A Uniform Error Bound for Stochastic Kriging: Properties and Implications on Simulation Experimental Design

XI CHEN, Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, United States

YUTONG ZHANG, Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, United States

GUANGRUI XIE, Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, United States

JINGTAO ZHANG, Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, United States

In this work, we propose a method to construct a uniform error bound for the SK predictor. In investigating the asymptotic properties of the proposed uniform error bound, we examine the convergence rate of SK's predictive variance under the supremum norm in both fixed and random design settings. Our analyses reveal that the large-sample properties of SK prediction depend on the design-point sampling scheme and the budget allocation scheme adopted. Appropriately controlling the order of noise variances through budget allocation is crucial for achieving a desirable convergence rate of SK's approximation error, as quantified by the uniform error bound, and for maintaining SK's numerical stability. Moreover, we investigate the impact of noise variance estimation on the uniform error bound's performance theoretically and numerically. We demonstrate the superiority of the proposed uniform bound to the Bonferroni correction-based simultaneous confidence interval under various experimental settings through numerical evaluations.

CCS Concepts: • **Computing methodologies** → **Uncertainty quantification**.

Additional Key Words and Phrases: Simulation Output Analysis, Simulation Theory, Metamodeling, Design of Experiments

1 Introduction

In the recent decades, research on metamodeling techniques for stochastic simulation experiments has received considerable attention from the simulation research community [2, 15, 34]. Several metamodeling methodologies, such as stochastic kriging (SK) [2, 10, 11], have been proposed for approximating the *mean response surface* implied by a stochastic simulation.

Built on Gaussian process regression (GPR), SK is suitable for simulation metamodeling under the impact of heteroscedasticity (i.e., the simulation output variance varies across the input space) [12, 46]. Similar to kriging, a popular GPR-based methodology for design and analysis of deterministic computer experiments [37], a key element of SK prediction is the use of conditional inference based on GP. At each prediction point in the input space, the conditional distribution of a GP is normal with the predictive mean and the predictive variance in

Authors' Contact Information: Xi Chen, Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia, United States; e-mail: xchen6@vt.edu; Yutong Zhang, Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia, United States; e-mail: yutongz@vt.edu; Guangrui Xie, Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia, United States; e-mail: guanx92@vt.edu; Jingtao Zhang, Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia, United States; e-mail: jingtaozhang@vt.edu.

closed form. A pointwise confidence interval of the mean response value at a given prediction point with a prescribed coverage probability can be constructed based on this conditional distribution. In many applications, however, it is often desirable to have a joint confidence region of the mean response values at an arbitrary set of prediction points with a prescribed simultaneous coverage probability. For example, in materials science and engineering, researchers are interested in chemical sensor calibration in a drifting environment via GPR-based statistical inference of the functional relationship between the sensor response and the analyte concentration and the environmental factors with rigorous uncertainty quantification, aiming at obtaining a calibration model of the desired quality [20]. To draw comprehensive and reliable inferences regarding the effectiveness of a treatment on individual patients, investigators in clinical trial studies aim to infer individualized treatment effects based on observational data via multi-task GPR modeling. Simultaneous confidence intervals are sought to provide measures of confidence in the estimates, a crucial aspect for realizing the full potential of precision medicine [1]. To the best of our knowledge, the methodology for constructing uniform error bounds for SK is relatively underdeveloped. Existing approaches to construct a joint confidence region for SK typically rely on bootstrapping [20, 26] and correction methods such as Bonferroni or Šidák [1, 14] and only apply to a finite number of prediction points. It would be desirable to develop a method capable of providing high-probability bounds on the maximum approximation error achieved by the SK predictor across the input space. Such a bound proves especially valuable in addressing predictive uncertainty, providing an overall measure of confidence in the model's predictions across the entire input space.

A closely related topic in nonparametric data analysis is simultaneous confidence bands (SCBs). SCBs play a crucial role in global inference for unknown functions, such as detecting the overall shape of an unknown function or testing whether the curve adheres to specific functional forms that describe the underlying input-output relationship [5, 6, 38]. SCBs were first constructed for nonparametric density estimation [4] and further extended to nonparametric regression in both the random design [22] and the fixed design settings [18]. Recent advancements in nonparametric methodology for constructing SCBs have significantly expanded in terms of techniques and objectives. For example, asymptotically correct SCBs were proposed for estimating either the mean or the variance function or both of a nonparametric regression model [6]. Various SCBs for single-index link functions [21], cumulative distribution functions [43], and functional data [8] also emerged. Adaptive SCBs that automatically adjust to the smoothness of the underlying function were also investigated [7]. Last but not least, various approaches were explored to construct bootstrap-based SCBs for enhancing the small sample size performance [24] or for obtaining the minimum band width [38].

Uniform error bounds for GPR are not new; they have been studied as an essential component of GP-upper confidence bound (GP-UCB) approaches. These methods are used to address stochastic multi-armed bandit problems with a continuous arm set, assuming that the underlying response function of interest is either a sample from a GP or has a bounded reproducing kernel Hilbert space norm [see, e.g., 13, 40]. Recently, there has been a growing interest in establishing uniform error bounds for various GPR models. These bounds are crucial in enhancing our understanding of the convergence rate and uncertainty quantification associated with the respective models. For example, Wang et al. (2020) derived uniform error bounds of the simple and universal kriging predictor built on noiseless observations [47]. Lederer and Umlauft (2019) proposed a uniform error bound for GPR under the impact of homoscedastic noise, where the true underlying function f is assumed to be a realization of a GP with kernel K satisfying some regularity conditions [29]. However, the exploration of a uniform error bound for GPR under heteroscedasticity has been limited.

In this work, we extend the approach of [29] and propose a method to construct a uniform error bound for the SK predictor. In investigating the asymptotic properties of the proposed uniform error bound, we examine the convergence rate of SK's predictive variance under the supremum norm in both fixed and random design settings. Previous work provided theoretical analyses of the predictive variance for GPR modeling with observations subject to zero [41] and homoscedastic noise [28, 44]. Our investigation fills the gap by conducting a counterpart

study of SK's predictive variance, which can be of independent interest. Our analyses reveal that the large-sample properties of SK prediction depend on the design-point set determined by the design-point sampling scheme and the budget allocation scheme adopted. Roughly speaking, adopting a quasi-uniform design facilitates faster predictive variance convergence rates. Given a sufficiently large simulation budget, the smoother the underlying function of interest is and the lower the input-space dimensionality is, the fewer distinct design points are required, but more replications are needed at each, and vice versa. Moreover, this work reveals that appropriately controlling the order of noise variances through budget allocation is crucial for achieving a desirable convergence rate of SK's approximation error, as quantified by the uniform error bound, and for maintaining SK's numerical stability. Last but not least, we evaluate the impact of noise variance estimation on the uniform error bound's performance theoretically and numerically.

The rest of the paper is organized as follows. Section 2 provides a review of SK and a summary of key notation used in this work. Section 3 presents the main results on the proposed uniform error bound for the SK predictor. Section 4 analyzes the convergence rate of SK's predictive variance under the supremum norm in the fixed design setting and studies the corresponding convergence order of the uniform error bound. The discussion of the random design setting is deferred to Appendix C. Section 5 reveals some implications on simulation experimental designs for SK and theoretically investigates the impact of noise variance estimation on the uniform error bound. Section 6 conducts numerical experiments to demonstrate the proposed uniform bound's performance. Finally, Section 7 concludes this work with a discussion of our major findings and avenues for future research.

2 Background

This section offers a concise review of SK and establishes some notation that will be useful in the rest of the work.

2.1 Review of Stochastic Kriging

In SK, the system output obtained at a point $\mathbf{x}$ in the d-dimensional input space $\mathcal{X} \subseteq \mathbb{R}^d$ on the jth simulation replication $\mathcal{Y}_j(\mathbf{x})$ is modeled as

$$\mathcal{Y}_j(\mathbf{x}) = f(\mathbf{x}) + \varepsilon_j(\mathbf{x}), \tag{1}$$

where $f(\mathbf{x})$ denotes the unknown true mean response that we intend to estimate at $\mathbf{x} \in \mathcal{X}$. For ease of exposition, we assume that $f : \mathcal{X} \mapsto \mathbb{R}$ represents a sample of a stationary mean-zero GP [37], in line with a vast array of GP literature [13, 32, 40, 44]. The spatial covariance between any two points in the GP is given by $\Sigma_{\mathsf{M}}(\mathbf{x}, \mathbf{x}') := \xi^2 K(\mathbf{x}, \mathbf{x}')$, where $\Sigma_{\mathsf{M}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ denotes the spatial covariance function, ξ^2 is the spatial variance of the GP, and $K(\mathbf{x}, \mathbf{x}')$ denotes the kernel function satisfying $K(\mathbf{x}, \mathbf{x}') \leq 1$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. The simulation noise terms incurred at $\mathbf{x}$ on different replications, $\varepsilon_j(\mathbf{x})$'s, are assumed to be independent and identically distributed (i.i.d.) normal random variables with mean zero and input-dependent variance $\mathsf{V}(\mathbf{x}) := \mathrm{Var}(\varepsilon_j(\mathbf{x}))$. The normality of $\varepsilon_j(\mathbf{x})$ could be anticipated since in a discrete-event simulation, the output $\mathcal{Y}_j(\mathbf{x})$ typically represents the average of a large number of more basic random variables obtained on the jth simulation replication.

