum_{j'=1}^{m}\mu_a^2\left\{\left[\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\phi_a(\mathbf{X}_j)\phi_b(\mathbf{X}_{j'}),\\
&\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_{\mathbf{X}^m}\left[\mathrm{MSE}^{(M)}_{\mathrm{opt}}(\mathbf{X}_0)\right] \le \mathrm{E}_{\mathbf{X}^m}\,\mathrm{tr}(\mathbf{B}) + \mathrm{tr}\left(\boldsymbol{\Sigma}_M^{(\zeta)}\right). \qquad (22)
\end{aligned}$$

Let $\boldsymbol{\Sigma}_M^{\mathrm{rem}} = \boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m) - \boldsymbol{\Phi}\mathbf{M}\boldsymbol{\Phi}^\top = \boldsymbol{\Phi}^{\mathrm{rem}}\mathbf{M}^{\mathrm{rem}}\boldsymbol{\Phi}^{\mathrm{rem}\top}$, which is a $m\times m$ positive semi-definite matrix. Then by the Woodbury formula (Rasmussen and Williams 2006, Appendix A.3), the matrix $\mathbf{B}$ can be written as

$$\begin{aligned}
\mathbf{B} &= \mathbf{M} - \mathbf{M}\boldsymbol{\Phi}^\top\left[\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\boldsymbol{\Phi}\mathbf{M}\\
&= \left[\mathbf{M}^{-1} + \boldsymbol{\Phi}^\top\left\{\boldsymbol{\Sigma}_M^{\mathrm{rem}}+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)\right\}^{-1}\boldsymbol{\Phi}\right]^{-1}. \qquad (23)
\end{aligned}$$

By Assumption A.1 and the definition of n, we have that $\boldsymbol{\Sigma}_\epsilon(\mathbf{x}^m)$ is diagonal and $\boldsymbol{\Sigma}_\epsilon(\mathbf{x}^m) \preceq \frac{\overline{\sigma}_0^2}{n}\mathbf{I}_m$ for any value of $\mathbf{x}^m$, where $\mathbf{I}_m$ is the $m\times m$ identity matrix. Therefore, from (23), we can apply

the Woodbury formula again to obtain that

$$\begin{aligned}
\mathbf{B} &\preceq \left[\mathbf{M}^{-1} + \mathbf{\Phi}^\top \left\{\mathbf{\Sigma}_M^{\text{rem}} + \frac{\overline{\sigma}_0^2}{n}\mathbf{I}_m\right\}^{-1} \mathbf{\Phi}\right]^{-1} \\
&= \frac{\overline{\sigma}_0^2}{mn}\left[\mathbf{I}_\zeta + \frac{\overline{\sigma}_0^2}{mn}\mathbf{M}^{-1} + \frac{1}{m}\mathbf{\Phi}^\top \left(\frac{n}{\overline{\sigma}_0^2}\mathbf{\Sigma}_M^{\text{rem}} + \mathbf{I}_m\right)^{-1} \mathbf{\Phi} - \mathbf{I}_\zeta\right]^{-1} \\
&= \frac{\overline{\sigma}_0^2}{mn}\mathbf{Q}^{-2}\left\{\mathbf{I}_\zeta + \mathbf{Q}^{-1}\left[\frac{1}{m}\mathbf{\Phi}^\top \left(\frac{n}{\overline{\sigma}_0^2}\mathbf{\Sigma}_M^{\text{rem}} + \mathbf{I}_m\right)^{-1} \mathbf{\Phi} - \mathbf{I}_\zeta\right]\mathbf{Q}^{-1}\right\}^{-1}, \qquad (24)
\end{aligned}$$

where $\mathbf{Q} = \left(\mathbf{I}_\zeta + \frac{\overline{\sigma}_0^2}{mn}\mathbf{M}^{-1}\right)^{1/2}$.

Define the event $\mathcal{E}_2 = \left\{\frac{n}{\overline{\sigma}_0^2}\mathbf{\Sigma}_M^{\text{rem}} \preceq \delta_2 \mathbf{I}_m\right\}$. Then since $\mathbf{\Sigma}_M^{\text{rem}}$ is positive semi-definite, we have the relation that

$$\left\{\text{tr}\left(\tfrac{n}{\overline{\sigma}_0^2}\mathbf{\Sigma}_M^{\text{rem}}\right) \le \delta_2\right\} \subseteq \left\{\lambda_{\max}\left(\tfrac{n}{\overline{\sigma}_0^2}\mathbf{\Sigma}_M^{\text{rem}}\right) \le \delta_2\right\} \subseteq \mathcal{E}_2.$$

Therefore, by Markov's inequality and the monotone convergence theorem, we have that

$$\begin{aligned}
&\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_2^c) \le \mathbb{P}_{\mathbf{X}^m}\left\{\text{tr}\left(\tfrac{n}{\overline{\sigma}_0^2}\mathbf{\Sigma}_M^{\text{rem}}\right) > \delta_2\right\} \le \frac{1}{\delta_2}\text{E}_{\mathbf{X}^m}\,\text{tr}\left(\tfrac{n}{\overline{\sigma}_0^2}\mathbf{\Sigma}_M^{\text{rem}}\right) \\
&= \frac{n}{\overline{\sigma}_0^2\delta_2}\sum_{i=1}^{m}\sum_{a=\zeta+1}^{\infty}\mu_a \text{E}_{\mathbf{X}^m}\phi_a^2(\mathbf{X}_i) = \frac{mn}{\overline{\sigma}_0^2\delta_2}\,\text{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right). \qquad (25)
\end{aligned}$$

On the other hand, we consider the event defined in Lemma 3 with $\delta = \delta_1$, i.e.

$$\mathcal{E}_1 = \left\{\left|\left|\left|\mathbf{Q}^{-1}\left(\frac{1}{m}\mathbf{\Phi}^\top\mathbf{\Phi} - \mathbf{I}_\zeta\right)\mathbf{Q}^{-1}\right|\right|\right| \le \delta_1\right\}.$$

On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have that

$$\begin{aligned}
&\mathbf{I}_\zeta + \mathbf{Q}^{-1}\left\{\frac{1}{m}\mathbf{\Phi}^\top\left(\tfrac{n}{\overline{\sigma}_0^2}\mathbf{\Sigma}_M^{\text{rem}} + \mathbf{I}_m\right)^{-1}\mathbf{\Phi} - \mathbf{I}_\zeta\right\}\mathbf{Q}^{-1} \\
&\overset{(i)}{\succeq} \mathbf{I}_\zeta + \mathbf{Q}^{-1}\left\{\frac{1}{m}\mathbf{\Phi}^\top\left(\delta_2\mathbf{I}_m + \mathbf{I}_m\right)^{-1}\mathbf{\Phi}^\top - \mathbf{I}_\zeta\right\}\mathbf{Q}^{-1} \\
&= \mathbf{I}_\zeta - \left(1 - \frac{1}{1+\delta_2}\right)\mathbf{Q}^{-2} + \frac{1}{1+\delta_2}\mathbf{Q}^{-1}\left\{\frac{1}{m}\mathbf{\Phi}^\top\mathbf{\Phi} - \mathbf{I}_\zeta\right\}\mathbf{Q}^{-1} \\
&\overset{(ii)}{\succeq} \mathbf{I}_\zeta - \left(1 - \frac{1}{1+\delta_2}\right)\mathbf{I}_\zeta - \frac{1}{1+\delta_2}\cdot\delta_1\mathbf{I}_\zeta = \frac{1-\delta_1}{1+\delta_2}\mathbf{I}_\zeta, \qquad (26)
\end{aligned}$$

where (i) follows on the event $\mathcal{E}_2$, and (ii) holds on the event $\mathcal{E}_1$ and from the fact $\mathbf{Q}^{-2} \preceq \mathbf{I}_\zeta$.

Therefore, by combining (25), (26), and the upper bound for $\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_1^c)$ given in Lemma 3 under our assumptions A.1-A.3, we obtain that

$$
\begin{aligned}
&\mathrm{E}_{\mathbf{X}^m} \operatorname{tr}(\mathbf{B}) \leq \mathrm{E}_{\mathbf{X}^m} \left\{\operatorname{tr}(\mathbf{B}) \mathbb{1}(\mathcal{E}_1 \cap \mathcal{E}_2)\right\} + \mathrm{E}_{\mathbf{X}^m} \left[\operatorname{tr}(\mathbf{B}) \left\{\mathbb{1}(\mathcal{E}_1^c) + \mathbb{1}(\mathcal{E}_2^c)\right\}\right] \\
&\overset{(i)}{\leq} \frac{1+\delta_2}{1-\delta_1} \frac{\overline{\sigma}_0^2}{mn} \operatorname{tr}\left(\mathbf{Q}^{-2}\right) + \operatorname{tr}(\mathbf{\Sigma}_M) \left\{\mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c)\right\} \\
&\overset{(ii)}{\leq} \frac{1+\delta_2}{1-\delta_1} \frac{\overline{\sigma}_0^2}{mn} \gamma\left(\frac{\overline{\sigma}_0^2}{mn}\right) + \frac{mn}{\overline{\sigma}_0^2 \delta_2} \operatorname{tr}(\mathbf{\Sigma}_M) \operatorname{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right) + \operatorname{tr}(\mathbf{\Sigma}_M) \left\{100 \rho_*^2 \frac{b(m, \zeta, r_*) \gamma(\frac{\overline{\sigma}_0^2}{mn})}{\delta_1 \sqrt{m}}\right\}^{r_*}, \qquad (27)
\end{aligned}
$$

where (i) follows from (26), and (ii) follows from (25), Lemma 3, and the fact that

$$
\operatorname{tr}\left(\mathbf{Q}^{-2}\right) = \operatorname{tr}\left\{\left(\mathbf{I}_\zeta + \frac{\overline{\sigma}_0^2}{mn} \mathbf{M}^{-1}\right)^{-1}\right\} = \sum_{a=1}^{\zeta} \left(1 + \frac{\overline{\sigma}_0^2}{mn \mu_a}\right)^{-1} = \sum_{a=1}^{\zeta} \frac{\mu_a}{\mu_a + \frac{\overline{\sigma}_0^2}{mn}} \leq \gamma\left(\frac{\overline{\sigma}_0^2}{mn}\right).
$$

Finally, we combine (22) and (27) to obtain that

$$
\begin{aligned}
&\mathrm{E}_{\mathbf{X}^m} \mathrm{E}_{\mathbf{X}_0} \left[\mathrm{MSE}_{\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] \leq \mathrm{E}_{\mathbf{X}^m} \operatorname{tr}(\mathbf{B}) + \operatorname{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right) \\
&\leq \frac{1+\delta_2}{1-\delta_1} \frac{\overline{\sigma}_0^2}{mn} \gamma\left(\frac{\overline{\sigma}_0^2}{mn}\right) + \left\{\frac{mn}{\overline{\sigma}_0^2 \delta_2} \operatorname{tr}(\mathbf{\Sigma}_M) + 1\right\} \operatorname{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right) + \operatorname{tr}(\mathbf{\Sigma}_M) \left\{100 \rho_*^2 \frac{b(m, \zeta, r_*) \gamma(\frac{\overline{\sigma}_0^2}{mn})}{\delta_1 \sqrt{m}}\right\}^{r_*}.
\end{aligned}
$$

Taking the infimum with respect to ζ and setting $\delta_1 = \delta_2 = 1/3$ leads to the conclusion. □

Proof of Theorem 2:

We define some additional notation. For abbreviation, we write $\sigma_j^2 = \sigma^2(\mathbf{x}_j)$, $j = 1, \ldots, m$. Let $\mathcal{F} = (\mathbf{f}(\mathbf{X}_1), \ldots, \mathbf{f}(\mathbf{X}_m))^\top = (\mathrm{f}_1(\mathbf{X}^m), \ldots, \mathrm{f}_q(\mathbf{X}^m))$ be the partition of $\mathcal{F}$ according to rows and columns, respectively. For the "bias" defined in (6) of the manuscript, let $\eta(\mathbf{x}) = (\eta_1(\mathbf{x}), \ldots, \eta_q(\mathbf{x}))^\top$ for any $\mathbf{x} \in \mathcal{X}$, where $\eta_s(\mathbf{x}) = \mathrm{f}_s(\mathbf{x}) - \mathrm{f}_s(\mathbf{x}^m)^\top \left(\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)\right)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x})$. Since by Assumption A.4, $\mathrm{f}_s(\cdot) \in \mathbb{H}$ for each $s = 1, \ldots, q$ and $\mathbf{\Sigma}(\mathbf{x}_j, \cdot) \in \mathbb{H}$ for each $j = 1, \ldots, m$, we have that the function $\eta_s(\cdot)$ also lies in $\mathbb{H}$. In the following, we investigate and provide upper bound for $\|\eta_s\|_2$, $s = 1, \ldots, q$. We first expand the function $\mathrm{f}_s(\mathbf{x})$ and $\eta_s(\mathbf{x})$ in terms of the orthonormal basis

$\{\phi_l(\mathbf{x}) : l = 1, 2, \ldots\}$:

$$\mathrm{f}_s(\mathbf{x}) = \sum_{l=1}^{\infty} \theta_{sl}\phi_l(\mathbf{x}), \qquad \eta_s(\mathbf{x}) = \sum_{l=1}^{\infty} \delta_{sl}\phi_l(\mathbf{x}), \tag{28}$$

for any $\mathbf{x} \in \mathcal{X}$ and $s = 1, \ldots, q$. For a fixed $\zeta \in \mathbb{N}$, define $\theta_s^{\downarrow} = (\theta_{s1}, \ldots, \theta_{s\zeta})^\top$, $\theta_s^{\uparrow} = (\theta_{s,\zeta+1}, \theta_{s,\zeta+2}, \ldots)^\top$, $\delta_s^{\downarrow} = (\delta_{s1}, \ldots, \delta_{s\zeta})^\top$, $\delta_s^{\uparrow} = (\delta_{s,\zeta+1}, \delta_{s,\zeta+2}, \ldots)^\top$. We also define the following quantities:

$$\begin{aligned}
&\mathbf{M} = \mathrm{diag}\left(\mu_1, \ldots, \mu_\zeta\right), \\
&\boldsymbol{\phi}_l = \left[\phi_l(\mathbf{X}_1), \ldots, \phi_l(\mathbf{X}_m)\right]^\top, \text{ for } l = 1, 2, \ldots, \\
&\boldsymbol{\Phi} = \left[\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_\zeta\right], \\
&\mathbf{v}_s = (v_{s1}, \ldots, v_{sm})^\top, \quad v_{sj} = \sum_{l=\zeta+1}^{\infty} \delta_{sl}\phi_l(\mathbf{X}_j), \text{ for } j = 1, \ldots, m.
\end{aligned}$$

Then based on Assumptions A.1-A.4, we can prove Lemma 1 and Lemma 2. On the other hand, from the definition of $\mathrm{MSE}_{\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{x}_0)$ in (6) of the manuscript, we have that

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] &= \mathrm{E}_{\mathbf{X}_0}\left[\eta(\mathbf{X}_0)^\top\left[\mathcal{F}^\top\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m) + \boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathcal{F}\right]^{-1}\eta(\mathbf{X}_0)\right] \\
&\leq \lambda_{\max}\left(\left[\mathcal{F}^\top\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m) + \boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathcal{F}\right]^{-1}\right) \cdot \mathrm{E}_{\mathbf{X}_0}\left[\eta(\mathbf{X}_0)^\top\eta(\mathbf{X}_0)\right] \\
&= \left[\lambda_{\min}\left(\mathcal{F}^\top\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m) + \boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathcal{F}\right)\right]^{-1} \cdot \mathrm{E}_{\mathbf{X}_0}\left[\sum_{s=1}^{q}\eta_s(\mathbf{X}_0)^\top\eta_s(\mathbf{X}_0)\right] \\
&= \left[\lambda_{\min}\left(\mathcal{F}^\top\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m) + \boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathcal{F}\right)\right]^{-1} \cdot \sum_{s=1}^{q}\|\eta_s\|_2^2 .
\end{aligned} \tag{29}$$

For simplicity, we define Γ_m to be the quantity inside the bracelets in Theorem 2:

$$\begin{aligned}
\Gamma_m = 8C_{\mathrm{f}}^2\frac{\overline{\sigma}_0^2}{mn} + \inf_{\zeta\in\mathbb{N}}\Bigg[&8C_{\mathrm{f}}^2\frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4\,\mathrm{tr}\left(\boldsymbol{\Sigma}_M\right)\mathrm{tr}\left(\boldsymbol{\Sigma}_M^{(\zeta)}\right) + C_{\mathrm{f}}^2\,\mathrm{tr}\left(\boldsymbol{\Sigma}_M^{(\zeta)}\right) \\
&+ C_{\mathrm{f}}^2\,\mathrm{tr}\left(\boldsymbol{\Sigma}_M\right)\left\{200\rho_*^2\frac{b(m, \zeta, r_*)\gamma(\frac{\overline{\sigma}_0^2}{mn})}{\sqrt{m}}\right\}^{r_*}\Bigg].
\end{aligned}$$

From the upper bound of $\mathrm{E}_{\mathbf{X}^m}\|\eta_s\|_2^2$ in Lemma 1, it is clear that $\mathrm{E}_{\mathbf{X}^m}\|\eta_s\|_2^2 \leq \Gamma_m$ for all $s = 1, \ldots, q$ since we can make the upper bound in Lemma 1 larger by replacing each $\|\mathrm{f}_s\|_{\mathbb{H}}$ with C_{f}. From the

Markov's inequality, for any $\xi \in (0, 1/4)$,

$$\mathbb{P}_{\mathbf{X}^m}\left(\sum_{s=1}^{q}\|\eta_s\|_2^2 \geq q\Gamma_m/\xi\right) \leq \frac{\sum_{s=1}^{q}\mathrm{E}_{\mathbf{X}^m}\|\eta_s\|_2^2}{q\Gamma_m/\xi} \leq \frac{q\Gamma_m}{q\Gamma_m/\xi} = \xi. \tag{30}$$

Then from Lemma 2, we have that for any $\xi \in (0, 1/4)$, for all $m > m_0$ (with m_0 dependent on $\xi, \mathbf{\Sigma}_M, \mathbf{f}, n, \overline{\sigma}_0^2, \rho_*$),

$$\mathbb{P}_{\mathbf{X}^m}\left(\left[\lambda_{\min}\left\{\mathcal{F}^\top\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathcal{F}\right\}\right]^{-1} > \frac{8\,\mathrm{tr}(\mathbf{\Sigma}_M)}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\right) < \xi. \tag{31}$$

We combine (29), (30) and (31) together to conclude that for any $\xi \in (0, 1/4)$, for all $m > m_0$, there exists a constant $c_\xi = 1/\xi$, such that

$$\begin{aligned}
&\mathbb{P}_{\mathbf{X}^m}\left(\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] > c_\xi\cdot\frac{8q\,\mathrm{tr}(\mathbf{\Sigma}_M)}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\Gamma_m\right)\\
&\leq \mathbb{P}_{\mathbf{X}^m}\left(\left[\lambda_{\min}\left(\mathcal{F}^\top\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathcal{F}\right)\right]^{-1}\cdot\sum_{s=1}^{q}\|\eta_s\|_2^2 > \frac{8\,\mathrm{tr}(\mathbf{\Sigma}_M)}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\cdot\frac{q\Gamma_m}{\xi}\right)\\
&\leq \mathbb{P}_{\mathbf{X}^m}\left(\sum_{s=1}^{q}\|\eta_s\|_2^2 \geq q\Gamma_m/\xi\right)\\
&\quad+\mathbb{P}_{\mathbf{X}^m}\left(\left[\lambda_{\min}\left\{\mathcal{F}^\top\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathcal{F}\right\}\right]^{-1} > \frac{8\,\mathrm{tr}(\mathbf{\Sigma}_M)}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\right)\\
&< \xi+\xi = 2\xi.
\end{aligned} \tag{32}$$

This has proved that $\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{8q\,\mathrm{tr}(\mathbf{\Sigma}_M)}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\Gamma_m$, which is the conclusion of Theorem 2. □

LEMMA 1. *Under Assumptions A.1-A.4, we have that for each $s = 1, \ldots, q$,*

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}^m}\|\eta_s\|_2^2 \leq \frac{8\|\mathrm{f}_s\|_{\mathbb{H}}^2\overline{\sigma}_0^2}{mn} + \inf_{\zeta\in\mathbb{N}}\Bigg[&\frac{8\|\mathrm{f}_s\|_{\mathbb{H}}^2 mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right) + \|\mathrm{f}_s\|_{\mathbb{H}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right)\\
&+\|\mathrm{f}_s\|_{\mathbb{H}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\left\{200\rho_*^2\frac{b(m,\zeta,r_*)\gamma(\frac{\overline{\sigma}_0^2}{mn})}{\sqrt{m}}\right\}^{r_*}\Bigg].
\end{aligned}$$

Proof of Lemma 1:

By Assumption A.1, we have that $\mathbf{\Sigma}_\epsilon(\mathbf{x}^m) = \mathrm{diag}\left(\sigma_1^2/n, \ldots, \sigma_m^2/n\right)$, where we let $\sigma_j^2 = \sigma^2(\mathbf{x}_j)$

for $j = 1, \ldots, m$. For any $\mathbf{x} \in \mathcal{X}$ and any $s \in \{1, \ldots, q\}$, we have the following relation:

$$
\begin{aligned}
&\sum_{j=1}^{m} \frac{n}{\sigma_j^2} \eta_s(\mathbf{x}_j) \mathbf{\Sigma}_M(\mathbf{x}_j, \mathbf{x}) \\
&= \sum_{j=1}^{m} \frac{n}{\sigma_j^2} \left\{ \mathrm{f}_s(\mathbf{x}_j) - \mathrm{f}_s(\mathbf{x}^m)^\top \left(\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{x}^m) \right)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}_j) \right\} \mathbf{\Sigma}_M(\mathbf{x}_j, \mathbf{x}) \\
&= \sum_{j=1}^{m} \frac{n}{\sigma_j^2} \mathrm{f}_s(\mathbf{x}_j) \mathbf{\Sigma}_M(\mathbf{x}_j, \mathbf{x}) - \sum_{j=1}^{m} \frac{n}{\sigma_j^2} \mathrm{f}_s(\mathbf{x}^m)^\top \left(\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{x}^m) \right)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}_j) \mathbf{\Sigma}_M(\mathbf{x}_j, \mathbf{x}) \\
&= \mathrm{f}_s(\mathbf{x}^m)^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}) \\
&\quad - \mathrm{f}_s(\mathbf{x}^m)^\top \left(\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{x}^m) \right)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}) \\
&= \mathrm{f}_s(\mathbf{x}^m)^\top \left(\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{x}^m) \right)^{-1} \left\{ \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{x}^m) - \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) \right\} \\
&\quad \cdot \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}) \\
&= \mathrm{f}_s(\mathbf{x}^m)^\top \left(\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{x}^m) \right)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}) \\
&= \mathrm{f}_s(\mathbf{x}) - \eta_s(\mathbf{x}).
\end{aligned} \tag{33}
$$

Therefore, we can rewrite (33) as

$$
\sum_{j=1}^{m} \frac{n}{\sigma_j^2} \eta_s(\mathbf{x}_j) \mathbf{\Sigma}_M(\mathbf{x}_j, \mathbf{x}) + \eta_s(\mathbf{x}) - \mathrm{f}_s(\mathbf{x}) = 0, \tag{34}
$$

for any $\mathbf{x} \in \mathcal{X}$ and any $s \in \{1, \ldots, q\}$.