Given a fixed simulation budget B to expend for approximating the mean response surface via SK, an experimental design for performing the simulation runs can be given as $\{(\mathbf{x}_i, n_i)_{i=1}^k : \sum_{i=1}^k n_i = B\}$, where k denotes the number of distinct design points selected from the input space $\mathcal{X}$, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ denote the k design points in the design-point set $\mathcal{D} := \{\mathbf{x}_i, i = 1, 2, \ldots, k\}$, and n_i represents the number of replications to apply at $\mathbf{x}_i$, $i = 1, 2, \ldots, k$. Based on the simulation dataset $\mathbb{D}_{k,B} = \{\mathbf{x}_i, \{\mathcal{Y}_j(\mathbf{x}_i)\}_{j=1}^{n_i}, i = 1, 2, \ldots, k : \sum_{i=1}^k n_i = B\}$ obtained, one can obtain the SK predictor of $f(\mathbf{x}_0)$ at any $\mathbf{x}_0 \in \mathcal{X}$ as follows:

$$\mu_{k,B}(\mathbf{x}_0) = \Sigma_{\mathsf{M}}(\mathbf{x}_0, \mathbf{X})^\top \left(\Sigma_{\mathsf{M}}(\mathbf{X}, \mathbf{X}) + \Sigma_\varepsilon \right)^{-1} \bar{\mathcal{Y}}; \tag{2}$$

and its corresponding predictive variance is given by

$$\sigma_{k,B}^2(\mathbf{x}_0) = \Sigma_{\mathsf{M}}(\mathbf{x}_0, \mathbf{x}_0) - \Sigma_{\mathsf{M}}(\mathbf{x}_0, \mathbf{X})^\top \left(\Sigma_{\mathsf{M}}(\mathbf{X}, \mathbf{X}) + \Sigma_\varepsilon \right)^{-1} \Sigma_{\mathsf{M}}(\mathbf{x}_0, \mathbf{X}), \tag{3}$$

where $\mathbf{X} = \left(\mathbf{x}_1^\top, \mathbf{x}_2^\top, \ldots, \mathbf{x}_k^\top\right)^\top$ denotes the $k \times d$ design matrix; $\bar{\mathcal{Y}} = \left(\bar{\mathcal{Y}}(\mathbf{x}_1), \bar{\mathcal{Y}}(\mathbf{x}_2), \ldots, \bar{\mathcal{Y}}(\mathbf{x}_k)\right)^\top$ denotes the $k \times 1$ vector of the sample averages of simulation outputs, with $\bar{\mathcal{Y}}(\mathbf{x}_i) = f(\mathbf{x}_i) + \bar{\varepsilon}(\mathbf{x}_i)$ and $\bar{\varepsilon}(\mathbf{x}_i)$ denoting the average random noise incurred at $\mathbf{x}_i$, $i = 1, 2, \ldots, k$. With a slight abuse of notation, we use $\Sigma_M(\mathbf{X}, \mathbf{X})$ to denote the $k \times k$ covariance matrix that records the spatial covariances across the k design points and $\Sigma_M(\mathbf{x}_0, \mathbf{X})$ to represent the $k \times 1$ vector that contains the spatial covariances between the k design points and the prediction point $\mathbf{x}_0$; and the kernel matrix $K(\mathbf{X}, \mathbf{X})$ and the kernel vector $K(\mathbf{x}_0, \mathbf{X})$ can be understood in the same manner. The $k \times k$ diagonal matrix Σ_ε denotes the variance-covariance matrix of the $k \times 1$ vector of average random noise terms $\bar{\varepsilon} = (\bar{\varepsilon}(\mathbf{x}_1), \bar{\varepsilon}(\mathbf{x}_2), \ldots, \bar{\varepsilon}(\mathbf{x}_k))^\top$ given by $\Sigma_\varepsilon = \mathrm{diag}\left(V(\mathbf{x}_1)/n_1, V(\mathbf{x}_2)/n_2, \ldots, V(\mathbf{x}_k)/n_k\right)$.

A pointwise confidence interval (CI) of $f(\mathbf{x}_0)$ at a given prediction point $\mathbf{x}_0 \in \mathcal{X}$ with a prescribed confidence level $(1 - \alpha)$ can be constructed as $\mu_{k,B}(\mathbf{x}_0) \pm z_{1-\alpha/2}\sigma_{k,B}(\mathbf{x}_0)$, where $\alpha \in (0, 1)$ and $z_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$-quantile of the standard normal distribution [25]. One way to construct a confidence region for $f(\cdot)$ at N prediction points (say, $\mathbf{x}_{0,i} \in \mathcal{X}$, $i = 1, 2, \ldots, N$) with confidence level $(1 - \alpha)$ is to apply the Bonferroni correction and obtain a simultaneous confidence bound comprising N pointwise confidence intervals [26]: $\mu_{k,B}(\mathbf{x}_{0,i}) \pm z_{1-\alpha/(2N)}\sigma_{k,B}(\mathbf{x}_{0,i})$, $i = 1, 2, \ldots, N$. Nevertheless, the literature on constructing a joint confidence bound for SK over an arbitrary set of prediction points is sparse.

2.2 Notation

In the remainder of this work, we employ the following notation consistently. Given a function $g : \mathcal{X} \mapsto \mathbb{R}$, let $\|g\|_\infty$ denote its supremum norm. For a vector $\mathbf{v}$, $\|\mathbf{v}\|$ denotes the Euclidean norm of $\mathbf{v}$. Let $\mathbf{I}_k$ denote the $k \times k$ identity matrix. For any symmetric matrix $\mathbf{Q}$, the ℓ_2-operator norm or spectral norm of $\mathbf{Q}$ is denoted as $\|\mathbf{Q}\|$, defined as $\|\mathbf{Q}\| := \lambda_{\max}(\mathbf{Q})$, where $\lambda_{\max}(\mathbf{Q})$ and $\lambda_{\min}(\mathbf{Q})$ respectively denote the maximum and minimum eigenvalues of $\mathbf{Q}$. Define $V_{k,\max} := \max_{1 \le i \le k} V(\mathbf{x}_i)/n_i$ and $V_{k,\min} := \min_{1 \le i \le k} V(\mathbf{x}_i)/n_i$ as the maximum and minimum noise variance at the k design points. Let $\lesssim$ and $\gtrsim$ denote inequalities up to a constant multiple and write $a \asymp b$ when both $a \lesssim b$ and $a \gtrsim b$ hold.

3 Uniform Error Bound for Stochastic Kriging

This section presents the main results regarding the construction of the proposed high-probability uniform error bound and then reveals its asymptotic order as the number of design points k and the budget B allocated approach infinity.

To deepen understanding, we sketch the idea behind constructing the uniform error bound, which essentially involves two main steps. First, consider any grid of points covering the input space $\mathcal{X}$ with the grid constant τ, denoted as $\mathbb{X}_\tau$, satisfying $\max_{\mathbf{x} \in \mathcal{X}} \min_{\mathbf{x}' \in \mathbb{X}_\tau} \|\mathbf{x} - \mathbf{x}'\| \le \tau$. Conceptually, we can visualize such a cover as a collection of balls of radius τ covering the input space $\mathcal{X}$. We establish a high-probability uniform bound for the approximation error $|\mu_{k,B}(\mathbf{x}') - f(\mathbf{x}')|$ that holds simultaneously at all points $\mathbf{x}'$ in the grid $\mathbb{X}_\tau$ by leveraging properties of the Gaussian distribution, which follows from model (1) and Expression (2) and accounts for both sampling and response-surface uncertainty. This step introduces a dependency on the number of points in the grid and is accounted for by the coefficient $\beta(\tau)$ to be defined in (5) which involves the covering number $M(\tau, \mathcal{X})$—the minimum number of points in $\mathbb{X}_\tau$ to cover $\mathcal{X}$ given the grid constant τ. The second step is to extend such an error bound from the points in the grid to any point in the input space $\mathcal{X}$, yielding the high-probability uniform error bound for SK. This extension is achieved by establishing the Lipschitz continuity of the predictive mean $\mu_{k,B}(\cdot)$ and the modulus of continuity $\omega_{\sigma_{k,B}}(\cdot)$ for the predictive standard deviation $\sigma_{k,B}(\cdot)$ based on the Lipschitz continuity and differentiability of the covariance function $\Sigma_M(\cdot, \cdot)$.

Theorem 1. *Consider a zero mean Gaussian process defined through the continuous covariance function $\Sigma_M(\cdot, \cdot)$ with Lipschitz constant L_Σ on the compact set $\mathcal{X} \subset \mathbb{R}^d$. Also consider a continuous unknown function $f : \mathcal{X} \to \mathbb{R}$ with*

Lipschitz constant L_f, and its noisy observations $\mathcal{Y}(\cdot)$'s satisfying the assumptions stipulated under model (1). *Then, the predictive mean function $\mu_{k,B}(\cdot)$ and the predictive standard deviation $\sigma_{k,B}(\cdot)$ obtained based on a simulation dataset $\mathbb{D}_{k,B} = \{\mathbf{x}_i, \{\mathcal{Y}_j(\mathbf{x}_i)\}_{j=1}^{n_i}, i = 1, 2, \ldots, k : \sum_{i=1}^{k} n_i = B\}$ are continuous with Lipschitz constant $L_{\mu_{k,B}}$ and modulus of continuity $\omega_{\sigma_{k,B}}(\cdot)$ on $\mathcal{X}$, which are respectively given by*

$$
\begin{aligned}
L_{\mu_{k,B}} &= L_\Sigma \sqrt{k} \left\| (\Sigma_M(\mathbf{X}, \mathbf{X}) + \Sigma_\varepsilon)^{-1} \bar{\mathcal{Y}} \right\|, \\
\omega_{\sigma_{k,B}}(h) &= \sqrt{2h L_\Sigma \left(1 + k \left\| (\Sigma_M(\mathbf{X}, \mathbf{X}) + \Sigma_\varepsilon)^{-1} \right\| \max_{\tilde{\mathbf{x}}, \tilde{\mathbf{x}}' \in \mathcal{X}} \Sigma_M(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') \right)},
\end{aligned}
\tag{4}
$$

where L_Σ, the Lipschitz constant of the covariance function $\Sigma_M(\cdot, \cdot)$, is defined as

$$
L_\Sigma := \max_{\tilde{\mathbf{x}}, \tilde{\mathbf{x}}' \in \mathcal{X}} \left\| \left(\frac{\partial \Sigma_M(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')}{\partial \tilde{x}_1} \ \frac{\partial \Sigma_M(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')}{\partial \tilde{x}_2} \ \cdots \ \frac{\partial \Sigma_M(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')}{\partial \tilde{x}_d} \right)^\top \right\|,
$$

and $\omega_{\sigma_{k,B}}(\cdot)$ measures how much the value of $\sigma_{k,B}(\cdot)$ can vary when moving from one point to another within a given distance $h > 0$. Given any $\alpha \in (0, 1)$, choose $\tau \in \mathbb{R}_+$ and set

$$
\beta(\tau) = 2 \log \left(M(\tau, \mathcal{X})/\alpha \right) \text{ and } \gamma_{k,B}(\tau) = (L_{\mu_{k,B}} + L_f)\tau + \sqrt{\beta(\tau)} \omega_{\sigma_{k,B}}(\tau),
\tag{5}
$$

then it holds true that

$$
P\left(|\mu_{k,B}(\mathbf{x}) - f(\mathbf{x})| \leq \sqrt{\beta(\tau)} \sigma_{k,B}(\mathbf{x}) + \gamma_{k,B}(\tau), \forall \mathbf{x} \in \mathcal{X} \right) \geq 1 - \alpha,
\tag{6}
$$

where L_f denotes a Lipschitz constant of $f(\cdot)$ that holds with probability of at least $1 - \alpha_L$ with $\alpha_L \in (0, 1)$ and can be given as

$$
L_f := \left\| \begin{bmatrix} \sqrt{2 \log \left(\frac{2d}{\alpha_L} \right)} \max_{\mathbf{x} \in \mathcal{X}} \sqrt{\Sigma_M^{\partial 1}(\mathbf{x}, \mathbf{x})} + 12\sqrt{6d} \max \left\{ \max_{\mathbf{x} \in \mathcal{X}} \sqrt{\Sigma_M^{\partial 1}(\mathbf{x}, \mathbf{x})}, \sqrt{r L_\Sigma^{\partial 1}} \right\} \\ \vdots \\ \sqrt{2 \log \left(\frac{2d}{\alpha_L} \right)} \max_{\mathbf{x} \in \mathcal{X}} \sqrt{\Sigma_M^{\partial d}(\mathbf{x}, \mathbf{x})} + 12\sqrt{6d} \max \left\{ \max_{\mathbf{x} \in \mathcal{X}} \sqrt{\Sigma_M^{\partial d}(\mathbf{x}, \mathbf{x})}, \sqrt{r L_\Sigma^{\partial d}} \right\} \end{bmatrix} \right\|;
$$

$L_\Sigma^{\partial i}$ denotes the Lipschitz constant of the partial derivative kernel $\Sigma_M^{\partial i}(\mathbf{x}, \mathbf{x})$ on the set $\mathcal{X}$ for $i = 1, 2, \ldots, d$, and $r := \max_{\tilde{\mathbf{x}}, \tilde{\mathbf{x}}' \in \mathcal{X}} \|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}'\|$.

Proof. Theorem 1 is inspired by Theorem 3.1 in [29], whose proof is in the same vein as those of Theorems 3.1 and 3.2 in [29]. Specifically, we replace the homogeneous noise variance-covariance matrix $\sigma^2 \mathbf{I}_k$ by the heteroscedastic noise variance-covariance matrix Σ_ε in the proof of Theorem 3.1. The Lipschitz constant L_f can be obtained following the proof of Theorem 3.2. □

Some comments follow immediately. First and foremost, we can construct a uniform bound for the function $f(\cdot)$ in light of Theorem 1: it holds with probability at least $1 - \alpha$ that

$$
LB(\mathbf{x}) \leq f(\mathbf{x}) \leq UB(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X},
\tag{7}
$$

where $LB(\mathbf{x}) := \mu_{k,B}(\mathbf{x}) - \sqrt{\beta(\tau)} \sigma_{k,B}(\mathbf{x}) - \gamma_{k,B}(\tau)$ and $UB(\mathbf{x}) := \mu_{k,B}(\mathbf{x}) + \sqrt{\beta(\tau)} \sigma_{k,B}(\mathbf{x}) + \gamma_{k,B}(\tau)$ are respectively the lower and upper limits of the bound at $\mathbf{x} \in \mathcal{X}$. Second, the grid constant τ is a parameter used in the derivation of Theorem 1. Given a particular value of τ, obtaining a closed-form expression for the covering number $M(\tau, \mathcal{X})$ can be challenging. However, it is typically convenient to upper bound $M(\tau, \mathcal{X})$. For example, for a hypercubic set $\mathcal{X} \subseteq \mathbb{R}^d$, an upper bound of $M(\tau, \mathcal{X})$ can be given as $(1 + r/\tau)^d$, where r denotes the edge length of the hypercube. Notice that $\beta(\tau)$ and $\gamma_{k,B}(\tau)$ in (5) can be obtained analytically given a simulation dataset $\mathbb{D}_{k,B}$ and the covariance

function $\Sigma_M(\cdot, \cdot)$. Therefore, a probabilistic uniform error bound that holds true simultaneously for all $\mathbf{x} \in \mathcal{X}$ can be constructed with a prescribed error level α and a choice of the grid constant τ.

The following result elucidates the order of the approximation error, quantified by the uniform error bound, as k and B approach infinity. Below we make the choice of grid constant τ depend on the number of design points k, so $\beta(\tau)$ and $\gamma_{k,B}(\tau)$ depend on k as well. The notation $\tau(k)$, $\beta_k(\tau)$, and $\gamma_{k,B}(\tau)$ will be used hereinafter whenever it is not cumbersome to emphasize the terms' dependence on the number of design points k.

Theorem 2. *Suppose that the assumptions of Theorem 1 are satisfied. The experimental design ensures that as the number of design points k increases, the number of observations obtained at each existing design point also grows. Furthermore, for any fixed $k \geq 1$, the numbers of replications allocated over the k design points satisfy that $V_{k,\max}/V_{k,\min} < \infty$. Assume that the absolute value of the unknown function $f : \mathcal{X} \to \mathbb{R}$ of interest is bounded above by $\bar{f} \in \mathbb{R}_+$. Given any $\alpha \in (0, 1)$, it holds that*

$$P \left(\sup_{\mathbf{x} \in \mathcal{X}} |\mu_{k,B}(\mathbf{x}) - f(\mathbf{x})| = O\left(\left(\beta_k(\tau) \cdot \tau(k) \cdot k V_{k,\min}^{-1} \right)^{\frac{1}{2}} + \tau(k) + \tau(k) \cdot k V_{k,\min}^{-\frac{1}{2}} + \sqrt{\beta_k(\tau)} \sup_{\mathbf{x} \in \mathcal{X}} \sigma_{k,B}(\mathbf{x}) \right) \right) \geq 1 - \alpha.$$

Proof. Due to Theorem 1 with $\beta_k(\tau) = 2 \log \left(M(\tau(k), \mathcal{X})/(\pi_k \alpha) \right)$ such that $\sum_{k=1}^{\infty} \pi_k = 1/2$ and the union bound over all $k \geq 1$, the following event holds true with probability of at least $1 - \alpha/2$:

$$\sup_{\mathbf{x} \in \mathcal{X}} |\mu_{k,B}(\mathbf{x}) - f(\mathbf{x})| \leq \sqrt{\beta_k(\tau)} \sup_{\mathbf{x} \in \mathcal{X}} \sigma_{k,B}(\mathbf{x}) + \gamma_{k,B}(\tau), \ \forall k \geq 1. \tag{8}$$

A trivial bound for the covering number can be obtained by considering a uniform grid over the hypercube containing $\mathcal{X}$. This leads to $M(\tau(k), \mathcal{X}) \leq (1 + r/\tau(k))^d$, where $r = \max_{\tilde{\mathbf{x}}, \tilde{\mathbf{x}}' \in \mathcal{X}} \|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}'\|$. Therefore, we have

$$\beta_k(\tau) \leq 2d \log \left(1 + r/\tau(k) \right) - 2 \log(\pi_k) - 2 \log(\alpha). \tag{9}$$

Furthermore, recall the definition of $L_{\mu_{k,B}}$ in (4) from Theorem 1,

$$L_{\mu_{k,B}} = L_{\Sigma} \sqrt{k} \left\| (\Sigma_M(\mathbf{X}, \mathbf{X}) + \Sigma_\varepsilon)^{-1} \bar{\mathcal{Y}} \right\|.$$

Since the matrix $\Sigma_M(\mathbf{X}, \mathbf{X}) + \Sigma_\varepsilon$ is positive definite and $f(\cdot)$ is bounded above by $\bar{f}$, it follows that

$$\left\| (\Sigma_M(\mathbf{X}, \mathbf{X}) + \Sigma_\varepsilon)^{-1} \bar{\mathcal{Y}} \right\| \leq \left\| \bar{\mathcal{Y}} \right\| / \lambda_{\min} \left(\Sigma_M(\mathbf{X}, \mathbf{X}) + \Sigma_\varepsilon \right) \leq (\sqrt{k} \bar{f} + \|\bar{\varepsilon}\|)/V_{k,\min},$$

where the second inequality follows from the triangular inequality, and recall $\bar{\varepsilon} = (\bar{\varepsilon}(\mathbf{x}_1), \bar{\varepsilon}(\mathbf{x}_2), \ldots, \bar{\varepsilon}(\mathbf{x}_k))^\top$. Given that $\bar{\varepsilon}$ is multivariate normally distributed with mean zero and variance-covariance matrix Σ_ε, $\|\bar{\varepsilon}\|^2$ is equal to $\sum_{i=1}^{k} a_i Z_i^2$ in distribution, where Z_i's are i.i.d. standard normal random variables and $a_i = V(\mathbf{x}_i)/n_i$, for $i = 1, 2, \ldots, k$, i.e., $\|\bar{\varepsilon}\|^2 \overset{\mathcal{D}}{=} \sum_{i=1}^{k} a_i Z_i^2$. Then by Lemma 1 of [27], we have

$$P \left(\|\bar{\varepsilon}\|^2 \geq 2 \left(\sum_{i=1}^{k} \frac{V^2(\mathbf{x}_i)}{n_i^2} \right)^{\frac{1}{2}} \sqrt{\eta_{k,B}} + 2 V_{k,\max} \eta_{k,B} + \sum_{i=1}^{k} \frac{V(\mathbf{x}_i)}{n_i} \right) \leq \exp(-\eta_{k,B})$$

for any $\eta_{k,B} > 0$. Therefore, with probability of at least $1 - \exp(-\eta_{k,B})$, we have

$$\|\bar{\varepsilon}\|^2 \leq 2 \left(\sum_{i=1}^{k} \frac{V^2(\mathbf{x}_i)}{n_i^2} \right)^{\frac{1}{2}} \sqrt{\eta_{k,B}} + 2 V_{k,\max} \eta_{k,B} + \sum_{i=1}^{k} \frac{V(\mathbf{x}_i)}{n_i}.$$