We proceed with (34) in two ways. On one hand, we can take the $\mathbb{H}$-norm of f_s in (34). Since $\eta_s \in \mathbb{H}$ and it has the expansion in (28), we can derive from (34) that

$$
\begin{aligned}
\mathrm{f}_s(\mathbf{x}) &= \sum_{j=1}^{m} \frac{n}{\sigma_j^2} \sum_{a=1}^{\infty} \delta_{sa} \phi_a(\mathbf{x}_j) \sum_{b=1}^{\infty} \mu_b \phi_b(\mathbf{x}_j) \phi_b(\mathbf{x}) + \sum_{b=1}^{\infty} \delta_{sb} \phi_b(\mathbf{x}) \\
&= \sum_{b=1}^{\infty} \left\{ \mu_b \sum_{j=1}^{m} \frac{n}{\sigma_j^2} \sum_{a=1}^{\infty} \delta_{sa} \phi_a(\mathbf{x}_j) \phi_b(\mathbf{x}_j) + \delta_{sb} \right\} \phi_b(\mathbf{x}), \\
\|\mathrm{f}_s\|_{\mathbb{H}}^2 &= \sum_{b=1}^{\infty} \frac{1}{\mu_b} \left\{ \mu_b \sum_{j=1}^{m} \frac{n}{\sigma_j^2} \sum_{a=1}^{\infty} \delta_{sa} \phi_a(\mathbf{x}_j) \phi_b(\mathbf{x}_j) + \delta_{sb} \right\}^2 \\
&= \sum_{b=1}^{\infty} \frac{\delta_{sb}^2}{\mu_b} + 2 \sum_{b=1}^{\infty} \sum_{j=1}^{m} \frac{n}{\sigma_j^2} \sum_{a=1}^{\infty} \delta_{sa} \delta_{sb} \phi_a(\mathbf{x}_j) \phi_b(\mathbf{x}_j) + \sum_{b=1}^{\infty} \mu_b \left\{ \sum_{j=1}^{m} \frac{n}{\sigma_j^2} \sum_{a=1}^{\infty} \delta_{sa} \phi_a(\mathbf{x}_j) \phi_b(\mathbf{x}_j) \right\}^2
\end{aligned}
$$

$$
\begin{aligned}
&= \|\eta_s\|_{\mathbb{H}}^2 + 2\sum_{j=1}^{m}\frac{n}{\sigma_j^2}\left\{\sum_{a=1}^{\infty}\delta_{sa}\phi_a(\mathbf{x}_j)\right\}^2 + \sum_{b=1}^{\infty}\mu_b\left\{\sum_{j=1}^{m}\frac{n}{\sigma_j^2}\sum_{a=1}^{\infty}\delta_{sa}\phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_j)\right\}^2 \\
&\geq \|\eta_s\|_{\mathbb{H}}^2, \\
\Longrightarrow \|\eta_s\|_{\mathbb{H}} &\leq \|\mathrm{f}_s\|_{\mathbb{H}}. \qquad (35)
\end{aligned}
$$

On the other hand, we take $\mathbb{H}$-inner product of the left-hand-side of (34) with $\phi_l(\mathbf{x})$ for any fixed l with $\mu_l > 0$, and obtain that

$$
\begin{aligned}
0 &= \sum_{j=1}^{m}\frac{n}{\sigma_j^2}\eta_s(\mathbf{x}_j)\langle\boldsymbol{\Sigma}_M(\mathbf{x}_j,\mathbf{x}),\phi_l(\mathbf{x})\rangle_{\mathbb{H}} + \langle\eta_s(\mathbf{x}),\phi_l(\mathbf{x})\rangle_{\mathbb{H}} - \langle\mathrm{f}_s(\mathbf{x}),\phi_l(\mathbf{x})\rangle_{\mathbb{H}}, \\
&= \sum_{j=1}^{m}\frac{n}{\sigma_j^2}\eta_s(\mathbf{x}_j)\phi_l(\mathbf{x}_j) + \frac{\delta_{sl}}{\mu_l} - \frac{\theta_{sl}}{\mu_l}, \\
&= \sum_{j=1}^{m}\frac{n}{\sigma_j^2}\sum_{a=1}^{\infty}\delta_{sa}\phi_a(\mathbf{x}_j)\phi_l(\mathbf{x}_j) + \frac{\delta_{sl}}{\mu_l} - \frac{\theta_{sl}}{\mu_l}, \\
&= \sum_{j=1}^{m}\sum_{a=1}^{\zeta}\frac{n}{\sigma_j^2}\delta_{sa}\phi_a(\mathbf{x}_j)\phi_l(\mathbf{x}_j) + \sum_{j=1}^{m}\frac{n}{\sigma_j^2}v_{sj}\phi_l(\mathbf{x}_j) + \frac{\delta_{sl}}{\mu_l} - \frac{\theta_{sl}}{\mu_l} \qquad (36)
\end{aligned}
$$

where we have used the reproducing property for the function $\phi_l \in \mathbb{H}$. We can then stack (36) in a column for $l = 1, \ldots, \zeta$ for some $\zeta \in \mathbb{N}$ with $\mu_\zeta > 0$, and obtain that

$$
\begin{aligned}
&\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\boldsymbol{\Phi}\delta_s^{\downarrow} + \boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s + \mathbf{M}^{-1}\delta_s^{\downarrow} - \mathbf{M}^{-1}\theta_s^{\downarrow} = 0, \\
\Longrightarrow\ &\delta_s^{\downarrow} = \left(\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\boldsymbol{\Phi} + \mathbf{M}^{-1}\right)^{-1}\left(\mathbf{M}^{-1}\theta_s^{\downarrow} - \boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right), \\
&= \left(\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\boldsymbol{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\left(\mathbf{Q}^{-1}\mathbf{M}^{-1}\theta_s^{\downarrow} - \mathbf{Q}^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right), \qquad (37)
\end{aligned}
$$

where $\mathbf{Q} = \left(\mathbf{I}_\zeta + \frac{\overline{\sigma}_0^2}{mn}\mathbf{M}^{-1}\right)^{1/2}$ as defined in Lemma 3. Therefore,

$$
\left\|\delta_s^{\downarrow}\right\| \leq \left\|\left(\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\boldsymbol{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right\|\left(\left\|\mathbf{Q}^{-1}\mathbf{M}^{-1}\theta_s^{\downarrow}\right\| + \left\|\mathbf{Q}^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\|\right). \qquad (38)
$$

By Assumption A.1, we have that $\boldsymbol{\Sigma}_\epsilon(\mathbf{x}^m) \preceq \frac{\overline{\sigma}_0^2}{n}\mathbf{I}_m$. Therefore, $\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\boldsymbol{\Phi} + \mathbf{M}^{-1} \succeq$

$\frac{n}{\overline{\sigma}_0^2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{M}^{-1} \succ 0$. This implies that

$$0 \prec \mathbf{Q}^{1/2}\left(\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}^{1/2} \preceq \mathbf{Q}^{1/2}\left(\frac{n}{\overline{\sigma}_0^2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}^{1/2}. \tag{39}$$

Note that the matrices $\left(\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}$, $\left(\frac{n}{\overline{\sigma}_0^2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}$, and $\mathbf{Q}$ are all symmetric and positive definite matrices. Furthermore, $\left(\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}$ is similar to the symmetric positive definite matrix $\mathbf{Q}^{1/2}\left(\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}^{1/2}$. Therefore,

$$\lambda_{\max}\left\{\left(\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right\} = \lambda_{\max}\left\{\mathbf{Q}^{1/2}\left(\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}^{1/2}\right\}, \tag{40}$$

and similarly

$$\lambda_{\max}\left\{\left(\frac{n}{\overline{\sigma}_0^2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right\} = \lambda_{\max}\left\{\mathbf{Q}^{1/2}\left(\frac{n}{\overline{\sigma}_0^2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}^{1/2}\right\}. \tag{41}$$

(39), (40), and (41) imply that

$$\begin{aligned}
&\left|\left|\left|\left(\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right|\right|\right| = \lambda_{\max}\left\{\left(\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right\} \\
&\leq \lambda_{\max}\left\{\left(\frac{n}{\overline{\sigma}_0^2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right\} = \left|\left|\left|\left(\frac{n}{\overline{\sigma}_0^2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right|\right|\right| \\
&= \frac{\overline{\sigma}_0^2}{mn}\left|\left|\left|\left\{\left(\mathbf{I}_\zeta + \frac{\overline{\sigma}_0^2}{mn}\mathbf{M}^{-1}\right) + \left(\frac{1}{m}\mathbf{\Phi}^\top\mathbf{\Phi} - \mathbf{I}_\zeta\right)\right\}^{-1}\mathbf{Q}\right|\right|\right| \\
&= \frac{\overline{\sigma}_0^2}{mn}\left|\left|\left|\mathbf{Q}^{-1}\left\{\mathbf{I}_\zeta + \mathbf{Q}^{-1}\left(\frac{1}{m}\mathbf{\Phi}^\top\mathbf{\Phi} - \mathbf{I}_\zeta\right)\mathbf{Q}^{-1}\right\}^{-1}\right|\right|\right|, \\
&\leq \frac{\overline{\sigma}_0^2}{mn}|||\mathbf{Q}^{-1}|||\left|\left|\left|\left\{\mathbf{I}_\zeta + \mathbf{Q}^{-1}\left(\frac{1}{m}\mathbf{\Phi}^\top\mathbf{\Phi} - \mathbf{I}_\zeta\right)\mathbf{Q}^{-1}\right\}^{-1}\right|\right|\right|
\end{aligned} \tag{42}$$

We consider the event defined in Lemma 3 with $\delta = 1/2$, i.e.

$$\mathcal{E}_3 = \left\{\left|\left|\left|\mathbf{Q}^{-1}\left(\frac{1}{m}\mathbf{\Phi}^\top\mathbf{\Phi} - \mathbf{I}_\zeta\right)\mathbf{Q}^{-1}\right|\right|\right| \leq \frac{1}{2}\right\}.$$

Then on the event $\mathcal{E}_3$, $\mathbf{I}_\zeta + \mathbf{Q}^{-1}\left(\frac{1}{m}\mathbf{\Phi}^\top\mathbf{\Phi} - \mathbf{I}_\zeta\right)\mathbf{Q}^{-1} \succeq (1 - 1/2)\mathbf{I}_\zeta = (1/2)\mathbf{I}_\zeta$. Moreover, $0 \prec \mathbf{Q}^{-1} \prec$

$\mathbf{I}_\zeta$. Therefore, (42) implies that

$$\left|\!\left|\!\left|\left(\mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right|\!\right|\!\right| \leq \frac{2\overline{\sigma}_0^2}{mn}\left|\!\left|\!\left|\mathbf{Q}^{-1}\right|\!\right|\!\right| \leq \frac{2\overline{\sigma}_0^2}{mn}. \tag{43}$$

In (38), the term $\left\|\mathbf{Q}^{-1}\mathbf{M}^{-1}\theta_s^\downarrow\right\|$ can be bounded as

$$\begin{aligned}
&\left\|\mathbf{Q}^{-1}\mathbf{M}^{-1}\theta_s^\downarrow\right\| = \sqrt{\left(\theta_s^\downarrow\right)^\top \mathbf{M}^{-1}\mathbf{Q}^{-2}\mathbf{M}^{-1}\theta_s^\downarrow} = \sqrt{\left(\theta_s^\downarrow\right)^\top \left(\mathbf{M}^2 + \frac{\overline{\sigma}_0^2}{mn}\mathbf{M}\right)^{-1}\theta_s^\downarrow} \\
&\leq \sqrt{\left(\theta_s^\downarrow\right)^\top \left(\frac{\overline{\sigma}_0^2}{mn}\mathbf{M}\right)^{-1}\theta_s^\downarrow} = \sqrt{\frac{mn}{\overline{\sigma}_0^2}}\sqrt{\sum_{l=1}^{\zeta}\frac{\theta_{sl}^2}{\mu_l^2}} \leq \sqrt{\frac{mn}{\overline{\sigma}_0^2}}\left\|\mathrm{f}_s\right\|_{\mathbb{H}}.
\end{aligned} \tag{44}$$

For the term $\left\|\mathbf{Q}^{-1}\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\|$ in (38), we first have that

$$\begin{aligned}
&\left\|\mathbf{Q}^{-1}\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\| = \left\|\left(\mathbf{M} + \frac{\overline{\sigma}_0^2}{mn}\mathbf{I}_\zeta\right)^{-1/2}\mathbf{M}^{1/2}\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\| \\
&\leq \left|\!\left|\!\left|\left(\mathbf{M} + \frac{\overline{\sigma}_0^2}{mn}\mathbf{I}_\zeta\right)^{-1/2}\right|\!\right|\!\right| \cdot \left\|\mathbf{M}^{1/2}\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\| = \frac{1}{\sqrt{\mu_\zeta + \frac{\overline{\sigma}_0^2}{mn}}}\left\|\mathbf{M}^{1/2}\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\| \\
&\leq \sqrt{\frac{mn}{\overline{\sigma}_0^2}}\left\|\mathbf{M}^{1/2}\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\| = \sqrt{\frac{mn}{\overline{\sigma}_0^2}}\sqrt{\sum_{l=1}^{\zeta}\mu_l\left(\boldsymbol{\phi}_l^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right)^2} \\
&\overset{(i)}{\leq} \sqrt{\frac{mn}{\overline{\sigma}_0^2}}\left\{\sum_{l=1}^{\zeta}\mu_l\left(\boldsymbol{\phi}_l^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\boldsymbol{\phi}_l\right)\left(\mathbf{v}_s^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right)\right\}^{1/2} \overset{(ii)}{\leq} \frac{n}{\underline{\sigma}_0^2}\sqrt{\frac{mn}{\overline{\sigma}_0^2}}\sqrt{\sum_{l=1}^{\zeta}\mu_l\left\|\boldsymbol{\phi}_l\right\|^2\left\|\mathbf{v}_s\right\|^2},
\end{aligned} \tag{45}$$

where (i) follows from the Cauchy-Schwarz inequality, and (ii) follows from Assumption A.1 that $\sigma_j^2 \geq \underline{\sigma}_0^2$ for all $j = 1, \ldots, m$ and hence $\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \preceq \frac{n}{\underline{\sigma}_0^2}\mathbf{I}_\zeta$.

We can combine (38), (43), (44), (45), and apply the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ to obtain that

$$\begin{aligned}
&\left\|\delta_s^\downarrow\right\|^2 \leq 2\left|\!\left|\!\left|\left(\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right|\!\right|\!\right|^2\left(\left\|\mathbf{Q}^{-1}\mathbf{M}^{-1}\theta_s^\downarrow\right\|^2 + \left\|\mathbf{Q}^{-1}\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\|^2\right) \\
&\leq 2\left(\frac{2\overline{\sigma}_0^2}{mn}\right)^2\left\{\frac{mn}{\overline{\sigma}_0^2}\left\|\mathrm{f}_s\right\|_{\mathbb{H}}^2 + \left(\frac{n}{\underline{\sigma}_0^2}\right)^2\frac{mn}{\overline{\sigma}_0^2}\sum_{l=1}^{d}\mu_l\left\|\boldsymbol{\phi}_l\right\|^2\left\|\mathbf{v}_s\right\|^2\right\} \\
&= 8\left\{\frac{\overline{\sigma}_0^2}{mn}\left\|\mathrm{f}_s\right\|_{\mathbb{H}}^2 + \frac{n\overline{\sigma}_0^2}{m\underline{\sigma}_0^4}\sum_{l=1}^{\zeta}\mu_l\left\|\boldsymbol{\phi}_l\right\|^2\left\|\mathbf{v}_s\right\|^2\right\}.
\end{aligned} \tag{46}$$

Now we evaluate the expectation $\mathrm{E}_{\mathbf{X}^m}\|\delta_s^{\downarrow}\|^2$. From (46), it suffices to control $\mathrm{E}_{\mathbf{X}^m}\left(\|\boldsymbol{\phi}_l\|^2\|\mathbf{v}_s\|^2\right)$ for $l=1,\ldots,d$. By the Cauchy-Schwarz inquality,

$$\mathrm{E}_{\mathbf{X}^m}\left(\|\boldsymbol{\phi}_l\|^2\|\mathbf{v}_s\|^2\right) \le \sqrt{\mathrm{E}_{\mathbf{X}^m}\left(\|\boldsymbol{\phi}_l\|^4\right)}\sqrt{\mathrm{E}_{\mathbf{X}^m}\left(\|\mathbf{v}_s\|^4\right)}. \tag{47}$$

By Assumption A.3, $\mathrm{E}_{\mathbf{P_X}}\left\{\phi_l^{2r_*}(\mathbf{X})\right\} \le \rho_*^{2r_*}$ for some $r_* \ge 2$. By Jensen's inequality, for all $l=1,2,\ldots,$

$$\mathrm{E}_{\mathbf{P_X}}\left\{\phi_l^4(\mathbf{X})\right\} \le \left[\mathrm{E}_{\mathbf{P_X}}\left\{\phi_l^{2r_*}(\mathbf{X})\right\}\right]^{2/r_*} \le \rho_*^{2r_*\cdot 2/r_*} = \rho_*^4.$$

Since $\mathbf{X}_1,\ldots,\mathbf{X}_m$ are i.i.d. distributed as $\mathbf{P_X}$ and $\mathrm{E}_{\mathbf{P_X}}\left\{\phi_l^4(\mathbf{X})\right\} \le \rho_*^4$ for all l, we have that

$$\begin{aligned}\mathrm{E}_{\mathbf{X}^m}\left(\|\boldsymbol{\phi}_l\|^4\right) &= \mathrm{E}_{\mathbf{X}^m}\left\{\left(\sum_{j=1}^m \phi_l^2(\mathbf{X}_j)\right)^2\right\} \\ &\le \mathrm{E}_{\mathbf{X}^m}\left(m\sum_{j=1}^m \phi_l^4(\mathbf{X}_j)\right) \le m^2\mathrm{E}_{\mathbf{X}^m}\left(\phi_l^4(\mathbf{X}_1)\right) \le m^2\rho_*^4.\end{aligned} \tag{48}$$

On the other hand, by applying the Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\mathrm{E}_{\mathbf{X}^m}\left(\|\mathbf{v}_s\|^4\right) = \mathrm{E}_{\mathbf{X}^m}\left\{\left(\sum_{j=1}^m v_{sj}^2\right)^2\right\} \le m\mathrm{E}_{\mathbf{X}^m}\left(\sum_{j=1}^m v_{sj}^4\right) = m^2\mathrm{E}_{\mathbf{X}^m}\left(v_{s1}^4\right) \\ &= m^2\mathrm{E}_{\mathbf{X}^m}\left\{\left(\sum_{l=\zeta+1}^{\infty}\delta_{sl}\phi_l(\mathbf{X}_1)\right)^4\right\} \le m^2\mathrm{E}_{\mathbf{X}^m}\left[\left\{\sum_{l=\zeta+1}^{\infty}\frac{\delta_{sl}^2}{\mu_l}\cdot\sum_{l=\zeta+1}^{\infty}\mu_l\phi_l^2(\mathbf{X}_1)\right\}^2\right].\end{aligned} \tag{49}$$

From (35), we can get an upper bound $\sum_{l=\zeta+1}^{\infty}\frac{\delta_{sl}^2}{\mu_l} \le \sum_{l=1}^{\infty}\frac{\delta_{sl}^2}{\mu_l} = \|\eta_s\|_{\mathbb{H}}^2 \le \|\mathrm{f}_s\|_{\mathbb{H}}^2$. Therefore, (49) further implies that

$$\begin{aligned}\mathrm{E}_{\mathbf{X}^m}\left(\|\mathbf{v}_s\|^4\right) &\le m^2\|\mathrm{f}_s\|_{\mathbb{H}}^4\cdot\mathrm{E}_{\mathbf{X}^m}\left[\left\{\sum_{l=\zeta+1}^{\infty}\mu_l\phi_l^2(\mathbf{X}_1)\right\}^2\right] \\ &= m^2\|\mathrm{f}_s\|_{\mathbb{H}}^4\cdot\mathrm{E}_{\mathbf{X}^m}\left\{\sum_{a=\zeta+1}^{\infty}\sum_{b=\zeta+1}^{\infty}\mu_a\mu_b\phi_a^2(\mathbf{X}_1)\phi_b^2(\mathbf{X}_1)\right\} \\ &\overset{(i)}{\le} m^2\|\mathrm{f}_s\|_{\mathbb{H}}^4\cdot\left\{\sum_{a=\zeta+1}^{\infty}\sum_{b=\zeta+1}^{\infty}\mu_a\mu_b\sqrt{\mathrm{E}_{\mathbf{X}^m}\phi_a^4(\mathbf{X}_1)\cdot\mathrm{E}_{\mathbf{X}^m}\phi_b^4(\mathbf{X}_1)}\right\}\end{aligned}$$

$$\overset{(ii)}{\leq} m^2\rho_*^4\|\mathrm{f}_s\|_{\mathbb{H}}^4 \sum_{a=\zeta+1}^{\infty}\sum_{b=\zeta+1}^{\infty}\mu_a\mu_b = m^2\rho_*^4\|\mathrm{f}_s\|_{\mathbb{H}}^4\left\{\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right)\right\}^2, \tag{50}$$

where (i) follows from the Cauchy-Schwarz inequality and the monotone convergence theorem, and (ii) follows from Assumption A.3.