Hence, if we set $\eta_{k,B} = \log\left(1/(\pi_k \alpha)\right)$ so that $\sum_{k=1}^{\infty} \pi_k = 1/2$, then applying the union bounds over all $k \geq 1$ yields

$$\left\|(\Sigma_{\mathsf{M}}(\mathbf{X},\mathbf{X}) + \Sigma_{\varepsilon})^{-1} \, \bar{\mathcal{Y}}\right\| \leq \left(\sqrt{k}\bar{f} + \left(2\left(\sum_{i=1}^{k} \frac{\mathsf{V}^2(\mathbf{x}_i)}{n_i^2}\right)^{\frac{1}{2}} \sqrt{\eta_{k,B}} + 2\mathsf{V}_{k,\max}\eta_{k,B} + \sum_{i=1}^{k} \frac{\mathsf{V}(\mathbf{x}_i)}{n_i}\right)^{\frac{1}{2}}\right)\mathsf{V}_{k,\min}^{-1} \tag{10}$$

for all $k \geq 1$ with probability of at least $1 - \alpha/2$. Hence, by Theorem 1, the Lipschitz constant of the predictive mean function $\mu_{k,B}(\cdot)$ satisfies

$$L_{\mu_{k,B}} \leq L_{\Sigma} \sqrt{k} \underbrace{\left(\sqrt{k}\bar{f} + \left(2\left(\sum_{i=1}^{k} \frac{\mathsf{V}^2(\mathbf{x}_i)}{n_i^2}\right)^{\frac{1}{2}} \sqrt{\eta_{k,B}} + 2\mathsf{V}_{k,\max}\eta_{k,B} + \sum_{i=1}^{k} \frac{\mathsf{V}(\mathbf{x}_i)}{n_i}\right)^{\frac{1}{2}}\right)\mathsf{V}_{k,\min}^{-1}}_{:=U_{k,B}}, \ \forall k \geq 1,$$

where $U_{k,B}$ denotes all the terms to the right of L_{Σ} in the inequality above and recall $\eta_{k,B} = \log\left(1/(\pi_k \alpha)\right)$. Since $\left(\sum_{i=1}^{k} \frac{\mathsf{V}^2(\mathbf{x}_i)}{n_i^2}\right)^{\frac{1}{2}} \leq \sqrt{k}\mathsf{V}_{k,\max}$, $\sum_{i=1}^{k} \frac{\mathsf{V}(\mathbf{x}_i)}{n_i} \leq k\mathsf{V}_{k,\max}$, and $\mathsf{V}_{k,\max}/\mathsf{V}_{k,\min} < \infty$, and $\eta_{k,B}$ grows slowly (typically logarithmically with k for some commonly used $\{\pi_k\}$ sequences), it follows that $L_{\mu_{k,B}} \lesssim U_{k,B} \asymp k\mathsf{V}_{k,\min}^{-\frac{1}{2}}$ with probability of at least $1 - \alpha/2$.

The modulus of continuity $\omega_{\sigma_{k,B}}(\cdot)$ of the predictive standard deviation defined in (4) satisfies

$$\omega_{\sigma_{k,B}}(\tau) \leq \left[2L_{\Sigma}\tau(k)\left(k \max_{\tilde{\mathbf{x}},\tilde{\mathbf{x}}' \in \mathcal{X}} \Sigma_{\mathsf{M}}(\tilde{\mathbf{x}},\tilde{\mathbf{x}}')\mathsf{V}_{k,\min}^{-1} + 1\right)\right]^{\frac{1}{2}}, \tag{11}$$

since $\left\|(\Sigma_{\mathsf{M}}(\mathbf{X},\mathbf{X}) + \Sigma_{\varepsilon})^{-1}\right\| = \lambda_{\max}\left((\Sigma_{\mathsf{M}}(\mathbf{X},\mathbf{X}) + \Sigma_{\varepsilon})^{-1}\right) = 1/\lambda_{\min}\left(\Sigma_{\mathsf{M}}(\mathbf{X},\mathbf{X}) + \Sigma_{\varepsilon}\right) \leq 1/\lambda_{\min}\left(\Sigma_{\varepsilon}\right) = 1/\mathsf{V}_{k,\min}$, where we have used $\lambda_{\min}\left(\Sigma_{\mathsf{M}}(\mathbf{X},\mathbf{X}) + \Sigma_{\varepsilon}\right) \geq \lambda_{\min}\left(\Sigma_{\mathsf{M}}(\mathbf{X},\mathbf{X})\right) + \lambda_{\min}\left(\Sigma_{\varepsilon}\right) \geq \lambda_{\min}\left(\Sigma_{\varepsilon}\right)$ due to Weyl's Theorem. Therefore, it follows that $\omega_{\sigma_{k,B}}(\tau) \lesssim \left(\tau(k) \cdot k\mathsf{V}_{k,\min}^{-1}\right)^{\frac{1}{2}}$.

Due to the union bound, we have that $\gamma_{k,B}(\tau)$ in (8) can be bounded above with probability of at least $1 - \alpha$ as follows:

$$\gamma_{k,B}(\tau) = \sqrt{\beta(\tau)}\omega_{\sigma_{k,B}}(\tau) + (L_{\mu_{k,B}} + L_f)\tau(k)$$

$$\leq \left[\beta_k(\tau) \cdot 2L_{\Sigma}\tau(k)\left(k \max_{\tilde{\mathbf{x}},\tilde{\mathbf{x}}' \in \mathcal{X}} \Sigma_{\mathsf{M}}(\tilde{\mathbf{x}},\tilde{\mathbf{x}}')/\mathsf{V}_{k,\min} + 1\right)\right]^{\frac{1}{2}} + L_f\tau(k) + L_{\Sigma}U_{k,B}\tau(k)$$

$$\lesssim \left(\beta_k(\tau) \cdot \tau(k) \cdot k\mathsf{V}_{k,\min}^{-1}\right)^{\frac{1}{2}} + \tau(k) + \tau(k) \cdot k\mathsf{V}_{k,\min}^{-\frac{1}{2}}. \tag{12}$$

Therefore, it follows from (8) and (12) that

$$\mathrm{P}\left(\sup_{\mathbf{x} \in \mathcal{X}} |\mu_{k,B}(\mathbf{x}) - f(\mathbf{x})| = O\left(\left(\beta_k(\tau) \cdot \tau(k) \cdot k\mathsf{V}_{k,\min}^{-1}\right)^{\frac{1}{2}} + \tau(k) + \tau(k) \cdot k\mathsf{V}_{k,\min}^{-\frac{1}{2}} + \sqrt{\beta_k(\tau)}\sup_{\mathbf{x} \in \mathcal{X}} \sigma_{k,B}(\mathbf{x})\right)\right) \geq 1 - \alpha.$$

□

We briefly comment on Theorem 2, which sets up the foundation for the subsequent analysis. First and foremost, the upper bound of the approximation error achieved at any $\mathbf{x} \in \mathcal{X}$, $\sqrt{\beta_k(\tau)}\sigma_{k,B}(\mathbf{x}) + \gamma_{k,B}(\tau)$, consists of two parts. From (9), we observe that the smaller the grid constant $\tau(k)$, the greater the coefficient $\beta_k(\tau)$. To guarantee a vanishing approximation error for $\forall \mathbf{x} \in \mathcal{X}$, the choice of $\tau(k)$ must ensure that $\gamma_{k,B}(\tau) \to 0$ and

$\sup_{\mathbf{x} \in \mathcal{X}} \sqrt{\beta_k(\tau)} \sigma_{k,B}(\mathbf{x}) \to 0$ as the number of design points k and the total simulation budget B approach infinity. It is worth noting that the grid constant $\tau(k)$ can be chosen arbitrarily small, making the magnitude of $\gamma_{k,B}(\tau)$ negligible compared to $\sqrt{\beta_k(\tau)} \sigma_{k,B}(\mathbf{x})$ for $\forall \mathbf{x} \in \mathcal{X}$. We will use this fact to study the asymptotic performance of the uniform error bound as $k, B \to \infty$. Second, given a fixed total budget B to allocate at k distinct design points, the budget allocation scheme adopted impacts the width of the uniform error bound through $\sigma_{k,B}(\mathbf{x})$ and $\gamma_{k,B}(\tau)$, indicated by (8) and the upper bounds of $\omega_{\sigma_{k,B}}(\tau)$ and $L_{\mu_{k,B}}$. In particular, efficient unequal budget allocation schemes can potentially reduce the magnitudes of $\sigma_{k,B}(\mathbf{x})$ and $\gamma_{k,B}(\tau)$ compared to the equal allocation scheme, leading to a tighter uniform error bound. However, if $\tau(k)$ is sufficiently small, the impact of budget allocation schemes may not be that obvious; see (12). Nonetheless, the explicit analysis of these impacts is challenging, and we resort to numerical studies in Section 6 for a thorough investigation. Last but not least, understanding the convergence rate of $\sup_{\mathbf{x} \in \mathcal{X}} \sigma_{k,B}(\mathbf{x})$ is crucial to determine the order of the vanishing approximation error. We will conduct an in-depth investigation of this in Section 4.

4 Analysis of the Predictive Variance and the Uniform Error Bound in the Fixed Design Setting

In this section, we analyze the convergence rate of the predictive variance $\sigma^2_{k,B}(\cdot)$ under the supremum norm in the fixed design setting and subsequently elucidate the corresponding order of the proposed uniform error bound. Due to space limitations, the discussion of the random design setting is deferred to Appendix C.

We concentrate on the fixed design setting in this section, commonly employed in the design and analysis of computer experiments [37]. In this setting, the design points are chosen according to specific sampling schemes of our choice. Such designs encompass a variety of approaches, including, but not limited to, grid designs, low-discrepancy sequences, maximin Latin hypercube designs, and optimal Latin hypercube designs. Our analysis of the convergence rate of SK's predictive variance in the fixed design setting leverages results from the scattered data analysis literature, where the goal is to approximate or interpolate an underlying fixed function; see, e.g., [36] and [49].

Notice that the predictive variance $\sigma^2_{k,B}(\mathbf{x}_0)$ can be upper bounded as follows:

$$\sigma^2_{k,B}(\mathbf{x}_0) \le \Sigma_{\mathsf{M}}(\mathbf{x}_0, \mathbf{x}_0) - \Sigma_{\mathsf{M}}(\mathbf{x}_0, \mathbf{X})^{\top} \left(\Sigma_{\mathsf{M}}(\mathbf{X}, \mathbf{X}) + \mathsf{V}_{k,\max} \mathbf{I}_k \right)^{-1} \Sigma_{\mathsf{M}}(\mathbf{x}_0, \mathbf{X}), \tag{13}$$

where recall $\mathsf{V}_{k,\max} := \max_{1 \le i \le k} \mathsf{V}(\mathbf{x}_i)/n_i$. *Define $\lambda := \mathsf{V}_{k,\max}/\xi^2$ hereinafter in this section.* We can rewrite the upper bound for $\sigma^2_{k,B}(\mathbf{x}_0)$ given by the right-hand side of (13) as $\xi^2 \left(K_{\mathbf{x}_0}(\mathbf{x}_0) - \widehat{K}_{\mathbf{x}_0}(\mathbf{x}_0) \right)$, where $K_{\mathbf{x}_0}(\mathbf{x}') := K(\mathbf{x}_0, \mathbf{x}')$ for any $\mathbf{x}' \in \mathcal{X}$ and recall from Section 2.1 that $K(\mathbf{x}, \mathbf{x}') = \Sigma_{\mathsf{M}}(\mathbf{x}, \mathbf{x}')/\xi^2$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ is the kernel function; in addition,

$$\widehat{K}_{\mathbf{x}_0}(\mathbf{x}') = K(\mathbf{x}', \mathbf{X})^{\top} \left(K(\mathbf{X}, \mathbf{X}) + \lambda \mathbf{I}_k \right)^{-1} K(\mathbf{X}, \mathbf{x}_0).$$

Inspired by [36] and [49], we realize that, for any given prediction point $\mathbf{x}_0 \in \mathcal{X}$, $\widehat{K}_{\mathbf{x}_0}(\cdot)$ is the solution to the following regularized kernel based approximation using the noiseless observations of the function $f := K_{\mathbf{x}_0}$ at the k design points in the design-point set $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\} \subseteq \mathcal{X}$:

$$s_\lambda = \widehat{K}_{\mathbf{x}_0} := \underset{s \in \mathcal{H}}{\operatorname{argmin}} \left\{ \sum_{j=1}^{k} \left(f(\mathbf{x}_j) - s(\mathbf{x}_j) \right)^2 + \lambda \|s\|^2_{\mathcal{H}} \right\}, \tag{14}$$

where the solution $s_\lambda(\cdot)$ depends on the choice of the regularization parameter $\lambda = \mathsf{V}_{k,\max}/\xi^2$, $\mathcal{H}$ denotes the reproducing kernel Hilbert space (RKHS) associated with kernel K, and $\|\cdot\|_{\mathcal{H}}$ denotes the norm equipped with $\mathcal{H}$ (see a brief introduction to RKHS in Appendix B). Therefore, we can analyze the error achieved by the estimator, specifically $\|f - s_\lambda\|_\infty = \|K_{\mathbf{x}_0} - \widehat{K}_{\mathbf{x}_0}\|_\infty$, to investigate the convergence properties of the predictive variance $\sigma^2_{k,B}(\cdot)$ under the supremum norm. To begin with, we introduce some useful definitions next.

Definition 1. *For $v \in \mathbb{N}$, $q \in [1, \infty]$, and a domain $\tilde{X} \subseteq \mathbb{R}^d$ which is non-empty, open, and connected, the integer order Sobolev space $W_q^v(\tilde{X})$ is defined as*

$$W_q^v(\tilde{X}) = \{f \in L_q(\tilde{X}) : \forall \boldsymbol{\alpha} \in \mathbb{N}^d, |\boldsymbol{\alpha}| \leq v, D^{\boldsymbol{\alpha}} f \in L_q(\tilde{X})\},$$

where $\mathbb{N}^d$ denotes the set of multi-indices of size d, $|\boldsymbol{\alpha}| := \sum_{i=1}^d \alpha_i$ for a multi-index $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d) \in \mathbb{N}^d$, and $D^{\boldsymbol{\alpha}}$ denotes the weak derivative operator corresponding to $\boldsymbol{\alpha}$ which is defined as $D^{\boldsymbol{\alpha}} f = \frac{\partial^{|\boldsymbol{\alpha}|}}{\partial \alpha_1 \partial \alpha_2 \ldots \partial \alpha_d} f$.

For any function $f \in W_q^v(\tilde{X})$, the (semi-)norms are given by

$$|f|_{W_q^v(\tilde{X})} := \left(\sum_{|\boldsymbol{\alpha}|=v} \|D^{\boldsymbol{\alpha}} f\|_{L_q(\tilde{X})}^q \right)^{1/q} \quad and \quad \|f\|_{W_q^v(\tilde{X})} := \left(\sum_{|\boldsymbol{\alpha}| \leq v} \|D^{\boldsymbol{\alpha}} f\|_{L_q(\tilde{X})}^q \right)^{1/q}, \; if\ q \in [1, \infty),$$

$$|f|_{W_\infty^v(\tilde{X})} := \sup_{|\boldsymbol{\alpha}|=v} \|D^{\boldsymbol{\alpha}} f\|_{L_\infty(\tilde{X})} \quad and \quad \|f\|_{W_\infty^v(\tilde{X})} := \sup_{|\boldsymbol{\alpha}| \leq v} \|D^{\boldsymbol{\alpha}} f\|_{L_\infty(\tilde{X})}, \; if\ q = \infty,$$

with $\| \cdot \|_{L_q(\tilde{X})}$ denoting the usual L_q norm.

Sobolev spaces of fractional orders ($v \notin \mathbb{N}$) are defined analogously; see [3, 49] for details. In particular, for $v > d/2$, the Sobolev space $W_2^v(\mathbb{R}^d)$ can be written as $W_2^v(\mathbb{R}^d) = \left\{ f \in L_2(\mathbb{R}^d) : \|f\|_{W_2^v(\mathbb{R}^d)} < \infty \right\}$. The Sobolev spaces on a subset (e.g., compact and convex) $X \subset \mathbb{R}^d$, $W_2^v(X)$, can be defined as $W_2^v(X) := \left\{ f : X \to \mathbb{R}^d : \exists \check{f} \in W_2^v(\mathbb{R}^d) \text{ such that } \check{f}(\mathbf{x}) = f(\mathbf{x}), \forall \mathbf{x} \in X \right\}$, with norm $\|f\|_{W_2^v(X)} := \inf \left\{ \|\check{f}\|_{W_2^v(\mathbb{R}^d)} : \check{f} \in W_2^v(\mathbb{R}^d) \text{ and } \check{f}(\mathbf{x}) = f(\mathbf{x}), \forall \mathbf{x} \in X \right\}$. The RKHS $\mathcal{H}$ associated with kernel K is said to be norm-equivalent to the Sobolev space $W_2^v(X)$, if there exists constants $c_1, c_2 > 0$ such that

$$c_1 \|f\|_{W_2^v(X)} \leq \|f\|_{\mathcal{H}} \leq c_2 \|f\|_{W_2^v(X)}, \quad \forall f \in \mathcal{H}.$$

That is, $\mathcal{H}$ is equivalent to $W_2^v(X)$ as a set of functions. In this case, the kernel function K is called v-smooth. One popular kernel function that is v-smooth is the Matérn kernel as given below.

For $v > d/2$, the Matérn kernel is given by

$$K(\mathbf{x}, \mathbf{x}') = \frac{2^{1-(v-d/2)}}{\Gamma(v-d/2)} \left(\sqrt{2(v-d/2)} \frac{\|\mathbf{x} - \mathbf{x}'\|}{l} \right)^{v-d/2} \mathcal{B}_{v-d/2} \left(\sqrt{2(v-d/2)} \frac{\|\mathbf{x} - \mathbf{x}'\|}{l} \right), \tag{15}$$

where Γ is the Gamma function, $\mathcal{B}_{v-d/2}$ is the modified Bessel function of the second kind of order $v - d/2$, and $l > 0$ is the lengthscale parameter. The parameter v controls the smoothness level of functions in the corresponding RKHS; the smaller its value, the rougher the functions.

Definition 2. *Given the bounded input space $X \subseteq \mathbb{R}^d$ and the design-point set $\mathcal{D} = \{\mathbf{x}_i, i = 1, 2, \ldots, k\} \subseteq X$, the fill distance $h_{\mathcal{D}, X}$ and the separation distance $q_{\mathcal{D}}$ are defined as*

$$h_{\mathcal{D}, X} := \sup_{\mathbf{x} \in X} \min_{1 \leq j \leq k} \|\mathbf{x} - \mathbf{x}_j\|, \quad q_{\mathcal{D}} := \frac{1}{2} \min_{\substack{1 \leq i, j \leq k \\ i \neq j}} \|\mathbf{x}_i - \mathbf{x}_j\|.$$

The design-point set $\mathcal{D}$ is said to be quasi-uniform with respect to a constant $c_{qu} > 0$ if $q_{\mathcal{D}} \leq h_{\mathcal{D}, X} \leq c_{qu} \cdot q_{\mathcal{D}}$.

REMARK 1. A design-point set $\mathcal{D}$ with a small fill distance $h_{\mathcal{D}, X}$ guarantees that every point in the input space X is close to some design point in $\mathcal{D}$, while $\mathcal{D}$ having a large separation distance $q_{\mathcal{D}}$ indicates that the design points in $\mathcal{D}$ are not clustered. Quasi-uniform designs achieve the optimal rates for the fill distance, namely, $h_{\mathcal{D}, X} \asymp k^{-1/d}, \forall k \in \mathbb{N}$ [48, 51]. For example, the design-point set comprising regular grid points in a hypercube $(0, 1)^d$ is quasi-uniform. Points selected sequentially to minimize the GPR's predictive variance for a v-smooth

kernel with $\nu > (d/2) + 1$ form a quasi-uniform design. It is known that low-discrepancy point sets, e.g., the Halton sequence and Sobol' nets have small fill distances; and sequences such as the Halton sequence have a fill distance $h_{\mathcal{D},\mathcal{X}} \lesssim k^{-1/d} (\log k)$, achieving the optimal rate up to a logarithmic factor [41]. Latin hypercube designs (LHDs) are not necessarily quasi-uniform. However, variants of LHDs, such as maximin and minimax LHDs, exist that optimize the fill or separation distance [51]. ∎

We will ground the upcoming analysis on the smoothness level of the kernel function K, derive the convergence rate of the predictive variance under the supremum norm, and investigate the asymptotic order of the approximation error as quantified by the uniform error bound in the fixed design setting. Subsections 4.1 and 4.2 elaborate on the finite smoothness and infinite smoothness cases, respectively.

4.1 Finite Smoothness Case

This subsection performs an analysis in the fixed design setting when the kernel K has a finite smoothness level. We start with the following result which provides an upper bound on the error of the regularized kernel based approximation s_λ given in (14).