We combine (46), (47), (48), and (50), and to obtain that

$$\begin{aligned}\mathrm{E}_{\mathbf{X}^m}\left(\left\|\delta_s^{\downarrow}\right\|^2 \;\middle|\; \mathcal{E}_3\right) &\leq 8\left\{\frac{\overline{\sigma}_0^2}{mn}\|\mathrm{f}_s\|_{\mathbb{H}}^2 + \frac{n\overline{\sigma}_0^2}{m\underline{\sigma}_0^4}\sum_{l=1}^{\zeta}\mu_l\cdot m^2\rho_*^4\|\mathrm{f}_s\|_{\mathbb{H}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right)\right\}\\ &\leq 8\|\mathrm{f}_s\|_{\mathbb{H}}^2\left\{\frac{\overline{\sigma}_0^2}{mn} + \frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right)\right\}. \end{aligned}\tag{51}$$

We also have the coarse upper bound for $\mathrm{E}_{\mathbf{X}^m}\|\delta_s^{\downarrow}\|^2$ using (35):

$$\mathrm{E}_{\mathbf{X}^m}\left\|\delta_s^{\downarrow}\right\|^2 = \sum_{l=1}^{\zeta}\delta_{sl}^2 \leq \sum_{l=1}^{\infty}\delta_{sl}^2 \leq \mu_1\sum_{l=1}^{\infty}\frac{\delta_{sl}^2}{\mu_l} = \mu_1\|\eta_s\|_{\mathbb{H}}^2 \leq \mu_1\|\mathrm{f}_s\|_{\mathbb{H}}^2 \leq \|\mathrm{f}_s\|_{\mathbb{H}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right). \tag{52}$$

This together with the upper bound for $\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_3^c)$ in Lemma 3 (with $\delta = 1/2$) implies that

$$\begin{aligned}\mathrm{E}_{\mathbf{X}^m}\left\|\delta_s^{\downarrow}\right\|^2 &= \mathrm{E}_{\mathbf{X}^m}\left\{\left\|\delta_s^{\downarrow}\right\|^2\mathbb{1}(\mathcal{E}_3)\right\} + \mathrm{E}_{\mathbf{X}^m}\left\{\left\|\delta_s^{\downarrow}\right\|^2\mathbb{1}(\mathcal{E}_3^c)\right\}\\ &\leq \mathrm{E}_{\mathbf{X}^m}\left(\left\|\delta_s^{\downarrow}\right\|^2 \;\middle|\; \mathcal{E}_3\right)\cdot\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_3) + \mathrm{E}_{\mathbf{X}^m}\left\|\delta_s^{\downarrow}\right\|^2\cdot\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_3^c)\\ &\leq \mathrm{E}_{\mathbf{X}^m}\left(\left\|\delta_s^{\downarrow}\right\|^2 \;\middle|\; \mathcal{E}_3\right) + \mathrm{E}_{\mathbf{X}^m}\left\|\delta_s^{\downarrow}\right\|^2\cdot\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_3^c)\\ &\leq 8\|\mathrm{f}_s\|_{\mathbb{H}}^2\left\{\frac{\overline{\sigma}_0^2}{mn} + \frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right)\right\}\\ &\quad + \|\mathrm{f}_s\|_{\mathbb{H}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\left\{200\rho_*^2\frac{b(m,\zeta,r_*)\gamma(\frac{\overline{\sigma}_0^2}{mn})}{\sqrt{m}}\right\}^{r_*}. \end{aligned}\tag{53}$$

On the other hand, from (35), we have that

$$\begin{aligned}\left\|\delta_s^{\uparrow}\right\|^2 &= \sum_{l=\zeta+1}^{\infty}\delta_{sl}^2 \leq \mu_{\zeta+1}\sum_{l=\zeta+1}^{\infty}\frac{\delta_{sl}^2}{\mu_l} \leq \mu_{\zeta+1}\sum_{l=1}^{\infty}\frac{\delta_{sl}^2}{\mu_l}\\ &= \mu_{\zeta+1}\|\eta_s\|_{\mathbb{H}}^2 \leq \mu_{\zeta+1}\|\mathrm{f}_s\|_{\mathbb{H}}^2 \leq \mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right)\|\mathrm{f}_s\|_{\mathbb{H}}^2. \end{aligned}\tag{54}$$

Therefore, (53) and (54) together imply that

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}^m}\|\eta_s\|_2^2 &= \mathrm{E}_{\mathbf{X}^m}\|\delta_s\|^2 = \mathrm{E}_{\mathbf{X}^m}\left\|\delta_s^{\downarrow}\right\|^2 + \mathrm{E}_{\mathbf{X}^m}\left\|\delta_s^{\uparrow}\right\|^2 \\
&\le 8\|\mathrm{f}_s\|_{\mathbb{H}}^2\frac{\overline{\sigma}_0^2}{mn} + 8\|\mathrm{f}_s\|_{\mathbb{H}}^2\frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right) + \|\mathrm{f}_s\|_{\mathbb{H}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right) \\
&\quad + \|\mathrm{f}_s\|_{\mathbb{H}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\left\{200\rho_*^2\frac{b(m,\zeta,r_*)\gamma(\frac{\overline{\sigma}_0^2}{mn})}{\sqrt{m}}\right\}^{r_*}. \qquad (55)
\end{aligned}$$

Taking the infimum with respect to ζ leads to the result. □

LEMMA 2. *Under Assumptions A.1-A.4, for any $\xi \in (0,1)$, there exists a large integer $m_0 \in \mathbb{N}$ that depends on ξ, $\mathbf{\Sigma}_M$, $\mathbf{f}$, n, $\overline{\sigma}_0^2$ in Assumption A.1, and ρ_* in Assumption A.3, such that for all $m > m_0$,*

$$\mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\min}\left\{\mathcal{F}^\top\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathcal{F}\right\} \ge \frac{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}{8\,\mathrm{tr}(\mathbf{\Sigma}_M)}\right) \ge 1-\xi.$$

Proof of Lemma 2:

$$\begin{aligned}
&\lambda_{\min}\left\{\mathcal{F}^\top\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathcal{F}\right\} = \min_{\|a\|=1} a^\top\mathcal{F}^\top\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathcal{F}a \\
&\ge \lambda_{\min}\left\{m\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\right\}\cdot\min_{\|a\|=1} a^\top\left(\frac{1}{m}\mathcal{F}^\top\mathcal{F}\right)a \\
&= \lambda_{\min}\left\{m\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\right\}\cdot\lambda_{\min}\left(\frac{1}{m}\mathcal{F}^\top\mathcal{F}\right).
\end{aligned}$$

Therefore, for any constants $c_1, c_2 > 0$,

$$\begin{aligned}
&\mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\min}\left\{\mathcal{F}^\top\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathcal{F}\right\} < c_1c_2\right) \\
&\le \mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\min}\left(\frac{1}{m}\mathcal{F}^\top\mathcal{F}\right) < c_1\right) + \mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\min}\left\{m\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\right\} < c_2\right) \\
&= \mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\min}\left(\frac{1}{m}\mathcal{F}^\top\mathcal{F}\right) < c_1\right) + \mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\max}\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right) > m/c_2\right). \qquad (56)
\end{aligned}$$

We choose the values of c_1 and c_2 and bound the two terms separately. Since $\frac{1}{m}\mathcal{F}^\top\mathcal{F} = \frac{1}{m}\sum_{j=1}^m \mathbf{f}(\mathbf{X}_j)\mathbf{f}(\mathbf{X}_j)^\top$, by the strong law of large numbers, $\frac{1}{m}\mathcal{F}^\top\mathcal{F} \xrightarrow{a.s.} \mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]$ as $m \to \infty$,

where $\xrightarrow{a.s.}$ means the almost sure convergence. Since $\lambda_{\min}(\cdot)$ is a continuous function, by the continuous mapping theorem, $\lambda_{\min}\left(\frac{1}{m}\mathcal{F}^\top\mathcal{F}\right) \xrightarrow{a.s.} \lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)$ as $m \to \infty$. Therefore, we can set $c_1 = \frac{1}{2}\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)$, and for any given constant $\xi \in (0,1)$, there exists a large integer $m_1 = m_1(\xi) \in \mathbb{N}$, such that for all $m \geq m_1$,

$$\begin{aligned}
&\mathbb{P}_{\mathbf{X}^m}\left(\left|\lambda_{\min}\left(\frac{1}{m}\mathcal{F}^\top\mathcal{F}\right) - \lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)\right| > \frac{1}{2}\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)\right) < \frac{\xi}{2} \\
\Longrightarrow\ &\mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\min}\left(\frac{1}{m}\mathcal{F}^\top\mathcal{F}\right) < \frac{1}{2}\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)\right) < \frac{\xi}{2}. \qquad (57)
\end{aligned}$$

On the other hand, we know that by Assumption A.1, $\mathbf{\Sigma}_\epsilon(\mathbf{X}^m) \preceq \frac{\overline{\sigma}_0^2}{n}\mathbf{I}_m$. Moreover, using the monotone convergence theorem, the expectation and the variance of $\mathrm{tr}(\mathbf{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m))$ can be controlled as follows:

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}^m}\,\mathrm{tr}\left(\mathbf{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m)\right) = \mathrm{E}_{\mathbf{X}^m}\left\{\sum_{j=1}^m\sum_{l=1}^\infty \mu_l\phi_l^2(\mathbf{X}_j)\right\} = \sum_{j=1}^m\sum_{l=1}^\infty \mu_l \mathrm{E}_{\mathbf{X}^m}\left\{\phi_l^2(\mathbf{X}_j)\right\} \\
&= m\sum_{l=1}^\infty \mu_l = m\,\mathrm{tr}(\mathbf{\Sigma}_M), \qquad (58) \\
&\mathrm{Var}_{\mathbf{X}^m}\left\{\mathrm{tr}\left(\mathbf{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m)\right)\right\} = \mathrm{Var}_{\mathbf{X}^m}\left\{\sum_{j=1}^m\sum_{l=1}^\infty \mu_l\phi_l^2(\mathbf{X}_j)\right\} \\
&\overset{(i)}{=} \sum_{j=1}^m \mathrm{Var}_{\mathbf{X}^m}\left\{\sum_{l=1}^\infty \mu_l\phi_l^2(\mathbf{X}_j)\right\} \overset{(ii)}{\leq} \sum_{j=1}^m \mathrm{E}_{\mathbf{X}^m}\left\{\sum_{l=1}^\infty \mu_l\phi_l^2(\mathbf{X}_j)\right\}^2 \\
&\overset{(iii)}{=} \sum_{j=1}^m\sum_{a=1}^\infty\sum_{a=1}^\infty \mu_a\mu_b \mathrm{E}_{\mathbf{X}^m}\left\{\phi_a^2(\mathbf{X}_j)\phi_b^2(\mathbf{X}_j)\right\} \\
&\overset{(iv)}{\leq} \sum_{j=1}^m\sum_{a=1}^\infty\sum_{b=1}^\infty \mu_a\mu_b\sqrt{\mathrm{E}_{\mathbf{X}^m}\phi_a^4(\mathbf{X}_j)\cdot\mathrm{E}_{\mathbf{X}^m}\phi_b^4(\mathbf{X}_j)} \\
&\overset{(v)}{\leq} \sum_{j=1}^m\sum_{a=1}^\infty\sum_{b=1}^\infty \mu_a\mu_b\rho_*^4 = m\rho_*^4\left\{\mathrm{tr}(\mathbf{\Sigma}_M)\right\}^2, \qquad (59)
\end{aligned}$$

where (i) follows from the independence between $\mathbf{X}_1, \ldots, \mathbf{X}_m$, (ii) follows from the inequality $\mathrm{Var}(Z) \leq \mathrm{E}(Z^2)$ for any random variable Z, (iii) follows from the monotone convergence theorem, (iv) follows from the Cauchy-Schwarz inequality, and (v) follows from Assumption A.3. Now we set $c_2 = 1/\left[4\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\right]$, and $m_2 = m_2(\xi) \equiv \max\left\{2c_2\frac{\overline{\sigma}_0^2}{n}, 2\rho_*^4/\xi\right\} = \max\left\{\frac{\overline{\sigma}_0^2}{2n\,\mathrm{tr}(\mathbf{\Sigma}_M)}, 2\rho_*^4/\xi\right\}$. Then

for all $m > m_2$, we have that

$$\begin{aligned}
&\mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\max}\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)>m/c_2\right)\leq\mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\max}\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right)+\frac{\overline{\sigma}_0^2}{n}>m/c_2\right)\\
&\overset{(i)}{\leq}\mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\max}\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right)>\frac{m}{2c_2}\right)\leq\mathbb{P}_{\mathbf{X}^m}\left(\operatorname{tr}\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right)>\frac{m}{2c_2}\right)\\
&\leq\mathbb{P}_{\mathbf{X}^m}\left(\left|\operatorname{tr}\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right)-\mathrm{E}\operatorname{tr}\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right)\right|>\frac{m}{2c_2}-\mathrm{E}\operatorname{tr}\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right)\right)\\
&\overset{(ii)}{=}\mathbb{P}_{\mathbf{X}^m}\left(\left|\operatorname{tr}\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right)-\mathrm{E}\operatorname{tr}\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right)\right|>m\operatorname{tr}\left(\mathbf{\Sigma}_M\right)\right)\\
&\overset{(iii)}{\leq}\frac{\operatorname{Var}\left\{\operatorname{tr}\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right)\right\}}{m^2\left\{\operatorname{tr}\left(\mathbf{\Sigma}_M\right)\right\}^2}\overset{(iv)}{\leq}\frac{\rho_*^4}{m}\overset{(v)}{<}\xi,
\end{aligned}\tag{60}$$

where (i) follows from the choice of m_2, (ii) follows from (58) and the choice of c_2, (iii) follows from the Chebyshev's inequality, (iv) follows from (59), and (v) follows from the choice of m_2 again.

We combine (56), (57), and (60) to obtain that for any given $\xi\in(0,1)$, for all $m > m_0 = m_0(\xi)\equiv\max\{m_1(\xi),m_2(\xi)\}$,

$$\mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\min}\left\{\mathcal{F}^\top\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathcal{F}\right\}<\frac{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}{8\operatorname{tr}(\mathbf{\Sigma}_M)}\right)<\frac{\xi}{2}+\frac{\xi}{2}=\xi.$$

Taking the probability of the complement leads to the conclusion. □

LEMMA 3. *(Zhang et al. (2015) Lemma 10) Let $\boldsymbol{\phi}_l=[\phi_l(\mathbf{X}_1),\ldots,\phi_l(\mathbf{X}_m)]^\top$, for $l=1,2,\ldots$. For a given $\zeta\in\mathbb{N}$, let $\mathbf{\Phi}=[\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_\zeta]$. Let $\mathbf{M}=diag\,(\mu_1,\ldots,\mu_\zeta)$ and let $\mathbf{Q}=\left(\mathbf{I}_\zeta+\frac{\overline{\sigma}_0^2}{mn}\mathbf{M}^{-1}\right)^{1/2}$ be the symmetric positive definite square root of $\mathbf{I}_\zeta+\frac{\overline{\sigma}_0^2}{mn}\mathbf{M}^{-1}$. For any given $\delta>0$, define the event*

$$\mathcal{E}=\left\{\left|\left|\left|\mathbf{Q}^{-1}\left(\frac{1}{m}\mathbf{\Phi}^\top\mathbf{\Phi}-\mathbf{I}_\zeta\right)\mathbf{Q}^{-1}\right|\right|\right|\leq\delta\right\}.$$

Then under Assumptions A.1-A.3,

$$\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}^c)\leq\left\{100\rho_*^2\frac{b(m,\zeta,r_*)\gamma(\frac{\overline{\sigma}_0^2}{mn})}{\delta\sqrt{m}}\right\}^{r_*},$$

where $b(m,\zeta,r_*)$ and $\gamma(\cdot)$ are defined in Theorem 1.

Proof of Theorem 3:

We use Theorems 1 and 2 to prove the results for three different types of kernels. The results of Theorems 1 and 2 will be applied to each of k individual design first and then combined.

First note that by the Markov's inequality, Theorem 1 implies that for the ith design ($i = 1, \ldots, k$),

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}^{(M)}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} & \frac{2\overline{\sigma}_0^2}{mn}\gamma_i\left(\frac{\overline{\sigma}_0^2}{mn}\right) \\
& + \inf_{\zeta\in\mathbb{N}}\left[\left\{\frac{3mn}{\overline{\sigma}_0^2}\operatorname{tr}(\boldsymbol{\Sigma}_{M,i})+1\right\}\operatorname{tr}\left(\boldsymbol{\Sigma}^{(\zeta)}_{M,i}\right)+\operatorname{tr}(\boldsymbol{\Sigma}_{M,i})\left\{300\rho_*^2\frac{b(m,\zeta,r_*)\gamma_i(\frac{\overline{\sigma}_0^2}{mn})}{\sqrt{m}}\right\}^{r_*}\right], \qquad (61)
\end{aligned}$$

where $\gamma_i(a) = \sum_{l=1}^{\infty}\mu_{i,l}/(\mu_{i,l}+a)$ for any $a > 0$. And similarly, Theorem 2 implies that for the ith design ($i = 1, \ldots, k$),

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}^{(\boldsymbol{\beta})}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} & \frac{8q\operatorname{tr}(\boldsymbol{\Sigma}_{M,i})}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\left\{8C_{\mathrm{f}}^2\frac{\overline{\sigma}_0^2}{mn}\right. \\
& + \inf_{\zeta\in\mathbb{N}}\left[8C_{\mathrm{f}}^2\frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4\operatorname{tr}\left(\boldsymbol{\Sigma}_{M,i}\right)\operatorname{tr}\left(\boldsymbol{\Sigma}^{(\zeta)}_{M,i}\right)\right. \\
& \left.\left. + C_{\mathrm{f}}^2\operatorname{tr}\left(\boldsymbol{\Sigma}^{(\zeta)}_{M,i}\right) + C_{\mathrm{f}}^2\operatorname{tr}\left(\boldsymbol{\Sigma}_{M,i}\right)\left\{200\rho_*^2\frac{b(m,\zeta,r_*)\gamma_i(\frac{\overline{\sigma}_0^2}{mn})}{\sqrt{m}}\right\}^{r_*}\right]\right\}, \qquad (62)
\end{aligned}$$

The subsequent proofs are based on evaluating the right-hand-sides of (61) and (62). To obtain the upper bound for the maximum IMSE over $i = 1, \ldots, k$, we notice that if for every $i = 1, \ldots, k$, $\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} a(m,n)$ for some sequence of $a(m,n)$ that does not depend on i, then $\max_{i\in\{1,\ldots,k\}}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} a(m,n)$.

Since we only care about the asymptotic orders of $\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right]$ in terms of m and n, in the analysis below, we will use $C_1, C_2, \ldots$ to denote the constants whose values may vary from case to case but do not depend on m and n.

(i) If the ith covariance kernel $\Sigma_{M,i}$ has finite rank l_{*i}, then for the $\inf_{\zeta\in\mathbb{N}}$ terms in both (61) and (62), we let $l_* = \max_{i\in\{1,\ldots,k\}} l_{*i}$ and choose $\zeta = l_*$, which leads to $\operatorname{tr}\left(\boldsymbol{\Sigma}^{(\zeta)}_{M,i}\right) = 0$ for all $i = 1, \ldots, k$. Furthermore, since $r_* \geq 2$ in Assumption A.3, with $\zeta = l_*$,

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

$$\gamma_i\left(\frac{\overline{\sigma}_0^2}{mn}\right) = \sum_{l=1}^{l_*} \frac{\mu_{i,l}}{\mu_{i,l} + \frac{\overline{\sigma}_0^2}{mn}} \leq l_*.$$

(61) and (62) imply that for every $i = 1, \ldots, k$,

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] &\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{2l_*\overline{\sigma}_0^2}{mn} + \mathrm{tr}(\mathbf{\Sigma}_{M,i})\left\{300\rho_*^2\frac{l_*\max(r_*, \log l_*)}{\sqrt{m}}\right\}^{r_*} \\
&\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{C_1}{mn} + \frac{C_2}{m^{r_*/2}}, \\
\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] &\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{8q\,\mathrm{tr}(\mathbf{\Sigma}_{M,i})}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\Bigg[8C_{\mathrm{f}}^2\frac{\overline{\sigma}_0^2}{mn} \\
&\qquad + C_{\mathrm{f}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_{M,i}\right)\left\{200\rho_*^2\frac{l_*\max(r_*, \log l_*)}{\sqrt{m}}\right\}^{r_*}\Bigg] \\
&\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{C_3}{mn} + \frac{C_4}{m^{r_*/2}},
\end{aligned}$$

where C_1, C_2, C_3, C_4 are constants (note that $\max_{i\in\{1,\ldots,k\}} \mathrm{tr}\left(\mathbf{\Sigma}_{M,i}\right)$ is also a finite constant by Assumption A.2). Therefore,

$$\begin{aligned}
\max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] &\leq \max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] + \max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] \\
&\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{C_5}{mn} + \frac{C_6}{m^{r_*/2}} \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \max\left(\frac{1}{mn}, \frac{1}{m^{r_*/2}}\right).
\end{aligned}$$

(ii) If the ith covariance kernel $\Sigma_{M,i}$ satisfies $\mu_{i,l} \leq c_{1i}\exp\left(-c_{2i}l^{\kappa_i/d}\right)$ for all $l \in \mathbb{N}$, then for the $\inf_{\zeta\in\mathbb{N}}$ terms in both (61) and (62), we can choose $\zeta = (mn)^2$. Let $c_{1*} = \max_{i\in\{1,\ldots,k\}} c_{1i}$, $c_{2*} = \min_{i\in\{1,\ldots,k\}} c_{2i}$, and $\kappa_* = \min_{i\in\{1,\ldots,k\}} \kappa_i$. This definition implies that for any $z \geq 1$, $c_{1i}\exp\left(-c_{2i}z^{\kappa_i/d}\right) \leq c_{1*}\exp\left(-c_{2*}z^{\kappa_*/d}\right)$. Then for sufficiently large m,

$$\begin{aligned}
b(m, \zeta, r_*) &= \max\left\{\sqrt{\max(r_*, \log\zeta)}, \frac{\max(r_*, \log\zeta)}{m^{1/2-1/r_*}}\right\} \\
&= \max\left\{\sqrt{\max(r_*, 2\log(mn))}, \frac{\max(r_*, 2\log(mn))}{m^{1/2-1/r_*}}\right\} \leq 2\log(mn), \\
\mathrm{tr}\left(\mathbf{\Sigma}_{M,i}^{(\zeta)}\right) &= \sum_{l=(mn)^2+1}^{\infty} \mu_{i,l} \leq \sum_{l=(mn)^2+1}^{\infty} c_{1i}\exp\left(-c_{2i}l^{\kappa_i/d}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \int_{(mn)^2}^{\infty} c_{1i} \exp\left(-c_{2i} z^{\kappa_i/d}\right) dz \leq \int_{(mn)^2}^{\infty} c_{1*} \exp\left(-c_{2*} z^{\kappa_*/d}\right) dz \\
&\overset{(i)}{\leq} \frac{c_{1*} d}{\kappa_*} \int_{(mn)^{2\kappa_*}}^{\infty} t^{\frac{d}{\kappa_*}-1} \exp\left(-c_{2*} t\right) dt,
\end{aligned}
$$

where in (i), we use the change of variable $t = z^{\kappa_*/d}$. If $\kappa_*/d \geq 1$, then since $t \geq (mn)^{2\kappa_*/d} \geq 1$, we have $t^{\frac{d}{\kappa_*}-1} \leq 1$. If $0 < \kappa_*/d < 1$, then there exists a large $m_0 \in \mathbb{N}$ that depends on only c_{2*}, κ_*, d, such that for all $m \geq m_0$ and $t \geq (mn)^{2\kappa_*/d} \geq m^{2\kappa_*/d}$, we have $t^{\frac{d}{\kappa_*}-1} \leq \exp(c_{2*}t/2)$. Therefore, in all cases,

$$
\operatorname{tr}\left(\mathbf{\Sigma}_{M,i}^{(\zeta)}\right) \leq \frac{c_{1*} d}{\kappa} \int_{(mn)^{2\kappa_*/d}}^{\infty} \exp\left(-c_{2*} t/2\right) dt = \frac{2c_{1*} d}{c_{2*}\kappa_*} \exp\left\{-c_{2*}(mn)^{2\kappa_*/d}/2\right\}. \tag{63}
$$

Let $l_1 = \left\{\frac{2}{c_{2*}} \log(mn)\right\}^{d/\kappa_*}$. For sufficiently large m and every $i = 1, \ldots, k$, $\gamma_i\left(\frac{\overline{\sigma}_0^2}{mn}\right)$ can be bounded by

$$
\begin{aligned}
\gamma_i\left(\frac{\overline{\sigma}_0^2}{mn}\right) &= \sum_{l=1}^{\infty} \frac{\mu_{i,l}}{\mu_{i,l} + \frac{\overline{\sigma}_0^2}{mn}} = \sum_{l=1}^{\lfloor l_1 \rfloor + 1} \frac{\mu_{i,l}}{\mu_{i,l} + \frac{\overline{\sigma}_0^2}{mn}} + \sum_{l=\lfloor l_1 \rfloor + 2}^{\infty} \frac{\mu_{i,l}}{\mu_{i,l} + \frac{\overline{\sigma}_0^2}{mn}} \\
&\leq l_1 + 1 + \frac{mn}{\overline{\sigma}_0^2} \sum_{l=\lfloor l_1 \rfloor + 1}^{\infty} c_{1i} \exp\left(-c_{2i} l^{\kappa_i/d}\right) \\
&\leq l_1 + 1 + \frac{mn}{\overline{\sigma}_0^2} \int_{l_1}^{\infty} c_{1*} \exp\left(-c_{2*} z^{\kappa_*/d}\right) dz \\
&= l_1 + 1 + \frac{mn c_{1*} d}{\kappa_* \overline{\sigma}_0^2} \int_{l_1^{\kappa_*/d}}^{\infty} t^{\frac{d}{\kappa_*}-1} \exp\left(-c_{2*} t\right) dt \\
&\leq l_1 + 1 + \frac{mn c_{1*} d}{\kappa_* \overline{\sigma}_0^2} \int_{l_1^{\kappa_*/d}}^{\infty} \exp\left(-c_{2*} t/2\right) dt \\
&= l_1 + 1 + \frac{mn c_{1*} d}{2c_{2*}\kappa_* \overline{\sigma}_0^2} \exp\left(-c_{2*} l_1^{\kappa_*/d}/2\right) \\
&= l_1 + 1 + \frac{c_{1*} d}{2c_{2*}\kappa \overline{\sigma}_0^2} \leq C_1 \log^{\frac{d}{\kappa_*}}(mn),
\end{aligned}
$$

for some constant $C_1 > 0$ that does not depend on i. Therefore, (61) and (62) imply that for every $i = 1, \ldots, k$,

$$
\begin{aligned}
\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} & \frac{2C_1 \overline{\sigma}_0^2 \log^{\frac{d}{\kappa_*}}(mn)}{mn} \\
& + \left\{\frac{3mn}{\overline{\sigma}_0^2} \operatorname{tr}(\mathbf{\Sigma}_{M,i}) + 1\right\} \frac{2c_{1*} d}{c_{2*}\kappa_*} \exp\left\{-\frac{c_{2*}}{2}(mn)^{2\kappa_*/d}\right\}
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \operatorname{tr}(\boldsymbol{\Sigma}_{M,i}) \left\{ 300\rho_*^2 \frac{2\log(mn)\cdot C_1 \log^{\frac{d}{\kappa_*}}(mn)}{\sqrt{m}} \right\}^{r_*} \\
&\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{C_2 \log^{\frac{d}{\kappa_*}}(mn)}{mn} + C_3 mn \exp\left\{-c_{2*}(mn)^{2\kappa_*/d}/2\right\} + C_4 \frac{\log^{\frac{r_*(\kappa_*+d)}{\kappa_*}}(mn)}{m^{r_*/2}}, \\
\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}^{(\boldsymbol{\beta})}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] &\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{8q \operatorname{tr}(\boldsymbol{\Sigma}_{M,i})}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)} \Bigg[8C_{\mathrm{f}}^2 \frac{\overline{\sigma}_0^2}{mn} \\
&\quad + 8C_{\mathrm{f}}^2 \frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4} \rho_*^4 \operatorname{tr}\left(\boldsymbol{\Sigma}_{M,i}\right) \frac{c_{1*}}{c_{2*}} \exp\left\{-c_{2*}(mn)^2\right\} \\
&\quad + C_{\mathrm{f}}^2 \frac{2c_{1*}d}{c_{2*}\kappa_*} \exp\left\{-c_{2*}(mn)^{2\kappa_*/d}/2\right\} \\
&\quad + C_{\mathrm{f}}^2 \operatorname{tr}\left(\boldsymbol{\Sigma}_{M,i}\right) \left\{ 200\rho_*^2 \frac{2\log(mn)\cdot C_1 \log^{\frac{d}{\kappa_*}}(mn)}{\sqrt{m}} \right\}^{r_*} \Bigg] \\
&\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{C_5}{mn} + C_6 mn \exp\left\{-c_{2*}(mn)^{2\kappa_*/d}/2\right\} + C_7 \frac{\log^{\frac{r_*(\kappa_*+d)}{\kappa_*}}(mn)}{m^{r_*/2}},
\end{aligned}
$$

for some positive constants $C_2, C_3, C_4, C_5, C_6, C_7$. Therefore,

$$
\begin{aligned}
&\max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \\
&\le \max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}^{(M)}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] + \max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}^{(\boldsymbol{\beta})}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \\
&\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{C_2 \log^{\frac{d}{\kappa_*}}(mn)}{mn} + C_3 mn \exp\left\{-c_{2*}(mn)^{2\kappa_*/d}/2\right\} + C_4 \frac{\log^{\frac{r_*(\kappa_*+d)}{\kappa_*}}(mn)}{m^{r_*/2}} \\
&\quad + \frac{C_5}{mn} + C_6 mn \exp\left\{-c_{2*}(mn)^{2\kappa_*/d}/2\right\} + C_7 \frac{\log^{\frac{r_*(\kappa_*+d)}{\kappa_*}}(mn)}{m^{r_*/2}} \\
&\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \max\left\{ \frac{\log^{\frac{d}{\kappa_*}}(mn)}{mn}, \frac{\log^{\frac{r_*(\kappa_*+d)}{\kappa_*}}(mn)}{m^{r_*/2}} \right\},
\end{aligned}
$$

where the last inequality follows because $\frac{mn\exp\{-c_{2*}(mn)^{2\kappa_*/d}/2\}}{\log^{\frac{d}{\kappa_*}}(mn)/(mn)} \to 0$ and $\frac{1/(mn)}{\log^{\frac{d}{\kappa_*}}(mn)/(mn)} \to 0$ as $mn \to \infty$.

(iii) If the ith covariance kernel $\Sigma_{M,i}$ satisfies $\mu_{i,l} \le c_i l^{-2\nu_i/d-1}$ for all $l \in \mathbb{N}$, then $\mu_{i,l} \le c_* l^{-2\nu_*/d-1}$ for all $l \in \mathbb{N}$ and all $i = 1, \ldots, k$, where $c_* = \max_{i\in\{1,\ldots,k\}} c_i$ and $\nu_* = \min_{i\in\{1,\ldots,k\}} \nu_i$. For the $\inf_{\zeta\in\mathbb{N}}$ terms in both (61) and (62), we choose $\zeta = \lfloor (mn)^{3d/(2\nu_*)} \rfloor$. Then for sufficiently large m,

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

$$= \max\left\{\sqrt{\max\left(r_*, \frac{3d}{2\nu_*}\log(mn)\right)}, \frac{\max\left(r_*, \frac{3d}{2\nu_*}\log(mn)\right)}{m^{1/2-1/r_*}}\right\} \leq \frac{3d}{2\nu_*}\log(mn),$$

$$\operatorname{tr}\left(\boldsymbol{\Sigma}_{M,i}^{(\zeta)}\right) = \sum_{l=\zeta+1}^{\infty}\mu_{i,l} \leq \sum_{l=\zeta+1}^{\infty} c_* l^{-2\nu_*/d-1} \leq \int_{\zeta}^{\infty} c_* z^{-2\nu_*/d-1}dz = \frac{c_* d}{2\nu_*}\zeta^{-2\nu_*/d} \leq \frac{c_* d}{2\nu_*}(mn)^{-3},$$

$$\begin{aligned}\gamma_i\left(\frac{\overline{\sigma}_0^2}{mn}\right) &= \sum_{l=1}^{\infty}\frac{1}{1+\frac{\overline{\sigma}_0^2}{mn\mu_{i,l}}} = \sum_{l=1}^{\infty}\frac{1}{1+\frac{\overline{\sigma}_0^2 l^{2\nu_*/d+1}}{c_* mn}} \leq (mn)^{d/(2\nu_*+d)} + 1 + \sum_{l=\lfloor (mn)^{d/(2\nu_*+d)}\rfloor+2}\frac{c_* mn}{\overline{\sigma}_0^2 l^{2\nu_*/d+1}} \\ &\leq (mn)^{d/(2\nu_*+d)} + 1 + \frac{c_* mn}{\overline{\sigma}_0^2}\int_{(mn)^{d/(2\nu_*+d)}}^{\infty}\frac{1}{z^{2\nu_*/d+1}}dz \\ &= (mn)^{d/(2\nu_*+d)} + 1 + \frac{c_* d mn}{2\nu_*\overline{\sigma}_0^2}(mn)^{-\frac{2\nu_*}{2\nu_*+d}} \leq C_1(mn)^{d/(2\nu_*+d)},\end{aligned}$$

for some large constant $C_1 > 0$ that does not depend on i. Therefore, (61) and (62) imply that for every $i = 1, \ldots, k$,

$$\begin{aligned}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] &\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{2C_1\overline{\sigma}_0^2(mn)^{d/(2\nu_*+d)}}{mn} + \left\{\frac{3mn}{\overline{\sigma}_0^2}\operatorname{tr}(\boldsymbol{\Sigma}_{M,i}) + 1\right\}\frac{c_* d}{2\nu_*}(mn)^{-3} \\ &\quad + \operatorname{tr}(\boldsymbol{\Sigma}_{M,i})\left\{300\rho_*^2\frac{\frac{3d}{2\nu_*}\log(mn)\cdot C_1(mn)^{d/(2\nu_*+d)}}{\sqrt{m}}\right\}^{r_*} \\ &\lesssim_{\mathbb{P}_{\mathbf{X}^m}} C_2(mn)^{-\frac{2\nu_*}{2\nu_*+d}} + C_3(mn)^{-2} + C_4\frac{n^{\frac{dr_*}{2\nu_*+d}}\log^{r_*}(mn)}{m^{\frac{r_*(2\nu_*-d)}{2(2\nu_*+d)}}}, \\ \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] &\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{8q\operatorname{tr}(\boldsymbol{\Sigma}_{M,i})}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\Bigg[8C_{\mathrm{f}}^2\frac{\overline{\sigma}_0^2}{mn} \\ &\quad + 8C_{\mathrm{f}}^2\frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4\operatorname{tr}\left(\boldsymbol{\Sigma}_{M,i}\right)\frac{c_* d}{2\nu_*}(mn)^{-3} + C_{\mathrm{f}}^2\frac{c_* d}{2\nu_*}(mn)^{-3} \\ &\quad + C_{\mathrm{f}}^2\operatorname{tr}\left(\boldsymbol{\Sigma}_{M,i}\right)\left\{200\rho_*^2\frac{\frac{3d}{2\nu_*}\log(mn)\cdot C_1(mn)^{d/(2\nu_*+d)}}{\sqrt{m}}\right\}^{r_*}\Bigg] \\ &\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{C_5}{mn} + C_6(mn)^{-2} + C_7\frac{n^{\frac{dr_*}{2\nu_*+d}}\log^{r_*}(mn)}{m^{\frac{r_*(2\nu_*-d)}{2\nu_*+d}}},\end{aligned}$$

for some positive constants $C_2, C_3, C_4, C_5, C_6, C_7$. Therefore,

$$\begin{aligned}\max_{i\in\{1,\ldots,k\}}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] &\leq \max_{i\in\{1,\ldots,k\}}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] + \max_{i\in\{1,\ldots,k\}}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] \\ &\lesssim_{\mathbb{P}_{\mathbf{X}^m}} C_2(mn)^{-\frac{2\nu_*}{2\nu_*+d}} + C_3(mn)^{-2} + C_4\frac{n^{\frac{dr_*}{2\nu_*+d}}\log^{r_*}(mn)}{m^{\frac{r_*(2\nu_*-d)}{2\nu_*+d}}} \\ &\quad + \frac{C_5}{mn} + C_6(mn)^{-2} + C_7\frac{n^{\frac{dr_*}{2\nu+d}}\log^{r_*}(mn)}{m^{\frac{r_*(2\nu_*-d)}{2\nu_*+d}}}\end{aligned}$$

$$\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \max\left\{ \frac{1}{(mn)^{\frac{2\nu_*}{2\nu_*+d}}}, \frac{n^{\frac{dr_*}{2\nu_*+d}} \log^{r_*}(mn)}{m^{\frac{r_*(2\nu_*-d)}{2\nu_*+d}}} \right\},$$

where the last inequality follows because $\frac{1/(mn)}{(mn)^{-\frac{2\nu_*}{2\nu_*+d}}} = (mn)^{-d/(2\nu_*+d)} \to 0$ and $\frac{1/(mn)^2}{(mn)^{-\frac{2\nu_*}{2\nu_*+d}}} = (mn)^{-d/(2\nu_*+d)-1} \to 0$ as $mn \to \infty$. □

Proof of Theorem 4:

For $i = 1, \ldots, k$, let $\widetilde{\mathbf{X}}_{i,0} = \left(\sqrt{b_i}, \mathbf{X}_0^\top\right)^\top$ be the $\mathbb{R}^{d+1}$ random vector version of $\widetilde{\mathbf{x}}_{i,0}$ with $\mathbf{X}_0$ following the distribution $\mathbb{P}_{\mathbf{X}}$. For the covariance kernel $\mathbf{\Sigma}_{M,i}(\mathbf{x}, \mathbf{x}') = a_i \left(\mathbf{x}^\top \mathbf{x}' + b_i\right)$, using the definition of $\widetilde{\mathbf{x}}_{i,0}$ and $\mathbf{Z}_i$ in Theorem 4, we have that

$$\begin{aligned}
\mathbf{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{X}_0) &= \left(\mathbf{\Sigma}_{M,i}(\mathbf{X}_1, \mathbf{X}_0), \ldots, \mathbf{\Sigma}_{M,i}(\mathbf{X}_m, \mathbf{X}_0)\right)^\top \\
&= \Big(a_i(\mathbf{X}_1^\top \mathbf{X}_0 + b_i), \ldots, a_i(\mathbf{X}_m^\top \mathbf{X}_0 + b_i)\Big)^\top = a_i \mathbf{Z}_i \widetilde{\mathbf{X}}_0, \\
\mathbf{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{X}^m) &= \begin{pmatrix} \mathbf{\Sigma}_{M,i}(\mathbf{X}_1, \mathbf{X}_1) & \ldots & \mathbf{\Sigma}_{M,i}(\mathbf{X}_1, \mathbf{X}_m) \\ & \ldots & \\ \mathbf{\Sigma}_{M,i}(\mathbf{X}_m, \mathbf{X}_1) & \ldots & \mathbf{\Sigma}_{M,i}(\mathbf{X}_m, \mathbf{X}_m) \end{pmatrix} \\
&= \begin{pmatrix} a_i\big(\mathbf{X}_1^\top \mathbf{X}_1 + b_i\big) & \ldots & a_i\big(\mathbf{X}_1^\top \mathbf{X}_m + b_i\big) \\ & \ldots & \\ a_i\big(\mathbf{X}_m^\top \mathbf{X}_1 + b_i\big) & \ldots & a_i\big(\mathbf{X}_m^\top \mathbf{X}_m + b_i\big) \end{pmatrix} \\
&= a_i \mathbf{Z}_i \mathbf{Z}_i^\top.
\end{aligned}$$

Therefore, we plug in $\mathbf{f}_i(\mathbf{X}) \equiv 0$ to (2) of the manuscript and obtain that

$$\begin{aligned}
\widehat{y}_i(\mathbf{X}_0) &= \mathbf{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{X}_0)^\top \left[\mathbf{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{X}^m) + \frac{\sigma^2}{n}\mathbf{I}_m\right]^{-1} \overline{Y}_i \\
&= a_i \widetilde{\mathbf{X}}_{i,0}^\top \mathbf{Z}_i^\top \left(a_i \mathbf{Z}_i \mathbf{Z}_i^\top + \frac{\sigma^2}{n}\mathbf{I}_m\right)^{-1} \overline{Y}_i.
\end{aligned}$$

Similarly we obtain from (3) of the manuscript that

$$\begin{aligned}
\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0) &= \mathbf{\Sigma}_{M,i}(\mathbf{X}_0, \mathbf{X}_0) - \mathbf{\Sigma}_{M,i}^\top(\mathbf{X}^m, \mathbf{X}_0) \left[\mathbf{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{X}^m) + \frac{\sigma^2}{n}\mathbf{I}_m\right]^{-1} \mathbf{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{X}_0) \\
&= a_i \widetilde{\mathbf{X}}_{i,0}^\top \widetilde{\mathbf{X}}_{i,0} - a_i \widetilde{\mathbf{X}}_{i,0}^\top \mathbf{Z}_i^\top \left(a_i \mathbf{Z}_i \mathbf{Z}_i^\top + \frac{\sigma^2}{n}\mathbf{I}_m\right)^{-1} a_i \mathbf{Z}_i \widetilde{\mathbf{X}}_{i,0}
\end{aligned}$$

$$= a_i \widetilde{\mathbf{X}}_{i,0}^\top \left[\mathbf{I}_{d+1} - \mathbf{Z}_i^\top \left(\mathbf{Z}_i \mathbf{Z}_i^\top + \frac{\sigma^2}{a_i n} \mathbf{I}_m \right)^{-1} \mathbf{Z}_i \right] \widetilde{\mathbf{X}}_{i,0}$$
$$\overset{(i)}{=} a_i \widetilde{\mathbf{X}}_{i,0}^\top \left(\mathbf{I}_{d+1} + \frac{a_i n}{\sigma^2} \mathbf{Z}_i^\top \mathbf{Z}_i \right)^{-1} \widetilde{\mathbf{X}}_{i,0},$$

where we have applied the Woodbury matrix inversion formula (Rasmussen and Williams 2006, Appendix A.3) in the step (i). This has proved (12) of the main text.