Proposition 1 (PROPOSITION 3.6 IN WENDLAND AND RIEGER [49]). *Let $\nu = \kappa + s$ with $\kappa > d/2$ and $s \in (0, 1]$. Suppose that $\mathcal{X} \subseteq \mathcal{R}^d$ is bounded and satisfies an interior cone condition—there exists an angle $\theta \in (0, \pi/2)$ and a radius $\tilde{r} > 0$ such that for every $\mathbf{x} \in \mathcal{X}$ there exists a unit vector $\zeta(\mathbf{x})$ such that the cone $C(\mathbf{x}, \zeta(\mathbf{x}), \theta, \tilde{r}) := \{\mathbf{x} + \gamma \mathbf{x}_0 : \mathbf{x}_0 \in \mathbb{R}^d, \|\mathbf{x}_0\| = 1, \mathbf{x}_0^\top \zeta(\mathbf{x}) \ge \cos \theta, \gamma \in [0, \tilde{r}]\}$ is contained in $\mathcal{X}$. Let $q \in [1, \infty]$. Suppose that s_λ is the solution given in (14) for the given design-point set $\mathcal{D} \subseteq \mathcal{X}$ and the target function $f \in W_2^\nu(\mathcal{X})$. Then the following error estimate holds:*

$$|f - s_\lambda|_{W_q^j(\mathcal{X})} \le C \left(h_{\mathcal{D},\mathcal{X}}^{\nu-j-d(\frac{1}{2}-\frac{1}{q})} + h_{\mathcal{D},\mathcal{X}}^{-j}\sqrt{\lambda} \right) \|f\|_{W_2^\nu(\mathcal{X})}, \quad 0 \le j < \kappa - d/2,$$

which gives that, for $\lambda \le h_{\mathcal{D},\mathcal{X}}^{2\nu-d(1-2/q)_+}$,

$$|f - s_\lambda|_{W_q^j(\mathcal{X})} \le C h_{\mathcal{D},\mathcal{X}}^{\nu-j-d(\frac{1}{2}-\frac{1}{q})_+} \|f\|_{W_2^\nu(\mathcal{X})}, \quad 0 \le j < \kappa - d/2,$$

where $(x)_+ = \max\{x, 0\}$.

We note that the interior cone condition essentially states that the input space $\mathcal{X}$ is sufficiently regular with no sharp corners or cusps; and the d-dimensional unit cube $\mathcal{X} = [0, 1]^d$ provides an example [41]. By setting $f := K_{\mathbf{x}_0}$, $s_\lambda := \widehat{K}_{\mathbf{x}_0}$, $j = 0$, and $q = \infty$ in Proposition 1, we obtain the following result which reveals the optimal convergence rate of the maximum discrepancy between $K_{\mathbf{x}_0}$ and its approximation $\widehat{K}_{\mathbf{x}_0}$.

Corollary 1. *Suppose that the conditions of Proposition 1 are fulfilled. If $\lambda \le h_{\mathcal{D},\mathcal{X}}^{2\nu-d}$ and the design-point set $\mathcal{D}$ is quasi-uniform so that $h_{\mathcal{D},\mathcal{X}} \asymp k^{-1/d}$, then the maximum discrepancy between the function $f := K_{\mathbf{x}_0}$ and its approximation $s_\lambda := \widehat{K}_{\mathbf{x}_0}$ achieves the optimal convergence rate:*

$$\|K_{\mathbf{x}_0} - \widehat{K}_{\mathbf{x}_0}\|_\infty \lesssim k^{-\frac{2\nu-d}{2d}}.$$

In the finite smoothness case under the fixed design setting, Corollary 1 indicates that $\sigma_{k,B}(\mathbf{x}_0) \lesssim k^{-\frac{\nu-d/2}{2d}}$ for any $\mathbf{x}_0 \in \mathcal{X}$. The condition on the smoothness parameter λ stipulated in Corollary 1 can be expressed as $V_{k,\max} \lesssim k^{1-\frac{2\nu}{d}}$ following the definition $\lambda = V_{k,\max}/\xi^2$. This condition indicates that: (i) the upper bound of $V_{k,\max}$ must decrease with the number of design points k; and (ii) given a fixed number of design points k, the higher the smoothness level ν (respectively, the higher the input-space dimensionality d) a given problem has, the faster (resp., the slower) the decreasing rate required on $V_{k,\max}$. The above sheds light on the following important aspects of a desirable simulation experimental design to achieve the optimal convergence rate of $\|\sigma_{k,B}^2\|_\infty$. First

and foremost, the design-point set must be quasi-uniform. Moreover, the numbers of simulation replications n_i's must grow with the number of design points k, and the smoother the kernel K, the faster n_i's must grow.

Given that the kernel has a finite smoothness level, the next result extends Theorem 2 by elucidating the order of the uniform error bound as $k, B \to \infty$ in the fixed design setting.

Proposition 2. *Suppose that the assumptions of Theorem 2 and Corollary 1 are fulfilled, the kernel K has $v > d/2$ smoothness level with $v < \infty$, and $V_{k,\max} \asymp k^{1-2v/d}$. Given any $\alpha \in (0, 1)$, it holds that*

$$\mathrm{P}\left(\sup_{\mathbf{x} \in \mathcal{X}} |\mu_{k,B}(\mathbf{x}) - f(\mathbf{x})| = O\left((\log k)^{\frac{1}{2}} k^{-\frac{v-d/2}{2d}}\right)\right) \geq 1 - \alpha.$$

Proof. We see from (8) that both $\gamma_{k,B}(\tau)$ and $\sqrt{\beta_k(\tau)} \sup_{\mathbf{x} \in \mathcal{X}} \sigma_{k,B}(\mathbf{x})$ must converge to zero as $k, B \to \infty$ to guarantee a vanishing approximation error at $\forall \mathbf{x} \in \mathcal{X}$. In the finite smoothness case under the fixed design setting, $\lim_{k,B \to \infty} \gamma_{k,B}(\tau) = 0$ can be achieved if the grid constant as a function of k, $\tau(k)$, decreases sufficiently fast. Recall from (12) that $\gamma_{k,B}(\tau) \lesssim \left(\left(\beta_k(\tau) \cdot \tau(k) \cdot kV_{k,\min}^{-1}\right)^{\frac{1}{2}} + \tau(k) + \tau(k) \cdot kV_{k,\min}^{-\frac{1}{2}}\right)$. Since $V_{k,\max}/V_{k,\min} < \infty$, $V_{k,\min} \asymp k^{1-2v/d}$, setting $\tau(k) = O(k^{-a})$ as $k, B \to \infty$ with $a = 3v/d$ ensures that $\lim_{k,B \to \infty} \gamma_{k,B}(\tau) = 0$ and moreover $\gamma_{k,B}(\tau) = o\left(\sqrt{\beta_k(\tau)} \sup_{\mathbf{x} \in \mathcal{X}} \sigma_{k,B}(\mathbf{x})\right)$ as $k, B \to \infty$. The latter follows from the fact that $\beta_k(\tau) = O(\log k)$ due to (9) and $\sigma_{k,B}(\mathbf{x}) = O\left(k^{-\frac{v-d/2}{2d}}\right)$ at $\forall \mathbf{x} \in \mathcal{X}$ thanks to Corollary 1. Therefore, $\gamma_{k,B}(\tau) + \sqrt{\beta_k(\tau)} \sup_{\mathbf{x} \in \mathcal{X}} \sigma_{k,B}(\mathbf{x}) \lesssim \sqrt{\beta_k(\tau)} \sup_{\mathbf{x} \in \mathcal{X}} \sigma_{k,B}(\mathbf{x}) = O\left((\log k)^{\frac{1}{2}} k^{-\frac{v-d/2}{2d}}\right)$ as $k, B \to \infty$. □

4.2 Infinite Smoothness Case

When the kernel K has infinite smoothness, the corresponding RKHS $\mathcal{H}$ can be continuously embedded into every classical Sobolev space $W_p^v(\mathcal{X})$ for all $v \in \mathbb{N}$ and some fixed $p \in [1, \infty)$ [36]. By restricting our attention to two popular infinitely smooth kernels, respectively the Gaussian kernel (abbreviated as G): $K(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/\theta)$ with $\theta > 0$ and the inverse multiquadrics kernel (abbreviated as M): $K(\mathbf{x}, \mathbf{x}') = (c^2 + \|\mathbf{x} - \mathbf{x}'\|^2)^{-\xi}$, where $c > 0$ and $\xi > d/2$ (page 76 of [48]), we reveal the convergence rate of $\sigma_{k,B}^2(\cdot)$ under the supremum norm and the asymptotic order of the uniform error bound as $k, B \to \infty$ in the fixed design setting. We first state a result that provides an upper bound on the error of the regularized kernel based approximation s_λ given in (14).

Proposition 3 (Theorem 6.1 and Corollary 6.3 in Rieger and Zwicknagl [36]). *If $\mathcal{X}$ is a compact cube with side length ℓ, then for all $q \in [1, \infty]$, there exist constants h_0, C, and $\tilde{C} > 0$ such that for all design-point sets $\mathcal{D} \subset \mathcal{X}$ with fill distance $h_{\mathcal{D},\mathcal{X}} \leq h_0$ and for all $f \in \mathcal{H}_G$ in the RKHS of a Gaussian kernel G,*

$$\|f - s_\lambda\|_{L_q(\mathcal{X})} \leq \left(2 \exp\left(C \log(h_{\mathcal{D},\mathcal{X}})/h_{\mathcal{D},\mathcal{X}}\right) + \sqrt{\lambda} \exp\left(\tilde{C}/h_{\mathcal{D},\mathcal{X}}\right)\right) \|f\|_{\mathcal{H}_G},$$

which gives that, for $\lambda \leq \exp\left(2\left(C \log(h_{\mathcal{D},\mathcal{X}}) - \tilde{C}\right)/h_{\mathcal{D},\mathcal{X}}\right)$,

$$\|f - s_\lambda\|_{L_q(\mathcal{X})} \leq 3 \exp\left(C \log(h_{\mathcal{D},\mathcal{X}})/h_{\mathcal{D},\mathcal{X}}\right) \|f\|_{\mathcal{H}_G}.$$

For all design-point sets $\mathcal{D} \subset \mathcal{X}$ with fill distance $h_{\mathcal{D},\mathcal{X}} \leq h_0$ and for all $f \in \mathcal{H}_M$ in the RKHS of an inverse multiquadrics kernel M,

$$\|f - s_\lambda\|_{L_q(\mathcal{X})} \leq \left(2 \exp\left(-C/h_{\mathcal{D},\mathcal{X}}\right) + \sqrt{\lambda} \exp\left(\tilde{C}/h_{\mathcal{D},\mathcal{X}}\right)\right) \|f\|_{\mathcal{H}_M},$$

which gives that, for $\lambda \le \exp\left(-2\left(C + \tilde{C}\right)/h_{\mathcal{D},\mathcal{X}}\right)$,

$$\|f - s_\lambda\|_{L_q(\mathcal{X})} \le 3 \exp\left(-C/h_{\mathcal{D},\mathcal{X}}\right) \|f\|_{\mathcal{H}_M}.$$

Notice that the constants C, $\tilde{C}$, and h_0 depend only on d, ℓ, and q.

Setting $f := K_{\mathbf{x}_0}$, $s_\lambda := \widehat{K}_{\mathbf{x}_0}$, and $q = \infty$ in Proposition 3 yields the following result on the optimal convergence rate of the maximum discrepancy between $K_{\mathbf{x}_0}$ and its approximation $\widehat{K}_{\mathbf{x}_0}$ in the infinite smoothness case.

Corollary 2. *Suppose that the assumptions of Proposition 3 are fulfilled. For a Gaussian kernel, if $\lambda \le \exp\left(2\left(C\log(h_{\mathcal{D},\mathcal{X}}) - \tilde{C}\right)/h_{\mathcal{D},\mathcal{X}}\right)$ and the design-point set $\mathcal{D}$ is quasi-uniform so that $h_{\mathcal{D},\mathcal{X}} \asymp k^{-1/d}$, then the maximum discrepancy between the function $f := K_{\mathbf{x}_0}$ and its approximation $s_\lambda := \widehat{K}_{\mathbf{x}_0}$ attains its optimal convergence rate:*

$$\|K_{\mathbf{x}_0} - \widehat{K}_{\mathbf{x}_0}\|_\infty \lesssim \exp\left(-Ck^{\frac{1}{d}}\left(\log k\right)\right).$$

For an inverse multiquardrics kernel, if $\lambda \le \exp\left(-2\left(C + \tilde{C}\right)/h_{\mathcal{D},\mathcal{X}}\right)$ and the design-point set $\mathcal{D}$ is quasi-uniform so that $h_{\mathcal{D},\mathcal{X}} \asymp k^{-1/d}$, then the maximum discrepancy between the function $f := K_{\mathbf{x}_0}$ and its approximation $s_\lambda := \widehat{K}_{\mathbf{x}_0}$ attains its optimal convergence rate:

$$\|K_{\mathbf{x}_0} - \widehat{K}_{\mathbf{x}_0}\|_\infty \lesssim \exp\left(-Ck^{\frac{1}{d}}\right).$$

In the infinite smoothness case under the fixed design setting, Corollary 2 indicates that $\sigma_{k,B}(\mathbf{x}_0) \lesssim \exp(-Ck^{\frac{1}{d}}/2)$ at any $\mathbf{x}_0 \in \mathcal{X}$. The condition on the regularization parameter $\lambda = \mathsf{V}_{k,\max}/\xi^2$ stipulated in Corollary 2 can be expressed as $\mathsf{V}_{k,\max} \lesssim \exp\left(-\left(2Cd^{-1}\left(\log k\right) + 2\tilde{C}\right)k^{1/d}\right)$ and $\mathsf{V}_{k,\max} \lesssim \exp\left(-2(C + \tilde{C})k^{1/d}\right)$ for a Gaussian kernel and an inverse multiquadrics kernel, or equivalently, $\mathsf{V}_{k,\max} \lesssim \exp\left(-\bar{C}k^{1/d}\right)$ with $\bar{C} = 2(C + \tilde{C})$ if the logarithmic factor term is ignored. It implies that: (i) the upper bound of $\mathsf{V}_{k,\max}$ must decrease exponentially with $k^{1/d}$; and (ii) given a fixed number of design points k, the higher the input-space dimensionality d of a given problem, the slower the decreasing rate required on $\mathsf{V}_{k,\max}$. The above also provides some insight into a desirable simulation experimental design to achieve the optimal convergence rate of $\|\sigma_{k,B}^2\|_\infty$. First, the design-point set must be quasi-uniform. Moreover, the numbers of simulation replications n_i's must grow with the number of design points k at a much faster rate in the infinitely smoothness case compared to in the finite smoothness case.

Given that the kernel has infinite smoothness, the next result extends Theorem 2 by elucidating the order of the uniform error bound as $k, B \to \infty$ in the fixed design setting.

Proposition 4. *Suppose that the assumptions of Theorem 2 and Corollary 2 are fulfilled, the kernel K has infinite smoothness, and $\mathsf{V}_{k,\max} \asymp \exp\left(-\bar{C}k^{1/d}\right)$ with $\bar{C} = 2(C + \tilde{C})$, where C and $\tilde{C}$ are defined in Proposition 3. Given any $\alpha \in (0, 1)$, it holds that*

$$P\left(\sup_{\mathbf{x} \in \mathcal{X}} |\mu_{k,B}(\mathbf{x}) - f(\mathbf{x})| = O\left(k^{\frac{1}{2d}} \exp(-Ck^{\frac{1}{d}}/2)\right)\right) \ge 1 - \alpha.$$

PROOF. We see from (8) that both $\gamma_{k,B}(\tau)$ and $\sqrt{\beta_k(\tau)} \sup_{\mathbf{x} \in \mathcal{X}} \sigma_{k,B}(\mathbf{x})$ must converge to zero as $k, B \to \infty$ to guarantee a vanishing approximation error at $\forall \mathbf{x} \in \mathcal{X}$. In the infinite smoothness case under the fixed design setting, $\lim_{k,B\to\infty} \gamma_{k,B}(\tau) = 0$ can be achieved if the grid constant as a function of k, $\tau(k)$, decreases sufficiently fast. Recall from (12) that $\gamma_{k,B}(\tau) \lesssim \left(\left(\beta_k(\tau) \cdot \tau(k) \cdot k\mathsf{V}_{k,\min}^{-1}\right)^{\frac{1}{2}} + \tau(k) + \tau(k) \cdot k\mathsf{V}_{k,\min}^{-\frac{1}{2}}\right)$. Since $\mathsf{V}_{k,\max}/\mathsf{V}_{k,\min} < \infty$ and $\mathsf{V}_{k,\min} \asymp \exp\left(-\bar{C}k^{1/d}\right)$ with $\bar{C} = 2(C + \tilde{C})$, setting $\tau(k) = O\left(\exp\left(-ak^{\frac{1}{d}}\right)\right)$ where $a > \bar{C} + C$ ensures

$\lim\limits_{k,B\to\infty} \gamma_{k,B}(\tau) = 0$ and moreover $\gamma_{k,B}(\tau) = o\left(\sqrt{\beta(\tau)}\sigma_{k,B}(\mathbf{x})\right)$ at $\forall \mathbf{x} \in \mathcal{X}$ as $k, B \to \infty$. The latter follows from the fact that $\beta_k(\tau) = O(k^{\frac{1}{d}})$ due to (9) and $\sigma_{k,B}(\mathbf{x}) = O\left(\exp(-Ck^{\frac{1}{d}}/2)\right)$ at $\forall \mathbf{x} \in \mathcal{X}$ thanks to Corollary 2. Therefore,

$$\gamma_{k,B}(\tau) + \sqrt{\beta_k(\tau)} \sup_{\mathbf{x}\in\mathcal{X}} \sigma_{k,B}(\mathbf{x}) \lesssim \sqrt{\beta_k(\tau)} \sup_{\mathbf{x}\in\mathcal{X}} \sigma_{k,B}(\mathbf{x}) = O\left(k^{\frac{1}{2d}}\exp(-Ck^{\frac{1}{d}}/2)\right) \text{ as } k, B \to \infty. \qquad \square$$

We close this section with some practical insights into suitable experimental designs for SK metamodeling. To achieve a desirable convergence rate of $\sigma_{k,B}(\cdot)$ under the supremum norm given a sufficiently large budget B to expend, an experimental design suitable for the infinite smoothness case is expected to use fewer design points but with more replications allocated to each design point to fulfill its condition on $\mathsf{V}_{k,\max}$, hence a "sparse and deep" design is recommended. On the other hand, an experimental design suitable for the finite smoothness case can adopt relatively more design points but with fewer replications allocated at each point to fulfill its condition on $\mathsf{V}_{k,\max}$; hence, a "dense and shallow" design is recommended. Such insights echo those reported in other contexts in the stochastic simulation metamodeling literature; see, e.g., [46] and [55].

5 Discussion and Extension

This section provides implications for simulation experimental designs for SK and theoretically investigates the impact of noise variance estimation on the performance of the uniform error bound.

5.1 Discussion

This subsection provides some further insight into the impact of experimental designs and budget allocation schemes on SK metamodeling in light of the results obtained in Section 4 for the fixed design setting.

Given a fixed number of design points k, it is not difficult to see that $\mathsf{V}_{k,\max} = \max\limits_{1\le i\le k} \mathsf{V}(\mathbf{x}_i)/n_i$ is minimized by setting the number of replications at design point $\mathbf{x}_i$ to

$$n_i = \frac{\mathsf{V}(\mathbf{x}_i)}{\sum_{j=1}^{k} \mathsf{V}(\mathbf{x}_j)} B, \ i = 1, 2, \ldots, k. \tag{16}$$

Under the budget allocation in (16), we have $\lambda = \mathsf{V}_{k,\max}/\xi^2 = (B\xi^2)^{-1}\sum_{i=1}^{k}\mathsf{V}(\mathbf{x}_i)$. When k becomes large, it follows that

$$\lambda = (B\xi^2)^{-1}k\left(\frac{1}{k}\sum_{i=1}^{k}\mathsf{V}(\mathbf{x}_i)\right) \approx (B\xi^2)^{-1}k\left(|\mathcal{X}|^{-1}\int_{\mathcal{X}}\mathsf{V}(\mathbf{u})d\mathbf{u}\right) \asymp k/B,$$

where the last step follows from $|\mathcal{X}|^{-1}\int_{\mathcal{X}}\mathsf{V}(\mathbf{u})d\mathbf{u} < \infty$.

On the other hand, if an equal budget allocation rule is applied, i.e., $n_i = B/k$ for $i = 1, 2, \ldots, k$, it is easy to see that in this case $\lambda = (B\xi^2)^{-1}k\max\limits_{1\le i\le k}\mathsf{V}(\mathbf{x}_i)$. When k becomes large, we have

$$\lambda = (B\xi^2)^{-1}k\left(\max_{1\le i\le k}\mathsf{V}(\mathbf{x}_i)\right) \approx (B\xi^2)^{-1}k\sup_{\mathcal{X}}\mathsf{V}(\mathbf{x}) \asymp k/B,$$

where in the last step we have used $\sup_{\mathcal{X}}\mathsf{V}(\mathbf{x}) < \infty$. In brief, when k is sufficiently large, different budget allocation rules lead to the same order of λ which only depends on the total budget B and the number of design points k.