Now we turn to (13) of the manuscript. Note that

$$\begin{aligned} \mathrm{E}_{\mathbf{X}_0} \left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0) \right] &= \mathrm{E}_{\mathbf{X}_0} \left[a_i \widetilde{\mathbf{X}}_{i,0}^\top \left(\mathbf{I}_{d+1} + \frac{a_i n}{\sigma^2} \mathbf{Z}_i^\top \mathbf{Z}_i \right)^{-1} \widetilde{\mathbf{X}}_{i,0} \right] \\ &= \mathrm{tr} \left\{ \left(\mathbf{I}_{d+1} + \frac{a_i n}{\sigma^2} \mathbf{Z}_i^\top \mathbf{Z}_i \right)^{-1} \cdot a_i \mathrm{E}_{\mathbf{X}_0} \left(\widetilde{\mathbf{X}}_{i,0} \widetilde{\mathbf{X}}_{i,0}^\top \right) \right\}. \end{aligned} \tag{64}$$

According to the definition of $\mathbf{Z}_i$ and the fact that $\mathbf{X}^m, \mathbf{X}_0$ are i.i.d. draws from $\mathbb{P}_{\mathbf{X}}$, by the strong law of large numbers, as $m \to \infty$, almost surely in $\mathbb{P}_{\mathbf{X}^m}$,

$$\begin{aligned} \frac{1}{m} \mathbf{Z}_i^\top \mathbf{Z}_i &= \begin{pmatrix} b_i & \frac{\sqrt{b_i}}{m} \sum_{j=1}^m \mathbf{X}_j^\top \\ \frac{\sqrt{b_i}}{m} \sum_{j=1}^m \mathbf{X}_j & \frac{1}{m} \sum_{j=1}^m \mathbf{X}_j \mathbf{X}_j^\top \end{pmatrix} \\ &\to \begin{pmatrix} b_i & \sqrt{b_i} \mathrm{E}_{\mathbf{X}_1}(\mathbf{X}_1^\top) \\ \sqrt{b_i} \mathrm{E}_{\mathbf{X}_1}(\mathbf{X}_1) & \mathrm{E}_{\mathbf{X}_1}(\mathbf{X}_1 \mathbf{X}_1^\top) \end{pmatrix} = \mathrm{E}_{\mathbf{X}_0} \left(\widetilde{\mathbf{X}}_{i,0} \widetilde{\mathbf{X}}_{i,0}^\top \right). \end{aligned} \tag{65}$$

Therefore, (64) and (65) together imply that for each $i = 1, \ldots, k$, as $m \to \infty$, almost surely in $\mathbb{P}_{\mathbf{X}^m}$,

$$\begin{aligned} mn \cdot \mathrm{E}_{\mathbf{X}_0} \left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0) \right] &= \mathrm{tr} \left\{ \left(\frac{1}{mn} \mathbf{I}_{d+1} + \frac{a_i}{\sigma^2} \cdot \frac{1}{m} \mathbf{Z}_i^\top \mathbf{Z}_i \right)^{-1} \cdot a_i \mathrm{E}_{\mathbf{X}_0} \left(\widetilde{\mathbf{X}}_{i,0} \widetilde{\mathbf{X}}_{i,0}^\top \right) \right\} \\ &\to \mathrm{tr} \left\{ \left[\frac{a_i}{\sigma^2} \cdot \mathrm{E}_{\mathbf{X}_0} \left(\widetilde{\mathbf{X}}_{i,0} \widetilde{\mathbf{X}}_{i,0}^\top \right) \right]^{-1} \cdot a_i \mathrm{E}_{\mathbf{X}_0} \left(\widetilde{\mathbf{X}}_{i,0} \widetilde{\mathbf{X}}_{i,0}^\top \right) \right\} \\ &= \mathrm{tr} \left(\sigma^2 \mathbf{I}_{d+1} \right) = (d+1)\sigma^2. \end{aligned} \tag{66}$$

Define the event $\mathcal{A}_i = \{\text{The convergence in (66) happens as } n \to \infty\}$ for $i = 1, \ldots, k$. Then the almost sure convergence in (66) implies $\mathbb{P}_{\mathbf{X}^m}(\mathcal{A}_i) = 1$ for every $i = 1, \ldots, k$. This further implies

that

$$\mathbb{P}_{\mathbf{X}^m}\left(\cap_{i=1}^k \mathcal{A}_i\right) = 1 - \mathbb{P}_{\mathbf{X}^m}\left(\cup_{i=1}^k \mathcal{A}_i^c\right) \geq 1 - \sum_{i=1}^k \mathbb{P}_{\mathbf{X}^m}\left(\mathcal{A}_i^c\right) = 1 - \sum_{i=1}^k 0 = 1,$$

which implies that $\mathbb{P}_{\mathbf{X}^m}\left(\cap_{i=1}^k \mathcal{A}_i\right) = 1$, i.e. the convergence in (66) happens jointly over $i = 1, \ldots, k$ as $m \to \infty$ almost surely in $\mathbb{P}_{\mathbf{X}^m}$. Therefore, on the event $\cap_{i=1}^k \mathcal{A}_i$, (66) implies that $mn \cdot \max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \to (d+1)\sigma^2$ as $m \to \infty$. This has proved (13) of the main text. □

Proof of Theorem 5:

We first derive a natural bound for $\mathrm{PFS}(\mathbf{x}_0)$. For any $\mathbf{x}_0 \in \mathcal{X}$, any $i, i' \in \{1, \ldots, k\}$, define the random variable $W_{i,i'}(\mathbf{x}_0) = [\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)] - [\widehat{y}_{i'}(\mathbf{x}_0) - y_{i'}(\mathbf{x}_0)]$ (so $W_{i,i}(\mathbf{x}_0) = 0$). According to our definition in (4) of the manuscript, $y^\circ(\mathbf{x}_0) = y_{i^\circ(\mathbf{x}_0)}(\mathbf{x}_0)$. Therefore,

$$\begin{aligned}
&\mathrm{PFS}(\mathbf{x}_0) = \mathbb{P}_\epsilon\left(y_{\widehat{i}^\circ(\mathbf{x}_0)}(\mathbf{x}_0) - y^\circ(\mathbf{x}_0) \geq \delta_0\right) \\
&= \mathbb{P}_\epsilon\left\{\left[\widehat{y}_{i^\circ(\mathbf{x}_0)}(\mathbf{x}_0) - y_{i^\circ(\mathbf{x}_0)}(\mathbf{x}_0)\right] - \left[\widehat{y}^\circ(\mathbf{x}_0) - y_{\widehat{i}^\circ(\mathbf{x}_0)}(\mathbf{x}_0)\right] \geq \delta_0 + \left[\widehat{y}_{i^\circ(\mathbf{x}_0)}(\mathbf{x}_0) - \widehat{y}^\circ(\mathbf{x}_0)\right]\right\} \\
&\overset{(i)}{\leq} \mathbb{P}_\epsilon\left(W_{i^\circ(\mathbf{x}_0),\widehat{i}^\circ(\mathbf{x}_0)}(\mathbf{x}_0) \geq \delta_0\right) \leq \mathbb{P}_\epsilon\left(\max_{1\leq i,i'\leq k} W_{i,i'}(\mathbf{x}_0) \geq \delta_0\right) \\
&\leq \sum_{1\leq i,i'\leq k} \mathbb{P}_\epsilon\left(W_{i,i'}(\mathbf{x}_0) \geq \delta_0\right) = \sum_{1\leq i<i'\leq k} \mathbb{P}_\epsilon\left(\left|W_{i,i'}(\mathbf{x}_0)\right| \geq \delta_0\right), \qquad (67)
\end{aligned}$$

where $\widehat{y}^\circ(\mathbf{x}_0) = \min_{i\in\{1,2,\ldots,k\}} \widehat{y}_i(\mathbf{x}_0)$. Inequality (i) holds because $\widehat{i}^\circ(\mathbf{x}_0) = \arg\min_{i\in\{1,\ldots,k\}} \widehat{y}_i(\mathbf{x}_0)$ and hence $\widehat{y}_{i^\circ(\mathbf{x}_0)}(\mathbf{x}_0) \geq \widehat{y}^\circ(\mathbf{x}_0)$. Now since $y_i(\mathbf{x}) = \mathbf{f}_i(\mathbf{x})^\top \boldsymbol{\beta}_i + M_i(\mathbf{x})$ in (1) of the manuscript includes $M_i(\mathbf{x})$, it is clear that $W_{i,i'}$ depends on $M_i(\cdot)$, $M_{i'}(\cdot)$, $\mathbf{X}^m$ and $\mathbf{x}_0$, which are all random. We first remove the randomness from $M_i(\mathbf{x})$'s ($i = 1, \ldots, k$) by taking the expectation of $\mathrm{PFS}(\mathbf{x}_0)$ with respect to the joint Gaussian measure $\mathbb{P}_M$ induced by the k independent Gaussian processes with mean zero and covariance function $\boldsymbol{\Sigma}_{M,i}(\cdot,\cdot)$ for $i = 1, \ldots, k$. Then from (67) we can obtain that

$$\begin{aligned}
&\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{x}_0)\right] \leq \mathrm{E}_M\Big[\sum_{1\leq i<i'\leq k} \mathbb{P}_\epsilon\left(\left|W_{i,i'}(\mathbf{x}_0)\right| \geq \delta_0\right)\Big] \\
&= \sum_{1\leq i<i'\leq k} \mathrm{E}_M\mathrm{E}_\epsilon\left[\mathbb{1}\left\{\left|W_{i,i'}(\mathbf{x}_0)\right| \geq \delta_0\right\}\right] = \sum_{1\leq i<i'\leq k} \mathbb{P}_{M,\epsilon}\left(\left|W_{i,i'}(\mathbf{x}_0)\right| \geq \delta_0\right), \qquad (68)
\end{aligned}$$

where $\mathbb{P}_{M,\epsilon}$ denotes the joint (independent) probability measure of all $M_i(\cdot)$'s from Gaussian processes and the error terms. The inequality of (68) allows us to directly consider all randomness in $W_{i,i'}$'s given fixed $\mathbf{X}^m$ and $\mathbf{X}_0$.

Let $M_i(\mathbf{X}^m) = (M_i(\mathbf{X}_1), \ldots, M_i(\mathbf{X}_m))^\top$ and $\overline{\epsilon}(\mathbf{X}^m) = (\overline{\epsilon}_i(\mathbf{X}_1), \ldots, \overline{\epsilon}_i(\mathbf{X}_m))^\top$, for $i = 1, \ldots, k$. Under the joint measure $\mathbb{P}_{M,\epsilon}$ (with expectation $\mathrm{E}_{M,\epsilon}$), based on (2) of the manuscript, we have that for any given $\mathbf{X}^m$ and $\mathbf{x}_0 \in \mathcal{X}$,

$$
\begin{aligned}
&\mathrm{E}_{M,\epsilon}(\overline{\mathbf{Y}}_i) = \mathrm{E}_{M,\epsilon}\left[\mathcal{F}_i\boldsymbol{\beta}_i + M_i(\mathbf{X}^m) + \overline{\epsilon}(\mathbf{X}^m)\right] = \mathcal{F}_i\boldsymbol{\beta}_i, \\
&\mathrm{E}_{M,\epsilon}\left[\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)\right] \\
&= \mathrm{E}_{M,\epsilon}\left[\mathbf{f}_i(\mathbf{x}_0)^\top\widehat{\boldsymbol{\beta}}_i + \boldsymbol{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{x}_0)^\top\boldsymbol{\Sigma}_{y,i}^{-1}\left(\overline{\mathbf{Y}}_i - \mathcal{F}_i\widehat{\boldsymbol{\beta}}_i\right) - \mathbf{f}_i(\mathbf{x}_0)^\top\boldsymbol{\beta}_i - M_i(\mathbf{x}_0)\right] \\
&= \mathbf{f}_i(\mathbf{x}_0)^\top\left(\mathcal{F}_i^\top\boldsymbol{\Sigma}_{y,i}^{-1}\mathcal{F}_i\right)^{-1}\mathcal{F}_i^\top\boldsymbol{\Sigma}_{y,i}^{-1}\mathcal{F}_i\boldsymbol{\beta}_i + \boldsymbol{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{x}_0)^\top\boldsymbol{\Sigma}_{y,i}^{-1}\mathcal{F}_i\boldsymbol{\beta}_i \\
&\quad - \boldsymbol{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{x}_0)^\top\boldsymbol{\Sigma}_{y,i}^{-1}\mathcal{F}_i\left(\mathcal{F}_i^\top\boldsymbol{\Sigma}_{y,i}^{-1}\mathcal{F}_i\right)^{-1}\mathcal{F}_i^\top\boldsymbol{\Sigma}_{y,i}^{-1}\mathcal{F}_i\boldsymbol{\beta}_i - \mathbf{f}_i(\mathbf{x}_0)^\top\boldsymbol{\beta}_i \\
&= 0.
\end{aligned}
$$

Hence $\mathrm{E}_{M,\epsilon}(W_{i,i'}) = 0$ for all $1 \le i < i' \le k$. Furthermore, the variance of $\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)$ is $\mathrm{Var}_{M,\epsilon}[\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)] = \mathrm{E}_{M,\epsilon}[\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)]^2$, which is the MSE of $\widehat{y}_i(\mathbf{x}_0)$ and hence is equal to $\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)$ given in (3) of the manuscript. For $W_{i,i'}$ $(1 \le i < i' \le k)$, the independence between different $M_i(\cdot)$'s and errors implies that

$$
\begin{aligned}
\mathrm{Var}_{M,\epsilon}(W_{i,i'}) &= \mathrm{Var}_{M,\epsilon}[\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)] + \mathrm{Var}_{M,\epsilon}[\widehat{y}_{i'}(\mathbf{x}_0) - y_{i'}(\mathbf{x}_0)] \\
&= \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) + \mathrm{MSE}_{i',\mathrm{opt}}(\mathbf{x}_0).
\end{aligned}
$$

From (68), we apply the Markov's inequality and obtain that

$$
\begin{aligned}
&\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right] \le \sum_{1\le i<i'\le k}\mathrm{E}_{\mathbf{X}_0}\left[\mathbb{P}_{M,\epsilon}\left(\left|W_{i,i'}(\mathbf{X}_0)\right| \ge \delta_0\right)\right] \\
&\le \sum_{1\le i<i'\le k}\mathrm{E}_{\mathbf{X}_0}\left[\frac{\mathrm{E}_{M,\epsilon}\left|W_{i,i'}(\mathbf{X}_0)\right|^2}{\delta_0^2}\right] = \sum_{1\le i<i'\le k}\mathrm{E}_{\mathbf{X}_0}\left[\frac{\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0) + \mathrm{MSE}_{i',\mathrm{opt}}(\mathbf{X}_0)}{\delta_0^2}\right] \\
&\le \frac{k(k-1)}{\delta_0^2}\max_{i\in\{1,\ldots,k\}}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \qquad (69)
\end{aligned}
$$

We now prove Part (i) of Theorem 5. Under Assumptions A.1-A.4, Part (i) of Theorem 3 says that $\max_{i\in\{1,2,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} R(m,n)$ as $m\to\infty$. This is to say that for any $\xi\in(0,1/2)$, there exist $m_0\geq 1$ and $c_1>0$ that depends on ξ, such that for all $m\geq m_0$,

$$\mathbb{P}_{\mathbf{X}^m}\left(\max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \leq c_1 R(m,n)\right) \geq 1-\xi. \tag{70}$$

(69) and (70) together implies that

$$\begin{aligned}&\mathbb{P}_{\mathbf{X}^m}\left(\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right] \leq \frac{c_1 k(k-1)}{\delta_0^2} R(m,n)\right)\\ \geq\ &\mathbb{P}_{\mathbf{X}^m}\left(\max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \leq c_1 R(m,n)\right) \geq 1-\xi.\end{aligned} \tag{71}$$

This is to say that for any $\xi\in(0,1/2)$, there exist $m_0\geq 1$ and $c_1>0$ that depends on ξ, such that for all $m\geq m_0$, the relation (71) holds. In other words, we have proved that $\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} R(m,n)$.

Next we prove Part (ii) of Theorem 5 with the additional Assumptions A.5 and A.6. The simulation errors $\epsilon_{il}(\mathbf{x})$'s are all normally distributed by Assumption A.5. Also $M_i(\mathbf{x})$'s are normally distributed due to the Gaussian process model. Hence we know that for given $\mathbf{X}^m$ and $\mathbf{x}_0$, $\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)$ as a linear function of $\overline{\mathbf{Y}}_i$ and $y_i(\mathbf{x}_0)$, is normally distributed as $N(0, \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0))$. The independence of $\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)$ and $\widehat{y}_{i'}(\mathbf{x}_0) - y_{i'}(\mathbf{x}_0)$ for $1\leq i<i'\leq k$ further implies that for given $\mathbf{X}^m$ and $\mathbf{x}_0$, $\mathrm{Var}_{M,\epsilon}(W_{i,i'}) = \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) + \mathrm{MSE}_{i',\mathrm{opt}}(\mathbf{x}_0)$ and thus $W_{i,i'} \sim N(0, \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) + \mathrm{MSE}_{i',\mathrm{opt}}(\mathbf{x}_0))$. We can apply the tail probability bound of normal distributions ($\mathbb{P}(|Z|>z)\leq \exp(-z^2/2)$ if $Z\sim N(0,1)$ and $z>0$) and obtain that

$$\mathbb{P}_{M,\epsilon}\left(\left|W_{i,i'}(\mathbf{x}_0)\right| \geq \delta_0\right) \leq \exp\left(-\frac{\delta_0^2}{2\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) + \mathrm{MSE}_{i',\mathrm{opt}}(\mathbf{x}_0)\right]}\right). \tag{72}$$

(68) and (72) together imply that

$$\begin{aligned}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{x}_0)\right] &\leq \sum_{1\leq i<i'\leq k} \exp\left(-\frac{\delta_0^2}{2\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) + \mathrm{MSE}_{i',\mathrm{opt}}(\mathbf{x}_0)\right]}\right)\\ &\leq \frac{k(k-1)}{2}\exp\left(-\frac{\delta_0^2}{4\max_{i\in\{1,\ldots,k\}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)}\right).\end{aligned} \tag{73}$$

For abbreviation, we let $V = \max_{i\in\{1,\ldots,k\}} \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)$. Assumption A.6 says that 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 $t > 0$, we have $\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_4) \geq 1 - \xi$, where $\mathcal{E}_4$ is defined as

$$\mathcal{E}_4 = \left\{ \mathbb{P}_{\mathbf{X}_0} \left(V \geq tR(m,n) \right) \leq w_1 \exp\left(-w_2 t\right) \right\}.$$

Conditional on the event $\mathcal{E}_4$, from (73), we can derive that

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right] \leq \mathrm{E}_{\mathbf{X}_0}\left[\frac{k(k-1)}{2}\exp\left(-\frac{\delta_0^2}{4V}\right)\right] \\
&\overset{(i)}{=} \frac{k(k-1)}{2}\int_0^{+\infty} \mathbb{P}_{\mathbf{X}_0}\left\{\exp\left(-\frac{\delta_0^2}{4V}\right) > u\right\}\mathrm{d}u \\
&= \frac{k(k-1)}{2}\int_0^{+\infty} \mathbb{P}_{\mathbf{X}_0}\left\{V > \frac{\delta_0^2}{-4\log u}\right\}\mathrm{d}u \\
&\overset{(ii)}{\leq} \frac{k(k-1)}{2}\int_0^{+\infty} w_1 \exp\left\{-w_2\left(\frac{\delta_0^2}{-4R(m,n)\log u}\right)\right\}\mathrm{d}u \\
&\overset{(iii)}{\leq} \frac{w_1 k(k-1)}{2}\int_0^{+\infty} \exp\left\{-v - \left(\frac{w_2\delta_0^2}{4R(m,n)}\right)\frac{1}{v}\right\}\mathrm{d}v, \\
&\overset{(iv)}{=} \frac{w_1 k(k-1)}{2}\cdot\sqrt{\frac{w_2\delta_0^2}{R(m,n)}}\cdot K_1\left(\sqrt{\frac{w_2\delta_0^2}{R(m,n)}}\right),
\end{aligned} \tag{74}$$

where (i) uses the relation $\mathrm{E}(Z) = \int_0^\infty P(Z > t)\mathrm{d}t$ for any nonnegative random variable Z, (ii) follows from Assumption A.6 and the relation on the event $\mathcal{E}_4$, and (iii) uses a change of variable $v = -\log u$ in the integral. (iv) follows because the integral in (74) can be recognized as the density of a generalized inverse Gaussian distribution without normalizing constant, and here $K_1(\cdot)$ is the modified Bessel function of the second kind with parameter 1.

Theorem 2.13 of Kreh (2012) has shown that

$$\lim_{x\to+\infty} \frac{K_1(x)}{\sqrt{\frac{\pi}{2x}}e^{-x}} = 1,$$

which implies that there exists a constant $x_0 > 0$, such that for all $x > x_0$, $K_1(x) < 2\sqrt{\frac{\pi}{2x}}e^{-x} = \sqrt{\frac{2\pi}{x}}e^{-x}$. Since $R(m,n) \to 0$ for fixed n as $m \to \infty$, we can take $m \geq m_1$ for some large integer

$m_1 \geq m_0$ such that $\sqrt{\frac{w_2\delta_0^2}{R(m,n)}} > x_0$ and meanwhile

$$[R(m,n)]^{-1/4} \leq \exp\left\{\frac{1}{2}w_2^{1/2}\delta_0\,[R(m,n)]^{-1/2}\right\}.$$

As a result, we can derive from (74) that on the event $\mathcal{E}_4$, for all $m > m_1$,

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right] \leq \frac{w_1k(k-1)}{2}\cdot\sqrt{\frac{w_2\delta_0^2}{R(m,n)}}\cdot\sqrt{\frac{2\pi}{\sqrt{\frac{w_2\delta_0^2}{R(m,n)}}}}\exp\left\{-\sqrt{\frac{w_2\delta_0^2}{R(m,n)}}\right\}\\
&\leq \sqrt{\frac{\pi}{2}}w_1w_2^{1/4}k(k-1)\delta_0^{1/2}\,[R(m,n)]^{-1/4}\exp\left\{-w_2^{1/2}\delta_0\,[R(m,n)]^{-1/2}\right\}\\
&\leq \sqrt{\frac{\pi}{2}}w_1w_2^{1/4}k(k-1)\delta_0^{1/2}\exp\left\{-\frac{1}{2}w_2^{1/2}\delta_0\,[R(m,n)]^{-1/2}\right\}. \qquad (75)
\end{aligned}$$

Thus, $\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \exp\left\{-\frac{1}{2}w_2^{1/2}\delta_0\,[R(m,n)]^{-1/2}\right\}$ with probability at least $1-\xi$ for all $m \geq m_1$, which has proved Part (ii) of Theorem 5.

Finally, we prove Part (iii) of Theorem 5 with the additional Assumptions A.5 and A.7. Similar to the derivation of Part (ii), we define the quantity $\widetilde{V} = \max_{i\in\{1,\ldots,k\}}\sup_{\mathbf{x}_0\in\mathcal{X}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)$ for abbreviation. Then Assumption A.7 says that 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 $m \geq m_0$, for any $t > 0$, we have $\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_5) \geq 1-\xi$, where $\mathcal{E}_5$ is defined as $\mathcal{E}_5 = \left\{\widetilde{V} \leq w_3R(m,n)\right\}$. Therefore, from (73), we can derive that on the event $\mathcal{E}_5$, for all $m \geq m_0$,

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right] \leq \frac{k(k-1)}{2}\mathrm{E}_{\mathbf{X}_0}\exp\left(-\frac{\delta_0^2}{4\max_{i\in\{1,\ldots,k\}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)}\right)\\
&\leq \frac{k(k-1)}{2}\sup_{\mathbf{x}_0\in\mathcal{X}}\exp\left(-\frac{\delta_0^2}{4\max_{i\in\{1,\ldots,k\}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)}\right)\\
&= \frac{k(k-1)}{2}\exp\left(-\frac{\delta_0^2}{4\max_{i\in\{1,\ldots,k\}}\sup_{\mathbf{x}_0\in\mathcal{X}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)}\right)\\
&= \frac{k(k-1)}{2}\exp\left(-\frac{\delta_0^2}{4\widetilde{V}}\right) \leq \frac{k(k-1)}{2}\exp\left(-\frac{\delta_0^2}{4w_3R(m,n)}\right).
\end{aligned}$$

Thus, $\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \exp\left\{-\frac{\delta_0^2}{4w_3}\,[R(m,n)]^{-1}\right\}$ with probability at least $1-\xi$ for all $m \geq m_0$, which has proved Part (iii) of Theorem 5. □

Now we discuss the restrictiveness of Assumptions A.6 and A.7 in the main text. We present Theorem 6 below to illustrate that A.6 and A.7 can hold, by using the finite-rank kernel example as described in Remark 2 and Theorem 4 of the main text.

THEOREM 6. *(Exponentially decaying IPFS for finite-rank kernels) For a fixed positive integer k, consider the same model setup in Remark 2 of the main text with k finite-rank kernels $\mathbf{\Sigma}_{M,i} = a_i\left(\mathbf{x}^\top \mathbf{x}' + b_i\right)$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X} \subseteq \mathbb{R}^d$, where $a_i > 0$ and $b_i > 0$ are known constants for $i = 1, \ldots, k$. Let $\mathbb{P}_{\mathbf{X}}$ be any non-degenerate sampling distribution on $\mathcal{X}$ for $\mathbf{X}^m$ and $\mathbf{X}_0$.*

(i) Suppose that there exist constants $c_1 > 0, c_2 > 0, t_0 > 0$, such that $\mathbb{P}_{\mathbf{X}}$ has the tail bound $\mathbb{P}_{\mathbf{X}}(\|\mathbf{X}\| > t) \leq c_1 \exp(-c_2 t^2)$ for all $t > t_0$. Then for the optimal MSE given in (12) of the main text, 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 all $m \geq m_0$, for any $t > 0$,

$$\mathbb{P}_{\mathbf{X}^m}\left\{\mathbb{P}_{\mathbf{X}_0}\left(mn \cdot \max_{i\in\{1,\ldots,k\}} \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0) \geq t\right) \leq w_1 \exp\left(-w_2 t\right)\right\} \geq 1 - \xi. \tag{76}$$

(ii) Suppose that $\mathcal{X}$ is a compact set in $\mathbb{R}^d$. Then for the optimal MSE given in (12) of the main text, 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 all $m \geq m_0$,

$$\mathbb{P}_{\mathbf{X}^m}\left\{mn \cdot \max_{i\in\{1,\ldots,k\}} \sup_{\mathbf{x}_0\in\mathcal{X}} \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) \leq w_3\right\} \geq 1 - \xi. \tag{77}$$

Note that the rate $1/(mn)$ here is a tight convergence rate given Theorem 4, in the sense that it cannot be improved to any faster rate. The tail condition in Part (i) of Theorem 6 is satisfied by any d-dimensional multivariate normal distribution by the Hanson-Wright inequality (Hsu et al. 2012). Theorem 6 shows that Assumption A.6 holds for the finite-rank kernel if the sampling distribution of $\mathbf{X}^m$ and $\mathbf{X}_0$ has tail decaying like the Gaussian distribution. Similarly, Assumption A.7 holds when the covariance kernel and the f-functions are continuous with a compact domain.

Proof of Theorem 6:

First we show Part (i). We note that the tail condition $\mathbb{P}_{\mathbf{X}}(\|\mathbf{X}\| > t) \leq c_1 \exp(-c_2 t^2)$ implies the finite second moment for $\mathbb{P}_{\mathbf{X}}$, because,

$$\mathrm{E}_{\mathbf{X}}\left[\|\mathbf{X}\|^2\right] = \int_0^{+\infty} \mathbb{P}_{\mathbf{X}}\left(\|\mathbf{X}\|^2 > u\right) \mathrm{d}u \leq \int_0^{+\infty} c_1 \exp(-c_2 u) \mathrm{d}u = \frac{c_1}{c_2} < +\infty.$$

Furthermore, since $\mathbb{P}_{\mathbf{X}}$ is a non-degenerate sampling distribution on $\mathbb{R}$, the covariance matrix $\mathbf{V}_{\mathbf{X}} \equiv \mathrm{E}_{\mathbf{X}_0}\left\{[\mathbf{X}_0 - \mathrm{E}_{\mathbf{X}_0}(\mathbf{X}_0)][\mathbf{X}_0 - \mathrm{E}_{\mathbf{X}_0}(\mathbf{X}_0)]^\top\right\}$ must be positive definite. This is because otherwise, there exists a vector $\mathbf{a} \in \mathbb{R}^d$, such that

$$0 = \mathbf{a}^\top \mathrm{E}_{\mathbf{X}_0}\left\{[\mathbf{X}_0 - \mathrm{E}_{\mathbf{X}_0}(\mathbf{X}_0)][\mathbf{X}_0 - \mathrm{E}_{\mathbf{X}_0}(\mathbf{X}_0)]^\top\right\}\mathbf{a} = \mathrm{E}_{\mathbf{X}_0}\left\{\mathbf{a}^\top[\mathbf{X}_0 - \mathrm{E}_{\mathbf{X}_0}(\mathbf{X}_0)]\right\}^2,$$

which implies that $\mathbf{a}^\top \mathbf{X}_0$ is almost surely a constant, contradicting the assumption that $\mathbb{P}_{\mathbf{X}}$ is not degenerate.

For every $i = 1, \ldots, k$, we define

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

and $\widetilde{\mathbf{X}}_{i,0}$ is the $\mathbb{R}^{d+1}$ random vector version of $\widetilde{\mathbf{x}}_{i,0}$ with $\mathbf{X}_0$ following the distribution $\mathbb{P}_{\mathbf{X}}$. Define $\widetilde{\mathbf{V}}_i = \mathrm{E}_{\mathbf{X}_0}\left(\widetilde{\mathbf{X}}_{i,0}\widetilde{\mathbf{X}}_{i,0}^\top\right)$ for $i = 1, \ldots, k$. Then we can write that

$$\begin{aligned}
\widetilde{\mathbf{V}}_i = \mathrm{E}_{\mathbf{X}_{i,0}}\left(\widetilde{\mathbf{X}}_{i,0}\widetilde{\mathbf{X}}_{i,0}^\top\right) &= \begin{pmatrix} b_i & \sqrt{b_i}\mathrm{E}_{\mathbf{X}_{i,0}}(\mathbf{X}_{i,0}^\top) \\ \sqrt{b_i}\mathrm{E}_{\mathbf{X}_{i,0}}(\mathbf{X}_{i,0}) & \mathrm{E}_{\mathbf{X}_{i,0}}(\mathbf{X}_{i,0}\mathbf{X}_{i,0}^\top) \end{pmatrix} \\
&= \begin{pmatrix} \sqrt{b_i} & 0 \\ \mathrm{E}_{\mathbf{X}_{i,0}}(\mathbf{X}_{i,0}) & \mathbf{I}_d \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{V}_{\mathbf{X}} \end{pmatrix} \begin{pmatrix} \sqrt{b_i} & \mathrm{E}_{\mathbf{X}_{i,0}}(\mathbf{X}_{i,0}^\top) \\ 0 & \mathbf{I}_d \end{pmatrix}.
\end{aligned}$$

From the last expression, we can see that the matrix $\widetilde{\mathbf{V}}_i = \mathrm{E}_{\mathbf{X}_{i,0}}\left(\widetilde{\mathbf{X}}_{i,0}\widetilde{\mathbf{X}}_{i,0}^\top\right)$ must be positive definite since it is congruent to a block diagonal matrix which is positive definite. Let $\underline{\lambda}_{\min} \equiv \min_{i\in\{1,\ldots,k\}} \lambda_{\min}\big(\widetilde{\mathbf{V}}_i\big)$ which is strictly positive.

Similar to the convergence in (65) in the proof of Theorem 4, by the strong law of large numbers,

$\frac{1}{m}\mathbf{Z}_i^\top \mathbf{Z}_i$ converges to $\widetilde{\mathbf{V}}_i = \mathrm{E}_{\mathbf{X}_{i,0}}\left(\widetilde{\mathbf{X}}_{i,0}\widetilde{\mathbf{X}}_{i,0}^\top\right)$ entry-wise as $m \to \infty$ for each $i = 1, \ldots, k$. Furthermore, for a fixed k, we have that for each $i = 1, \ldots, k$, for any given $\xi \in (0, 1/2)$, there exists a large integer $m_{i,0} > 0$ that depends on ξ, such that for all $m \geq m_{i,0}$,

$$\mathbb{P}_{\mathbf{X}^m}\left(\left|\left|\left|\frac{1}{m}\mathbf{Z}_i^\top \mathbf{Z}_i - \widetilde{\mathbf{V}}_i\right|\right|\right| > \frac{1}{2}\lambda_{\min}\left(\widetilde{\mathbf{V}}_i\right)\right) < \frac{\xi}{k}.$$

Taking a union bound over all k designs implies that for all $m \geq m_0 \equiv \max_{i\in\{1,\ldots,k\}} m_{i,0}$ implies that

$$\mathbb{P}_{\mathbf{X}^m}\left(\left|\left|\left|\frac{1}{m}\mathbf{Z}_i^\top \mathbf{Z}_i - \widetilde{\mathbf{V}}_i\right|\right|\right| > \frac{1}{2}\lambda_{\min}\left(\widetilde{\mathbf{V}}_i\right), \text{ for all } i = 1, \ldots, k\right) < \sum_{i=1}^{k}\frac{\xi}{k} = \xi.$$

This further implies that with $\mathbb{P}_{\mathbf{X}^m}$-probability at least $1 - \xi$, for all $m \geq m_0$,

$$\begin{aligned}
\lambda_{\min}\left(\mathbf{I}_{d+1} + \frac{a_i n}{\sigma^2}\mathbf{Z}_i^\top \mathbf{Z}_i\right) &= \lambda_{\min}\left[\mathbf{I}_{d+1} + \frac{a_i mn}{\sigma^2}\left\{\frac{1}{m}\mathbf{Z}_i^\top \mathbf{Z}_i - \widetilde{\mathbf{V}}_i\right\} + \frac{a_i mn}{\sigma^2}\widetilde{\mathbf{V}}_i\right] \\
&\geq \lambda_{\min}\left(\mathbf{I}_{d+1} + \frac{a_i mn}{\sigma^2}\left[\widetilde{\mathbf{V}}_i - \frac{1}{2}\lambda_{\min}\left(\widetilde{\mathbf{V}}_i\right)\mathbf{I}_{d+1}\right]\right) \\
&\geq \lambda_{\min}\left[\mathbf{I}_{d+1} + \frac{a_i mn}{2\sigma^2}\lambda_{\min}\left(\widetilde{\mathbf{V}}_i\right)\mathbf{I}_{d+1}\right] \\
&> \frac{a_i mn}{2\sigma^2}\underline{\lambda}_{\min}.
\end{aligned}$$

Therefore, using the expression of $\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)$ derived in Theorem 4, we have that with $\mathbb{P}_{\mathbf{X}^m}$-probability at least $1 - \xi$, for any $t > 0$, for all $m \geq m_0$,

$$\begin{aligned}
&\mathbb{P}_{\mathbf{X}_0}\left(mn \cdot \max_{i\in\{1,\ldots,k\}} \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0) \geq t\right) \\
&= \mathbb{P}_{\mathbf{X}_0}\left(mn \cdot a_i \widetilde{\mathbf{X}}_{i,0}^\top \left(\mathbf{I}_{d+1} + \frac{a_i n}{\sigma^2}\mathbf{Z}_i^\top \mathbf{Z}_i\right)^{-1}\widetilde{\mathbf{X}}_{i,0} \geq t\right) \\
&\leq \mathbb{P}_{\mathbf{X}_0}\left(mna_i \cdot \left(\frac{a_i mn}{2\sigma^2}\underline{\lambda}_{\min}\right)^{-1}\widetilde{\mathbf{X}}_{i,0}^\top \widetilde{\mathbf{X}}_{i,0} \geq t\right) \\
&= \mathbb{P}_{\mathbf{X}_0}\left(b_i + \|\mathbf{X}_0\|^2 \geq \frac{\underline{\lambda}_{\min}}{2\sigma^2}t\right) \qquad (78)
\end{aligned}$$

Let $\overline{b} = \max_{i\in\{1,\ldots,k\}} b_i$. If $t > \max\left(4\sigma^2\underline{\lambda}_{\min}^{-1}\overline{b}, t_0\right)$, then for all $i = 1, \ldots, k$,

$$\frac{\underline{\lambda}_{\min}}{2\sigma^2}t - b_i > \frac{\underline{\lambda}_{\min}}{4\sigma^2}t,$$

and from the tail assumption $\mathbb{P}_{\mathbf{X}}(\|\mathbf{X}\| > t) \le c_1 \exp(-c_2 t^2)$ in Theorem 6, we have that

$$\begin{aligned}
&\mathbb{P}_{\mathbf{X}_0}\left(b_i + \|\mathbf{X}_0\|^2 \ge \frac{\underline{\lambda}_{\min}}{2\sigma^2} t\right) \le \mathbb{P}_{\mathbf{X}_0}\left(\|\mathbf{X}_0\|^2 \ge \frac{\underline{\lambda}_{\min}}{4\sigma^2} t\right) \\
&= \mathbb{P}_{\mathbf{X}_0}\left(\|\mathbf{X}_0\| \ge \frac{\sqrt{\underline{\lambda}_{\min}}}{2\sigma}\sqrt{t}\right) \le c_1 \exp\left(-\frac{c_2 \underline{\lambda}_{\min}}{4\sigma^2} t\right). \qquad (79)
\end{aligned}$$

If $0 < t \le \max\left(4\sigma^2 \underline{\lambda}_{\min}^{-1} \bar{b}, t_0\right)$, then we use the simple bound

$$\mathbb{P}_{\mathbf{X}_0}\left(b_i + \|\mathbf{X}_0\|^2 \ge \frac{\underline{\lambda}_{\min}}{2\sigma^2} t\right) \le 1 \le e^{c_2+1} \cdot \exp\left\{-t/\max\left(4\sigma^2 \underline{\lambda}_{\min}^{-1} \bar{b}, t_0\right)\right\}. \qquad (80)$$

Now let $w_1 = \max(e^{c_2+1}, c_1)$, $w_2 = \min\left\{c_2 \underline{\lambda}_{\min}/(4\sigma^2), \underline{\lambda}_{\min}/(4\sigma^2 \bar{b}), 1/t_0\right\}$, then (78), (79), and (80) together imply that with $\mathbb{P}_{\mathbf{X}^m}$-probability at least $1 - \xi$, for any $t > 0$, for all $m \ge m_0$,

$$\begin{aligned}
&\mathbb{P}_{\mathbf{X}_0}\left(mn \cdot \max_{i \in \{1,\ldots,k\}} \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0) \ge t\right) \\
&\le \mathbb{1}\left(0 < t \le 4\sigma^2 \underline{\lambda}_{\min}^{-1} \bar{b}\right) \cdot e^{c_2+1} \cdot \exp\left\{-t/\max\left(4\sigma^2 \underline{\lambda}_{\min}^{-1} \bar{b}, t_0\right)\right\} \\
&\quad + \mathbb{1}\left(t > 4\sigma^2 \underline{\lambda}_{\min}^{-1} \bar{b}\right) \cdot c_1 \exp\left(-\frac{c_2 \underline{\lambda}_{\min}}{4\sigma^2} t\right) \\
&\le w_1 \exp(-w_2 t).
\end{aligned}$$

This has proved Part (i) of Theorem 6.

Next we show Part (ii). Let $\underline{a} = \min_{i \in \{1,\ldots,k\}} a_i$ which is strictly positive given a fixed k. From the proof above, with $\mathbb{P}_{\mathbf{X}^m}$-probability at least $1 - \xi$, there exists a large integer m_0 such that uniformly for all $i = 1, \ldots, k$ and all $m \ge m_0$,

$$\lambda_{\min}\left(\mathbf{I}_{d+1} + \frac{a_i n}{\sigma^2} \mathbf{Z}_i^\top \mathbf{Z}_i\right) > \frac{a_i mn}{2\sigma^2} \underline{\lambda}_{\min}.$$

Since $\mathcal{X}$ is a compact set, there exists a constant $c_3 > 0$ such that $|\mathbf{x}| \le c_3$ for all $\mathbf{x} \in \mathcal{X}$. Recall that $\widetilde{\mathbf{x}}_{i,0} = (\sqrt{b_i}, \mathbf{x}_0^\top)^\top$ for any $\mathbf{x}_0 \in \mathcal{X}$. Therefore, with $\mathbb{P}_{\mathbf{X}^m}$-probability at least $1 - \xi$, for all $m \ge m_0$,

$$\begin{aligned}
&mn \cdot \max_{i \in \{1,\ldots,k\}} \sup_{\mathbf{x}_0 \in \mathcal{X}} \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) \\
&= mn \cdot \max_{i \in \{1,\ldots,k\}} \sup_{\mathbf{x}_0 \in \mathcal{X}} \widetilde{\mathbf{x}}_{i,0}^\top \left(\mathbf{I}_{d+1} + \frac{a_i n}{\sigma^2} \mathbf{Z}_i^\top \mathbf{Z}_i\right)^{-1} \widetilde{\mathbf{x}}_{i,0}
\end{aligned}$$

$$
\begin{aligned}
&\leq mn \cdot \max_{i\in\{1,\ldots,k\}} \left\{ \lambda_{\min}^{-1}\left(\mathbf{I}_{d+1} + \frac{a_i n}{\sigma^2}\mathbf{Z}_i^\top \mathbf{Z}_i\right) \sup_{\mathbf{x}_0\in\mathcal{X}} \widetilde{\mathbf{x}}_{i,0}^\top \widetilde{\mathbf{x}}_{i,0} \right\} \\
&\leq mn \cdot \max_{i\in\{1,\ldots,k\}} \left\{ \frac{2\sigma^2}{a_i mn}\underline{\lambda}_{\min}^{-1} \cdot \sup_{\mathbf{x}_0\in\mathcal{X}} \left(b_i + \|\mathbf{x}_0\|^2\right) \right\} \\
&\leq \frac{2\sigma^2(\bar{b} + c_3^2)\underline{\lambda}_{\min}^{-1}}{\underline{a}}.
\end{aligned}
$$

Set $w_3 = 2\sigma^2(\bar{b} + c_3^2)\underline{\lambda}_{\min}^{-1}/\underline{a}$ and then Part (ii) of Theorem 6 is proved. □

C Estimators of IMSE and IPFS

In this section, we propose simple estimators of IMSE and IPFS based on Monte Carlo draws from the sampling distribution $\mathbb{P}_{\mathbf{X}}$. Suppose that we already have the covariate sample $\mathbf{X}^m = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$. To estimate MSE, we draw another random sample $\widetilde{\mathbf{X}}^{m'} = \{\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_{m'}\}$ from the distribution $\mathbb{P}_{\mathbf{X}}$. The two samples $\mathbf{X}^m$ and $\tilde{\mathbf{X}}^m$ are independent. The sample size m' can be different from m. Then, according the definition of MSE in Equation (3) of the main text, we estimate the IMSE under the ith design ($i = 1, \ldots, k$) as

$$
\begin{aligned}
&\widehat{\mathrm{IMSE}}_i = \frac{1}{m'}\sum_{j=1}^{m'} \mathrm{MSE}_{i,\mathrm{opt}}(\widetilde{\mathbf{X}}_j), \qquad \text{where for } \; j = 1, \ldots, m', \\
&\mathrm{MSE}_{i,\mathrm{opt}}(\widetilde{\mathbf{X}}_j) = \boldsymbol{\Sigma}_{M,i}(\widetilde{\mathbf{X}}_j, \widetilde{\mathbf{X}}_j) - \boldsymbol{\Sigma}_{M,i}^\top(\mathbf{X}^m, \widetilde{\mathbf{X}}_j)\left[\boldsymbol{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{X}^m) + \boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{X}^m)\right]^{-1}\boldsymbol{\Sigma}_{M,i}(\mathbf{X}^m, \widetilde{\mathbf{X}}_j) \\
&\quad + \eta_i(\widetilde{\mathbf{X}}_j)^\top \left[\mathcal{F}_i^\top \left(\boldsymbol{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{X}^m) + \boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{X}^m)\right)^{-1}\mathcal{F}_i\right]^{-1}\eta_i(\widetilde{\mathbf{X}}_j), \\
&\text{and } \; \eta_i(\tilde{\mathbf{X}}_j) = \mathbf{f}_i(\widetilde{\mathbf{X}}_j) - \mathcal{F}_i^\top\left[\boldsymbol{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{X}^m) + \boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{X}^m)\right]^{-1}\boldsymbol{\Sigma}_{M,i}(\mathbf{X}^m, \widetilde{\mathbf{X}}_j). \qquad (81)
\end{aligned}
$$

It is straightforward to see that since $\widetilde{\mathbf{X}}^{m'}$ is an i.i.d. sample from $\mathbb{P}_{\mathbf{X}}$ and is independent of the sample $\mathbf{X}^m$, the proposed estimator $\widehat{\mathrm{IMSE}}_i$ in (81) is unbiased for the IMSE defined as $\mathrm{E}_{\mathbf{X}^m}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right]$. The maximal IMSE among the k designs can be then estimated by $\max_{i\in\{1,\ldots,k\}}\widehat{\mathrm{IMSE}}_i$.

For IPFS with an IZ parameter $\delta_0 > 0$, we first need to estimate the PFS at a given covariate

point $\mathbf{x}_0$, which can be approximated by the following quantity:

$$\mathrm{APFS}(\mathbf{x}_0) = \sum_{i \neq \widehat{i}^{\circ}(\mathbf{x}_0)} \mathbb{P}\left(N(0,1) < -\frac{\widehat{y}_i(\mathbf{x}_0) - \widehat{y}_{\widehat{i}^{\circ}(\mathbf{x}_0)}(\mathbf{x}_0) + \delta_0}{\sqrt{\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) + \mathrm{MSE}_{\widehat{i}^{\circ}(\mathbf{x}_0),\mathrm{opt}}(\mathbf{x}_0)}} \right), \tag{82}$$

where $\widehat{i}^{\circ}(\mathbf{x}_0)$ and $\widehat{y}_i(\mathbf{x}_0)$ are defined in Equation (4) of the main text, $\widehat{y}_i(\mathbf{x}_0)$ is defined in Equation (2) of the main text, and $\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)$ is defined in Equation (3) of the main text. Then, based on the random sample $\widetilde{\mathbf{X}}^{m'} = \{\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_{m'}\}$ from the distribution $\mathbb{P}_{\mathbf{X}}$ independent of $\mathbf{X}^m$, we can estimate the IPFS as

$$\widehat{\mathrm{IPFS}} = \frac{1}{m'} \sum_{j=1}^{m'} \mathrm{APFS}(\widetilde{\mathbf{X}}_j), \tag{83}$$

where $\mathrm{APFS}(\cdot)$ is defined in (82). The $\widehat{\mathrm{IPFS}}$ in (83) is a consistent estimator of $\mathrm{IPFS} = \mathrm{E}_M \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{PFS}(\mathbf{X}_0)\right]$.

D Analysis for the Case of Unequal n_i's

Let n_i be the number of simulation replications allocated to each of the m covariate points with design i, $i = 1, \ldots, k$. In this section, we fix the number of covariate points m, allow n_i to be unequal among different designs i, and develop a ranking and selection (R&S) framework for optimizing the simulation budget allocation n_i's in simulation with covariates introduced in the main text.

Suppose that the m covariate points collected are $\mathbf{x}_1, \ldots, \mathbf{x}_m$, and the total simulation budget to be allocated among pairs of covariate points and designs is n_{tot}, i.e., $m \sum_{i=1}^{k} n_i = n_{tot}$. With the target measures of the maximal IMSE and IPFS, the corresponding R&S problems can be formulated as

$$\begin{aligned} &\min \max_{i \in \{1,2,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \\ &\text{s.t. } m \sum_{i=1}^{k} n_i = n_{tot}, \text{ and } n_i \geq 0, \text{ for } i = 1, \ldots, k, \end{aligned} \tag{84}$$

and

$$\min \mathrm{E}_M \mathrm{E}_{\mathbf{X}_0} \left[\mathrm{PFS}(\mathbf{X}_0)\right] \tag{85}$$
$$\text{s.t. } m\sum_{i=1}^{k} n_i = n_{tot}, \text{ and } n_i \geq 0, \text{ for } i = 1, \ldots, k,$$

However, both optimization problems (84) and (85) cannot be directly solved due to the lack of analytical expressions of the objective functions $\max_{i\in\{1,2,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right]$ and $\mathrm{E}_M \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{PFS}(\mathbf{X}_0)\right]$. Here we propose two methods to approximate them.

Our first proposal is to replace the maximal IMSE and IPFS in (84) and (85) with their Monte Carlo estimators proposed in Section C. Both $\max_{i\in\{1,\ldots,k\}} \widehat{\mathrm{IMSE}}_i$ defined in (81) and $\widehat{\mathrm{IPFS}}$ defined in (83) have already taken into account the unequal n_i's in the matrix $\boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{X}^m)$. We can choose the Monte Carlo sample size m' according to the optimization budget. Then (84) and (85) can be solved using numerical optimization methods.

Our second proposal is to approximate them by the analytical upper bounds in our Theorems 1 and 2. Note that analytical approximations are common in solving R&S problems, especially in the OCBA method (Chen et al. 2000, 2008). They make the optimization problem tractable, and can often lead to efficient budget allocation rules.

Let $\{\mu_{i,l} : l = 1, 2, \ldots\}$ be the eigenvalues of the linear operator $T_{\boldsymbol{\Sigma}_{M,i}}$ defined in Section 2.1 of the main text. We recall from the second paragraph after Assumptions A.1-A.4 that the constants r_* and ρ_* in Assumption A.3 can be made common for all the k designs. Using the results in Theorems 1, 2 and 5, we can prove the following proposition.

PROPOSITION 1. *Suppose that Assumptions A.1 - A.4 in the main text hold for all the k designs. Let $\varrho_i = mn_i/n_{tot}$. For any $0 \leq \varrho \leq 1$, define the following quantities for $i = 1, \ldots, k$:*

$$\begin{aligned} R_i(\varrho) = {} & \frac{2\overline{\sigma}_0^2}{n_{tot}\varrho}\gamma_i\left(\frac{\overline{\sigma}_0^2}{n_{tot}\varrho}\right) + \frac{64C_i^\dagger q\overline{\sigma}_0^2 \operatorname{tr}(\boldsymbol{\Sigma}_{M,i})}{n_{tot}\varrho} \\ & + \inf_{\zeta\in\mathbb{N}}\left[\left\{\frac{64C_i^\dagger q\rho_*^4\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\operatorname{tr}\left(\boldsymbol{\Sigma}_{M,i}\right)^2 + 8C_i^\dagger q\operatorname{tr}\left(\boldsymbol{\Sigma}_{M,i}\right) + \frac{3}{\overline{\sigma}_0^2}\operatorname{tr}(\boldsymbol{\Sigma}_{M,i}) + 1\right\}\operatorname{tr}\left(\boldsymbol{\Sigma}_{M,i}^{(\zeta)}\right)n_{tot}\varrho\right. \\ & \left. + \left[8C_i^\dagger q\operatorname{tr}\left(\boldsymbol{\Sigma}_{M,i}\right)^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{M,i}\right)\right]\left\{300\rho_*^2\frac{b(m,\zeta,r_*)}{\sqrt{m}}\gamma_i\left(\frac{\overline{\sigma}_0^2}{n_{tot}\varrho}\right)\right\}^{r_*}\right], \end{aligned} \tag{86}$$

where A is the universal constant in Theorem 1 and

$$C_i^{\dagger} = C_{\mathrm{f},i}^2/\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^{\top}]\right), \quad C_{\mathrm{f},i} = \max_{1\le s\le q}\|\mathrm{f}_s\|_{\mathbb{H}_i}$$

$$\gamma_i(a) = \sum_{l=1}^{\infty}\frac{\mu_{i,l}}{\mu_{i,l}+a} \text{ for any } a>0,$$

$$\mathrm{tr}\left(\mathbf{\Sigma}_{M,i}\right) = \sum_{l=1}^{\infty}\mu_{i,l}, \quad \mathrm{tr}\left(\mathbf{\Sigma}_{M,i}^{(\zeta)}\right) = \sum_{l=\zeta+1}^{\infty}\mu_{i,l} \text{ for any } \zeta\in\mathbb{N},$$

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

Then, for the measures of the maximal IMSE and IPFS, we have

$$\max_{i\in\{1,\ldots,k\}}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \max_{i\in\{1,\ldots,k\}} R_i(\varrho_i), \tag{87}$$

$$\mathrm{E}_M\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{PFS}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \max_{i\in\{1,\ldots,k\}} R_i(\varrho_i),. \tag{88}$$

Proof of Proposition 1:

By directly combining the upper bounds in Theorems 1 and 2 together with the MSE decomposition in Equation (6) of the main text, we have that with $\mathbb{P}_{\mathbf{X}^m}-$ probability approaching 1, for each $i=1,\ldots,k$,

$$\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] = \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] + \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] \le R_i(mn_i/n_{tot}) = R_i(\varrho_i),$$

where $R_i(\cdot)$ is defined in (86) above and is slightly larger than the combined upper bounds from Theorems 1 and 2 by adjusting some constants. This implies the following upper bound

$$\max_{i\in\{1,\ldots,k\}}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \max_{i\in\{1,\ldots,k\}} R_i(\varrho_i),$$

which proves (87).

For the IPFS measure, we notice that for each of the three cases in Theorem 5, the upper bound is a monotone increasing function of $R(m,n)$, which is defined as a probablistic upper bound for $\max_{i\in\{1,\ldots,k\}}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right]$ in Theorem 3 of the main text. Therefore, the inequality in (88) holds as well. □

With Proposition (1), we can build an analytical R&S model for both (84) and (85),

$$\begin{aligned} &\min \max_{i\in\{1,\ldots,k\}} R_i(\varrho_i) \\ &\text{s.t.} \sum_{i=1}^{k} \varrho_i = 1, \text{ and } \varrho_i \geq 0, \text{ for } i = 1, \ldots, k. \end{aligned} \tag{89}$$

This is a typical nonlinear optimization problem. Its optimal solution ϱ_i^* gives us an approximately optimal allocation of the simulation budget among pairs of covariate points and designs with $n_i^* = \frac{\varrho_i^* n_{tot}}{m}$, $i = 1, 2, ..., k$.

Problem (89) involves a number of constants that depend on the properties of the covariance kernels used in the k designs. These constants can be made concrete when the covariate space $\mathcal{X}$, the covariance kernels $\mathbf{\Sigma}_{M,i}$, the sampling distribution $\mathbb{P}_{\mathbf{X}}$, and the regression functions $\mathbf{f}_{i1}, \ldots, \mathbf{f}_{iq}$ are fully specified in practice. Problem (89) is not necessarily a convex optimization problem. Since it is built based on a different setting (fixed m and unequal n_i's) from that of the main questions in this research, we do not pursue further development of it in this paper. We emphasize that our proposed theoretical analysis and results can be used to formulate and solve R&S type of problems that arise in simulation with covariates.

E Additional Numerical Results

This section provides additional numerical results to the main text. Section E.1 plots the two test functions in Section 5.1 of the main text. Section E.2 compares our static sampling with an adaptive design procedure, under the target measures of the maximal IMSE and IPFS. Section E.3 provides a procedure that can help the analyst make the design decision for achieving a target precision of the maximal IMSE.

E.1 Plots of the Test Functions used in Section 5.1 of the Main Text

The 1-d De Jong's function and Griewank's function without noise are shown in Figure 7. The 2-d De Jong's function under selected designs without noise is shown in Figure 8.

In the 1-d case (Figure 7), we present ten curves for the two test functions, each corresponding to $M(\mathbf{x})$ at one of the ten designs. It can be observed that no design can dominate the others in the

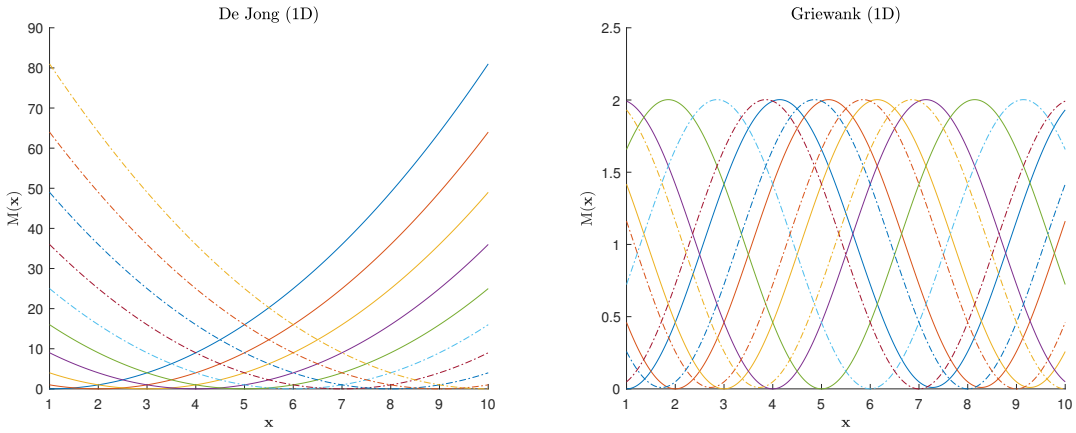

Figure 7: Plots of the two test functions for 1-dimensional $\mathbf{x}$. Ten different curves stand for the ten designs.

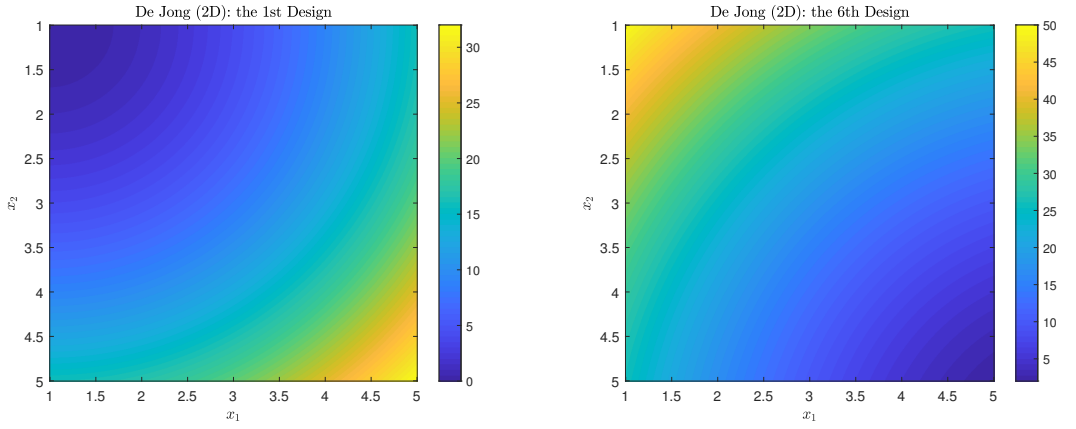

Figure 8: Heatmaps of De Jong's function for 2-dimensional $\mathbf{x}$ under the 1st and 6th design ($i = 1$ and $i = 6$).

tested functions, and the best design might not be unique for some $\mathbf{x}$. The De Jong's functions are smooth while 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. In the 2-d case (Figure 8), we present the heatmap of the 1st and 6th design ($i = 1$ and $i = 6$) for the De Jong's function. We can see that $M(\mathbf{x})$ varies a lot with $\mathbf{x}$.

E.2 Comparison between Static Sampling and an Adaptive Procedure

In this section, we compare our static sampling (i.e., fixed-distribution sampling; this is the sampling method studied in this research) with an intuitive adaptive design procedure. The adaptive procedure works in a greedy manner and iteratively collects the covariate point that maximizes the largest MSE of the fitted SK models. In this way, it sequentially explores the whole covariate space and reduces the overall MSE of the SK prediction. We call it Adaptive MSE Procedure.

Suppose that m_1 covariate points $\mathbf{x}^{m_1} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{m_1}\}$ have been sampled already. For the illustration purpose, we will use the superscript $[m_1]$ to indicate that the SK estimators are derived from the current simulation samples $\mathbf{x}^{m_1}$. From Equation (3) in the main text, the mean squared error of the current-stage SK predictor of design i at $\mathbf{x}_0$ is

$$\begin{aligned}\mathrm{MSE}_{i,\mathrm{opt}}^{[m_1]}(\mathbf{x}_0) = {} & \boldsymbol{\Sigma}_{M,i}(\mathbf{x}_0, \mathbf{x}_0) - \boldsymbol{\Sigma}_{M,i}^{\top}(\mathbf{x}^{m_1}, \mathbf{x}_0) \left[\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^{m_1}, \mathbf{x}^{m_1}) + \boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^{m_1})\right]^{-1} \boldsymbol{\Sigma}_{M,i}(\mathbf{x}^{m_1}, \mathbf{x}_0) \\ & + \eta_i^{[m_1]}(\mathbf{x}_0)^{\top} \left[(\mathcal{F}_i^{[m_1]})^{\top} \left(\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^{m_1}, \mathbf{x}^{m_1}) + \boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^{m_1})\right)^{-1} \mathcal{F}_i^{[m_1]}\right]^{-1} \eta_i^{[m_1]}(\mathbf{x}_0), \quad (90)\end{aligned}$$

where $\eta_i^{[m_1]}(\mathbf{x}_0) = \mathbf{f}_i(\mathbf{x}_0) - (\mathcal{F}_i^{[m_1]})^{\top} \left(\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^{m_1}, \mathbf{x}^{m_1}) + \boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^{m_1})\right)^{-1} \boldsymbol{\Sigma}_{M,i}(\mathbf{x}^{m_1}, \mathbf{x}_0)$, $\mathcal{F}_i^{[m_1]} = (\mathbf{f}_i(\mathbf{x}_1), \ldots, \mathbf{f}_i(\mathbf{x}_{m_1}))^{\top}$, and $\boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^{m_1})$ is the $m_1 \times m_1$ covariance matrix of the averaged simulation errors across m_1 covariate points under design i.

The Adaptive MSE Procedure samples the next covariate point $\mathbf{x}_{m_1+1}$ with the largest maximal $\mathrm{MSE}_{i,\mathrm{opt}}^{[m_1]}(\mathbf{x}_0)$, where "largest" is over the covariate space $\mathcal{X}$ and "maximal" is over the k SK models. That is,

$$\mathbf{x}_{m_1+1} = \arg\max_{\mathbf{x}_0 \in \mathcal{X}} \max_{i \in \{1,\ldots,k\}} \mathrm{MSE}_{i,\mathrm{opt}}^{[m_1]}(\mathbf{x}_0). \quad (91)$$

The formal description of the Adaptive MSE Procedure is given as follows.

Adaptive MSE Procedure

1. Specify the covariate space $\mathcal{X}$ and the total number of covariate points m. Perform n_0 replications for the pair of the center point of the covariate space and design i, $i = 1, \ldots, k$. $m_1 \leftarrow 0$.

2. If $m_1 > m$, stop. Otherwise,

   a. Obtain $\mathbf{x}_{m_1+1}$ by (91).

   b. Perform n_0 replications for the pair of covariate point $\mathbf{x}_{m_1+1}$ and design i, $i = 1, \ldots, k$.

   c. Update the SK model for each design $i = 1, \ldots, k$. $m_1 \leftarrow m_1 + 1$.

We use the De Jong's and Griewank's functions under the same parameter settings as in Section 5.1 of the main text for testing, i.e., the covariate space is $\mathcal{X} = [1, 10]^d$ and there are $k = 10$ designs. Meanwhile, we vary the domain dimension d and the sampling distribution $\mathbb{P}_{\mathbf{X}}$. Specifically, we test the following examples on our static sampling from $\mathbb{P}_{\mathbf{X}}$ and the Adaptive MSE Procedure:

(i) De Jong's functions, for dimension $d = 1$, $\mathbb{P}_{\mathbf{X}}$ being the truncated $N(5.5, 1^2)$;

(ii) De Jong's functions, for dimension $d = 1$, $\mathbb{P}_{\mathbf{X}}$ being the truncated $N(5.5, 0.25^2)$;

(iii) De Jong's functions, for dimension $d = 2$, $\mathbb{P}_{\mathbf{X}}$ being the truncated $N(5.5, 0.3^2)$ in each dimension;

(iv) De Jong's functions, for dimension $d = 3$, $\mathbb{P}_{\mathbf{X}}$ being the truncated $N(2.5, 0.3^2)$ in each dimension;

(v) Griewank's functions, for dimension $d = 1$, $\mathbb{P}_{\mathbf{X}}$ being the uniform distribution on $[1, 10]$;

(vi) Griewank's functions, for dimension $d = 1$, $\mathbb{P}_{\mathbf{X}}$ being the truncated $N(5.5, 1^2)$;

(vii) Griewank's functions, for dimension $d = 10$, $\mathbb{P}_{\mathbf{X}}$ being the uniform distribution on $[1, 10]^{10}$;

(viii) Griewank's functions, for dimension $d = 10$, $\mathbb{P}_{\mathbf{X}}$ being the truncated $N(2.5, 0.75^2)$ in each dimension.

Figures 9-16 report the comparison results for our static sampling and the Adaptive MSE Procedure under the measures of the maximal IMSE and IPFS. We have the following observations:

- For Case (i), where $d = 1$ and the De Jong's functions are smooth enough, the Adaptive MSE Procedure has smaller maximal IMSE and IPFS than the static sampling from $\mathbb{P}_{\mathbf{X}}$.

- For Cases (ii), (iii), and (iv) with the De Jong's functions in dimension $d = 1, 2, 3$, where the normal variance becomes smaller, i.e., the sampling distribution $\mathbb{P}_{\mathbf{X}}$ becomes more concentrated, the static sampling from $\mathbb{P}_{\mathbf{X}}$ has slightly better performance than the Adaptive MSE Procedure, but overall their performances are similar.

- For Case (v), where $d = 1$, the sampling distribution $\mathbb{P}_{\mathbf{X}}$ is uniform, and the target is the Griewank's functions, we can see from Figure 13 that the Adaptive MSE Procedure has slightly smaller maximal IMSE and IPFS than the static sampling, but overall their performances are similar.

- For Case (vi), where $d = 1$, the sampling distribution $\mathbb{P}_{\mathbf{X}}$ is truncated normal with a moderately large variance, and the target Griewank's functions have strong oscillation, we can see from Figure 14 that the static sampling almost always yields smaller maximal IMSE and IPFS than the Adaptive MSE Procedure.

- For Cases (vii) and (viii), where the dimension is high ($d = 10$) and the target Griewank's functions have strong oscillation, we can see from Figures 15 and 16 that the static sampling always yields much smaller maximal IMSE and IPFS than the Adaptive MSE Procedure, for both the uniform distribution and the truncated normal distribution.

In conclusion, the static sampling from $\mathbb{P}_{\mathbf{X}}$ seems to yield comparable performance to the Adaptive MSE Procedure under the two measures in general. The static sampling tends to perform better than the Adaptive MSE Procedure when the target function has strong oscillation, the dimension becomes higher, and the covariate distribution $\mathbb{P}_{\mathbf{X}}$ becomes more concentrated.

E.3 Achieving a Target Precision of the Maximal IMSE

Based on the linear decreasing trend of the maximal IMSE in Figure 5 of the main text, we propose a simple procedure to determine the sample size m_0 such that the maximal IMSE satisfies $\max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] = c_0$ for a target precision c_0. Suppose that we have already drawn m covariate points $\mathbf{X}^m$ from $\mathbb{P}_{\mathbf{X}}$ and each covariate point has n simulation replications.

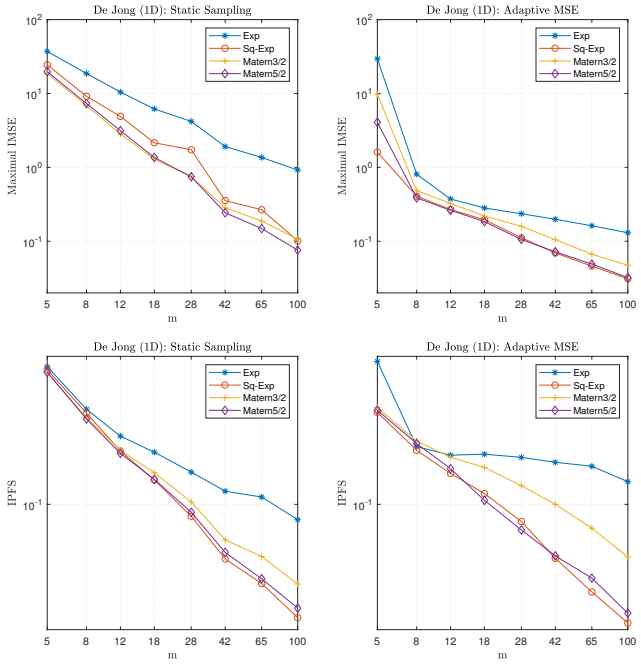

Figure 9: Truncated $N(5.5, 1^2)$ of $d = 1$: The maximal IMSE and IPFS for the 1-dimensional De Jong's functions and four covariance kernels.

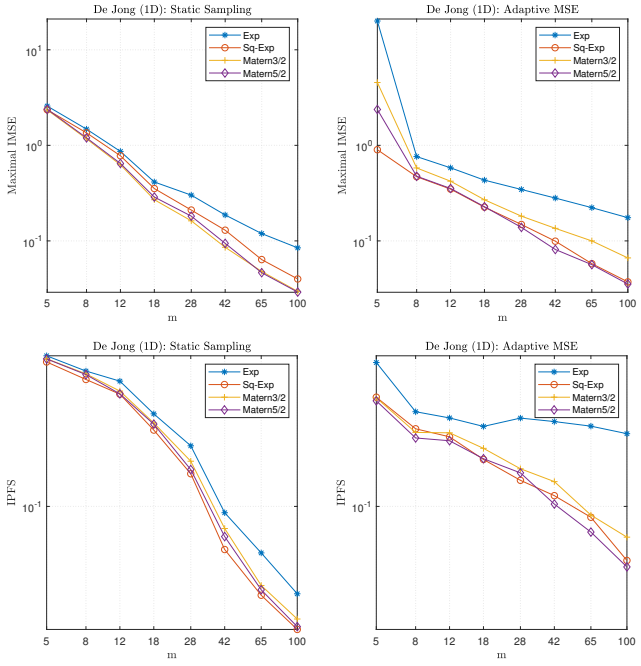

Figure 10: Truncated $N(5.5, 0.25^2)$ of $d = 1$: The maximal IMSE and IPFS for the 1-dimensional De Jong's functions and four covariance kernels.

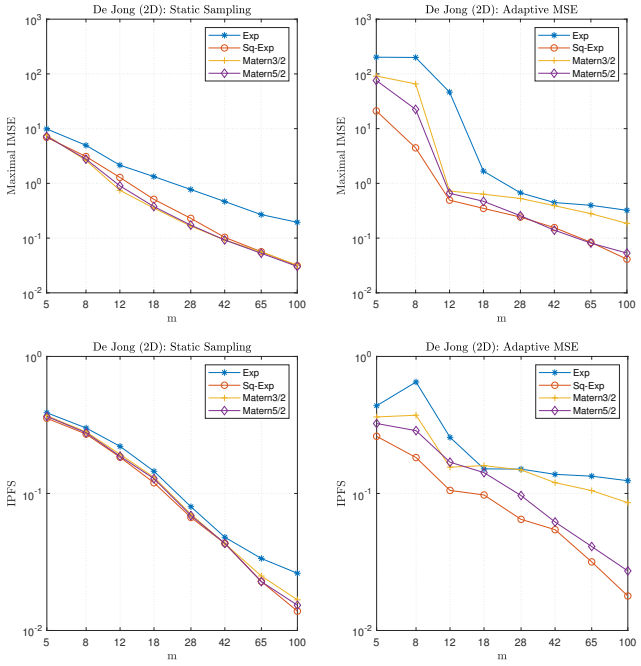

Figure 11: Truncated $N(5.5, 0.3^2)$ on each dimension of $d = 2$: The maximal IMSE and IPFS for the 2-dimensional De Jong's functions and four covariance kernels.

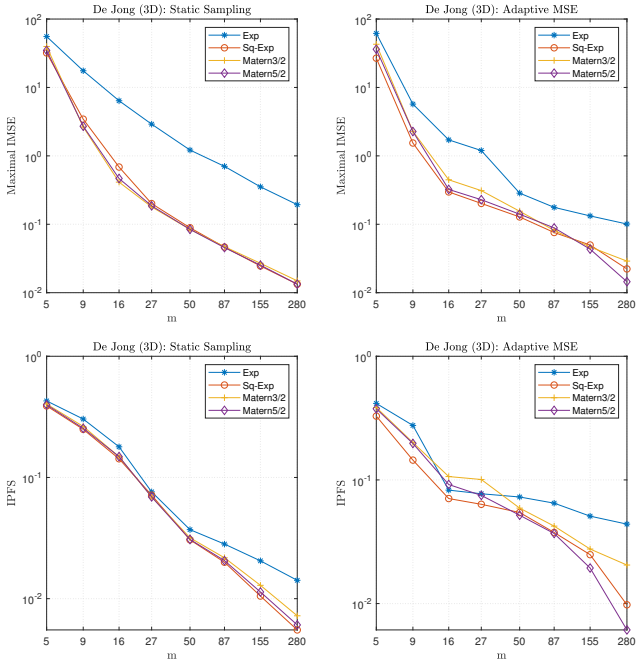

Figure 12: Truncated $N(2.5, 0.3^2)$ on each dimension of $d = 3$: The maximal IMSE and IPFS for the 3-dimensional De Jong's functions and four covariance kernels.

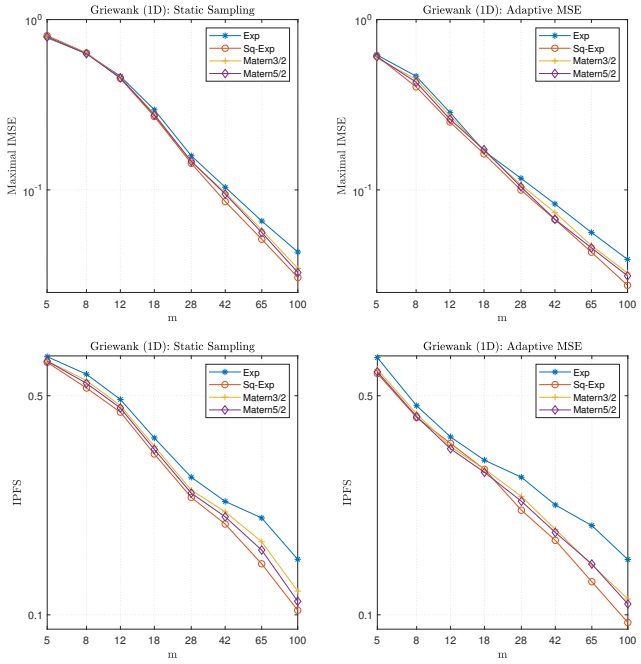

Figure 13: Uniform distribution of $d = 1$: The maximal IMSE and IPFS for the 1-dimensional Griewank's functions and four covariance kernels.

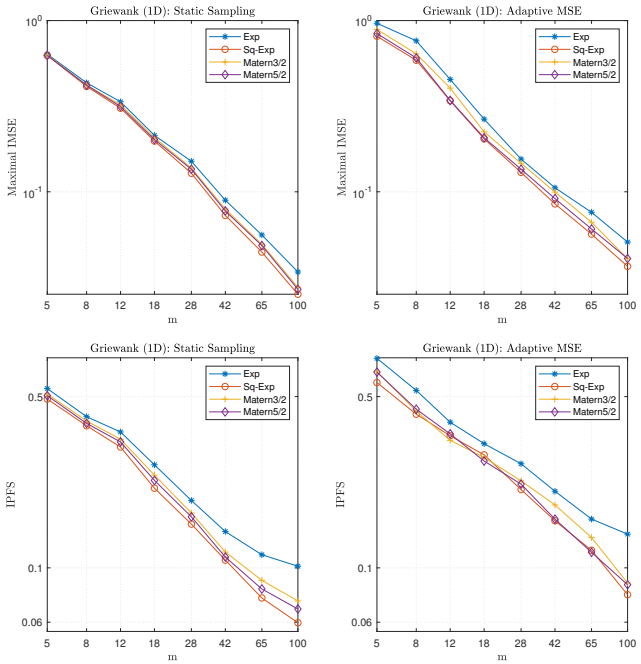

Figure 14: Truncated $N(5.5, 1^2)$ of $d = 1$: The maximal IMSE and IPFS for the 1-dimensional Griewank's functions and four covariance kernels.

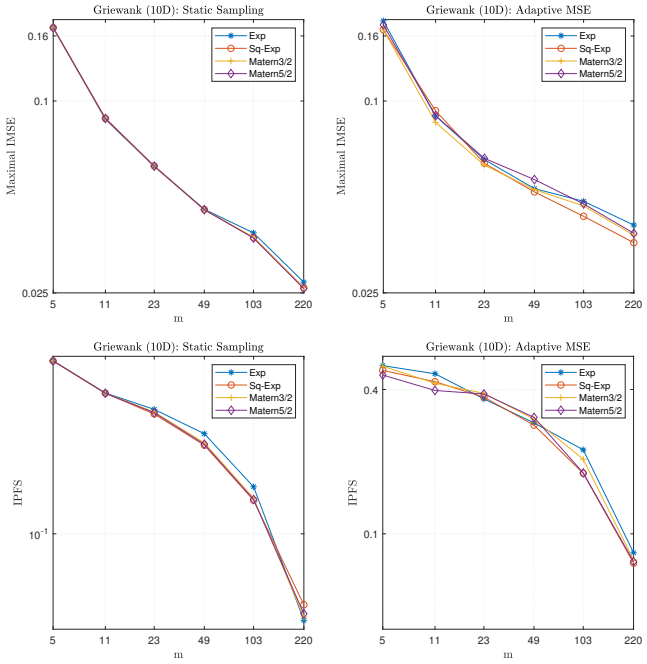

Figure 15: Uniform distribution of $d = 10$: The maximal IMSE and IPFS for the 10-dimensional Griewank's functions and four covariance kernels.

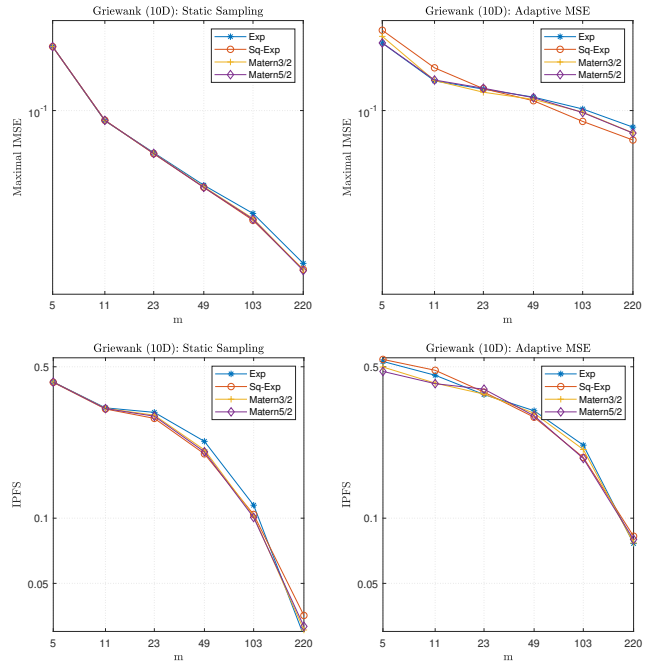

Figure 16: Truncated $N(2.5, 0.75^2)$ on each dimension of $d = 10$: The maximal IMSE and IPFS for the 10-dimensional Griewank's functions and four covariance kernels.

Then, for an integer $L \geq 3$, we draw $L-1$ subsamples of sizes $m_1 < \ldots < m_{L-1} (< m_L \equiv m)$ from $\mathbf{X}^m$ without replacement. Denote these subsets as $\mathbf{X}^{m_1}, \ldots, \mathbf{X}^{m_{L-1}}$. We then fit $(L-1)k$ SK models based on each dataset of $\mathbf{X}^{m_1}, \ldots, \mathbf{X}^{m_{L-1}}$, and estimate the maximal IMSE for each subset using the Monte Carlo estimator described in Section 3 of the Online Supplement. We repeat this subsampling-fitting-estimating process for multiple times and take the average of the estimated maximal IMSE's at each size $m_1, \ldots, m_{L-1}, m_L$, denoted by $\max_{i\in\{1,\ldots,k\}} \widehat{\text{IMSE}}_i(m_l)$, $l = 1, \ldots, L$. Finally, we fit the linear model $\log(\max_{i\in\{1,\ldots,k\}} \text{IMSE}_i) = c_1 + c_2 \log m + \text{error}$ using the pairs $\left\{ \left(\max_{i\in\{1,\ldots,k\}} \widehat{\text{IMSE}}_l(m_l), m_l \right) : l = 1, \ldots, L \right\}$, and predict m_0 by $\widehat{m}_0 = \exp\{(\log c_0 - \widehat{c}_1)/\widehat{c}_2\}$, where $\widehat{c}_1, \widehat{c}_2$ are the fitted linear coefficients.

Next, we apply this procedure to the M/M/1 queue example. We draw $m = 80$ covariate points from the sampling distribution $\mathbb{P}_{\mathbf{X}}$ with $n = 10$ replications, and estimate the maximal IMSE with subsample sizes $\{10, 15, 23, 35, 53, 80\}$. For example, for the squared exponential kernel and uniform sampling distribution, we obtain the fitted linear regression model $\log(\max_{i\in\{1,\ldots,k\}} \text{IMSE}_i) = -1.03 \log(m) - 4.58$ and the predicted $\widehat{m}_0 \approx 119$ such that $\max_{i\in\{1,\ldots,k\}} \text{IMSE}_i = c_0 = 7.5 \times 10^{-5}$. To numerically verify whether the true maximal IMSE is around 7.5×10^{-5} at sample size $\widehat{m}_0 = 119$, we randomly draw another 39 covariate points from the uniform distribution, establish the SK models based on the union of the 39 new points and the 80 existing points, and compute the maximal IMSE. We repeat this process for 40 macro Monte Carlo replications. We find that the median maximal IMSE over the 40 macro replications is 7.37×10^{-5}. The numerical results for the two tested sampling distributions and four covariance kernels are summarized in Table 2. In almost all cases, the predicted $\widehat{m}_0$ values yield very similar or smaller maximal IMSE's compared to the target values. This demonstrates that our theory can help the decision makers determine the number of additional covariate points needed to achieve a target precision.

References

Ahmed, M. A., T. M. Alkhamis. 2009. Simulation optimization for an emergency department healthcare unit in Kuwait. *Eur J Oper Res*, 198, 936–942.

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

Table 2: Prediction of sample size m_0 for a maximal IMSE precision c_0 based on $m = 80$ covariate points.

	Kernels	c_0	$\widehat{c}_1$	$\widehat{c}_2$	$\widehat{m}_0$	Mean	Median
uniform, $n = 10$	`SqExp`	7.5×10^{-5}	-1.03	-4.58	119	7.37×10^{-5}	7.37×10^{-5}
	`Matern 5/2`	7.5×10^{-5}	-1.12	-4.06	130	7.07×10^{-5}	6.95×10^{-5}
	`Matern 3/2`	7.5×10^{-5}	-1.12	-3.92	147	7.19×10^{-5}	6.95×10^{-5}
	`Exp`	2.0×10^{-4}	-0.95	-3.99	118	2.23×10^{-4}	2.09×10^{-4}
truncated normal, $n = 10$	`SqExp`	7.5×10^{-5}	-1.00	-4.70	122	7.44×10^{-5}	7.04×10^{-5}
	`Matern 5/2`	7.5×10^{-5}	-1.06	-4.24	144	6.57×10^{-5}	6.45×10^{-5}
	`Matern 3/2`	7.5×10^{-5}	-1.08	-4.04	160	6.67×10^{-5}	6.69×10^{-5}
	`Exp`	2.0×10^{-4}	-0.95	-4.00	117	2.34×10^{-4}	2.11×10^{-4}

Notes: *"Mean" is the sample average of the maximal IMSE over 40 macro Monte Carlo replications. "Median" is the sample median of the maximal IMSE over 40 macro Monte Carlo replications.*

Benini, L., R. Hodgson, P. Siegel. 1998. System-level power estimation and optimization. *Proc 1998 International Symposium on Low Power Electronics and Design*, 173–178.

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

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

Chen, C. H., L. H. Lee. 2011. *Stochastic Simulation Optimization: An Optimal Computing Budget Allocation*. Singapore: World Scientific Publishing.

Chen, C. H., J. Lin, E. Yücesan, S. E. Chick. 2000. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dyn S*, 10, 251–270.

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

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

Ding, H., L. Benyoucef, X. Xie. 2005. A simulation optimization methodology for supplier selection problem. *Int J Comput Integ Manuf*, 18, 210–224.

Frazier, P. I., W. B. Powell, S. Dayanik. 2008. A knowledge-gradient policy for sequential information collection. *SIAM J Contr Optim*, 47(5), 2410–2439.

Gao, S., W. Chen. 2017. Efficient feasibility determination with multiple performance measure constraints. *IEEE Trans Automat Contr*, 62, 113–122.

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

Gao, S., J. Du, C.-H. Chen. 2019a. Selecting the optimal system design under covariates. In *2019 ieee 15th international conference on automation science and engineering (case)*, 547–552. IEEE.

Gao, S., C. Li, J. Du. 2019b. Rate analysis for offline simulation online application. In *Proc. 2019 Winter Simulation Conf.*, 3468–3479.

Garud, S. S., I. A. Karimi, M. Kraft. 2017. Design of computer experiments: A review. *Computers and Chemical Engineering*, 106, 71–95.

Glynn, P., S. Juneja. 2004. A large deviations perspective on ordinal optimization. In *Proc. 2004 Winter Simulation Conf.*, 577–585.

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

Hensman, J., N. Fusi, N. Lawrence. 2014. Gaussian processes for big data. In *Proc. 29th Conference on Uncertainty in Artificial Intelligence*, 282–290.

Hong, L. J., G. Jiang. 2019. Offine simulation online application: a new framework of simulation-based decision making. *Asia Pac J Oper Res*, 36(6), 1940015.

Hsing, T., R. Eubank. 2015. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear operators*. John Wiley & Sons.

Hsu, D., S. M. Kakade, T. Zhang. 2012. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52), 1–6.

Kim, S. H., B. L. Nelson. 2006. Selecting the best system. In Henderson, S. G., B. L. Nelson, editors, *Simulation*, Handbooks in Operations Research and Management Science, chapter 13, 501–534. Elsevier, Amsterdam, Netherlands.

Kleijnen, J. P. C. 1993. Simulation and optimization in production planning: A case study. *Decis Support Syst*, 9(3), 269–280.

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

Kreh, M. 2012. *Bessel Functions. Lecture Notes, Penn State - Göttingen Summer School on Number Theory.*

Law, A. M. 2015. *Simulation Modeling and Analysis*. 5th edition. McGraw-Hill, New York.

Luo, Y., R. Duraiswami. 2013. Fast near-GRID Gaussian process regression. In *Proc. 16th International Conference on Artificial Intelligence and Statistics*, 424–432.

Ni, E. C., D. F. Ciocan, S. G. Henderson, S. R. Hunter. 2017. Efficient ranking and selection in parallel computing environments. *Oper Res*, 65(3), 821–836.

Qu, H., M. C. Fu. 2014. Gradient extrapolated stochastic kriging. *ACM Trans Model Comput Simul*, 24(4). Article 3.

Rasmussen, C. E., C. K. Williams. 2006. *Gaussian Process for Machine Learning*. MIT press.

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

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

Santin, G., R. Schaback. 2016. Approximation of eigenfunctions in kernel-based spaces. *Advances in Computational Mathematics*, 42(4), 973–993.

Shen, H., L. J. Hong, X. Zhang. 2021. Ranking and selection with covariates for personalized decision making. *INFORMS Journal on Computing*, 33(4), 1500–1519.

Stein, M. L. 1999. *Interpolation for Spatial Data: Some Theory for Kriging*. Springer, New York.

Steinwart, I., D. Hush, C. Scovel. 2009. Optimal rates for regularized least squares regression. In *Proc. 22nd Annual Conference on Learning Theory*, 79–93.

van der Vaart, A. W., J. H. van Zanten. 2011. Information rates of nonparametric Gaussian process methods. *J Mach Learn Res*, 12, 2095–2119.

Van Trees, H. L. 2001. *Detection, Estimation, and Modulation Theory*. John Wiley & Sons.

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

Wilson, A., H. Nickisch. 2015. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *International Conference on Machine Learning*, 1775–1784.

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

Zhang, Y., J. C. Duchi, M. J. Wainwright. 2015. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *J Mach Learn Res*, 16, 3299–3340.

Zhou, E., W. Xie. 2015. Simulation optimization when facing input uncertainty. In *Proc. 2015 Winter Simulation Conf.*, 3714–3724.