The discussion above sheds some light on the relationship between the number of design points k and the total budget B to achieve the optimal decreasing rate of $\|\sigma_{k,B}\|_\infty$. In light of the results for the fixed design setting in Section 4, we have the following remarks regarding the finite and the infinite smoothness cases. When the underlying kernel function has a smoothness level $\nu < \infty$, the discussion above and Corollary 1 indicate that $\lambda \asymp k/B \lesssim k^{1-\frac{2\nu}{d}}$, or equivalently, $k \lesssim B^{\frac{d}{2\nu}}$, when the budget B is sufficiently large. Therefore, in addition to

adopting a quasi-uniform design-point set, to achieve the optimal convergence rate of $\sigma_{k,B}(\cdot)$, one should use a large number of design points k if the input-space dimensionality d of a given problem is high (respectively, the smoothness level ν is low), and vice versa. On the other hand, when the kernel function has an infinite smoothness level, the discussion above and Corollary 2 indicate that $\lambda \asymp k/B \lesssim \exp\left(-\left(2Cd^{-1}\left(\log k\right) + 2\tilde{C}\right)k^{1/d}\right)$ (for a Gaussian kernel), or equivalently, $k \lesssim (\log B)^d$ if the logarithmic factor terms are ignored, when the budget B is sufficiently large. Hence, besides adopting a quasi-uniform design-point set, to achieve an optimal convergence rate of $\|\sigma_{k,B}\|_\infty$ in this case, one should use a relatively large number of design points k if the input space dimensionality d of a given problem is high. However, the number of design points required is expected to be much smaller than that in the finite smoothness case.

Last but not least, we note that the root mean squared error of SK is upper bounded by the predictive standard deviation under the supremum norm:

$$\text{RMSE} := \left(\int_\mathcal{X} \sigma_{k,B}^2(\mathbf{x})\mathrm{d}\mathbf{x}\right)^{1/2} \leq \|\sigma_{k,B}\|_\infty.$$

Therefore, the aforementioned desirable simulation experimental designs for achieving the optimal convergence rate of $\|\sigma_{k,B}\|_\infty$ also apply if the target is to achieve an adequate global response surface fitting with a guaranteed RMSE convergence rate.

5.2 Impact of Noise Variance Estimation on the Uniform Error Bound

In this subsection, we conduct a theoretical investigation of the impact of noise variance estimation on the performance of the uniform error bound. Recall that, in light of Theorem 1, we can obtain a high-probability uniform bound as given in (7) for $f(\cdot)$, assuming the knowledge of true noise variances $\mathsf{V}(\mathbf{x}_i)$ for $i = 1, 2, \ldots, k$. Refer to this bound as the *nominal* uniform bound. Denote $\widehat{\mathsf{V}}(\mathbf{x}_i)$ as the sample variance calculated based on $n_i \geq 2$ replications simulated at $\mathbf{x}_i$ for $i = 1, 2, \ldots, k$. Define $\widehat{LB}(\mathbf{x}) := \widehat{\mu}_{k,B}(\mathbf{x}) - \sqrt{\beta(\tau)}\widehat{\sigma}_{k,B}(\mathbf{x}) - \widehat{\gamma}_B(\tau)$ and $\widehat{UB}(\mathbf{x}) := \widehat{\mu}_{k,B}(\mathbf{x}) + \sqrt{\beta(\tau)}\widehat{\sigma}_{k,B}(\mathbf{x}) + \widehat{\gamma}_B(\tau)$, which are constructed using the sample variance $\widehat{\mathsf{V}}(\mathbf{x}_i)$ in place of $\mathsf{V}(\mathbf{x}_i)$ for $i = 1, 2, \ldots, k$. We refer to $[\widehat{LB}(\mathbf{x}), \widehat{UB}(\mathbf{x})]$ as the *empirical* uniform bound for $f(\mathbf{x})$ at all $\mathbf{x} \in \mathcal{X}$ and study its performance as k and B tend to infinity.

Define $n_k^* := \min_{1 \leq i \leq k} n_i$ as the minimum number of replications allocated across the k design points, and recall $\mathsf{V}_{k,\max} = \max_{1 \leq i \leq k} \mathsf{V}(\mathbf{x}_i)/n_i$. The following result provides a high-probability bound of the maximum discrepancy between the diagonal entries in the true noise variance-covariance matrix Σ_ε and those in its estimate $\widehat{\Sigma}_\varepsilon := \mathrm{diag}\left(\widehat{\mathsf{V}}(\mathbf{x}_1)/n_1, \widehat{\mathsf{V}}(\mathbf{x}_2)/n_2, \ldots, \widehat{\mathsf{V}}(\mathbf{x}_k)/n_k\right)$, which holds for all $k \geq 1$.

Proposition 5. *Fix $\alpha \in (0, 1)$. Let $\eta_{k,B} = \log\left((\pi_k\alpha)^{-1}\right)$ so that $\sum_{k=1}^{\infty} \pi_k = 1/3$. Suppose that the noise terms incurred at $\mathbf{x}_i$, $\varepsilon_j(\mathbf{x}_i)$'s, are i.i.d. normal with mean zero and variance $\mathsf{V}(\mathbf{x}_i)$ at $i = 1, 2, \ldots, k$, and $n_k^* \geq 2$. Then, with probability at least $1 - \alpha/3$, it holds that,*

$$\max_{1 \leq i \leq k}\left|\frac{\widehat{\mathsf{V}}(\mathbf{x}_i) - \mathsf{V}(\mathbf{x}_i)}{n_i}\right| \leq \sqrt{\frac{8\left(\eta_{k,B} + \log(2k)\right)}{(n_k^* - 1)}} \cdot \mathsf{V}_{k,\max}, \quad \forall k \geq 1.$$

The proof of Proposition 5 is provided in Appendix A.1. Based on Proposition 5, we can analyze the difference between the empirical uniform bound and the nominal uniform bound as k and B tend to infinity.

Theorem 3. *Fix $\alpha \in (0, 1)$. Suppose that the assumptions of Theorem 2 and Proposition 5 are satisfied and the design-point set $\mathcal{D}$ is quasi-uniform.*

If the kernel K has $v > d/2$ smoothness with $v < \infty$, $V_{k,\max} \asymp k^{1-2v/d}$, and $k^{\frac{6v}{d}-1}(\log k) = o(n_k^*)$, it holds with probability at least $1 - \alpha$ that, $\sup_{\mathbf{x} \in \mathcal{X}} \left| \frac{\widehat{LB}(\mathbf{x}) - LB(\mathbf{x})}{\sqrt{\beta(\tau)}\sigma_{k,B}(\mathbf{x}) + \gamma_{k,B}(\tau)} \right| = o(1)$ and $\sup_{\mathbf{x} \in \mathcal{X}} \left| \frac{\widehat{UB}(\mathbf{x}) - UB(\mathbf{x})}{\sqrt{\beta(\tau)}\sigma_{k,B}(\mathbf{x}) + \gamma_{k,B}(\tau)} \right| = o(1)$ as $k, B \to \infty$.

If the kernel K has infinite smoothness, $V_{k,\max} \asymp \exp\left(-\bar{C}k^{1/d}\right)$ for some $\bar{C} > 0$, and $k^2(\log k)\exp(2(\bar{C}+C)k^{\frac{1}{d}}) = o(n_k^*)$ with $C, \bar{C} > 0$ as specified in the proof of Proposition 4, then it holds with probability at least $1 - \alpha$ that, $\sup_{\mathbf{x} \in \mathcal{X}} \left| \frac{\widehat{LB}(\mathbf{x}) - LB(\mathbf{x})}{\sqrt{\beta(\tau)}\sigma_{k,B}(\mathbf{x}) + \gamma_{k,B}(\tau)} \right| = o(1)$ and $\sup_{\mathbf{x} \in \mathcal{X}} \left| \frac{\widehat{UB}(\mathbf{x}) - UB(\mathbf{x})}{\sqrt{\beta(\tau)}\sigma_{k,B}(\mathbf{x}) + \gamma_{k,B}(\tau)} \right| = o(1)$ as $k, B \to \infty$.

The proof of Theorem 3 is deferred to Appendix A.2. We see from Theorem 3 that the difference between the empirical uniform bound and the nominal uniform bound becomes negligible so long as the numbers of replications at all design points grow sufficiently fast with the expansion of the design-point set $\mathcal{D}$. While the conditions stipulated on the minimum number of replications n_k^* may seem stringent, the numerical results in Section 6 indicate that the empirical and nominal uniform bounds perform similarly in practice, unless the numbers of replications at some design points become very small. We note that the conservativeness in the conditions arises from the techniques used to upper bound various terms involving the inverse of $\Sigma := (\Sigma_M + \Sigma_\varepsilon)$ in $|\widehat{LB}(\mathbf{x}) - LB(\mathbf{x})|$ and $|\widehat{UB}(\mathbf{x}) - UB(\mathbf{x})|$ and to ensure that these terms vanish faster than the nominal uniform bound's half-width for all $\mathbf{x} \in \mathcal{X}$. Other proof techniques may provide less stringent conditions. The proof of Theorem 3 also reveals the importance of stabilizing the SK metamodeling process as the design-point set expands and the budget allocated grows. To be more precise, the condition of matrix Σ can become increasingly worse as $k, B \to \infty$, which is reflected by the magnitude of $\|\Sigma^{-1}\| = 1/\lambda_{\min}(\Sigma)$ in the proof of Theorem 3. This also leads us to briefly discuss the condition number of Σ, $\text{cond}(\Sigma) = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$. It is known that the maximum eigenvalue typically incurs no problem and behaves nicely compared to the minimum eigenvalue [Chapter 12 of 48]. Since $\lambda_{\min}(\Sigma) \geq \lambda_{\min}(\Sigma_M) + \lambda_{\min}(\Sigma_\varepsilon) \geq \max\{\lambda_{\min}(\Sigma_\varepsilon), \lambda_{\min}(\Sigma_M)\}$, we know more about the stability and accuracy of SK's prediction if we know more about the lower bounds for $\lambda_{\min}(\Sigma_M)$ and $\lambda_{\min}(\Sigma_\varepsilon)$. Lower bounds for $\lambda_{\min}(\Sigma_M)$ are known to depend on the separation distance $q_\mathcal{D}$. Recall from Definition 2 that $q_\mathcal{D} \asymp h_{\mathcal{D},\mathcal{X}} = O(k^{-1/d})$ if the design-point set $\mathcal{D}$ is quasi-uniform. In this case, it is known that $\lambda_{\min}(\Sigma_M) \gtrsim q_\mathcal{D}^{(2v-d)} = O(k^{1-2v/d})$ in the finite smoothness case [51], whereas in the infinite smoothness case, $\lambda_{\min}(\Sigma_M) \gtrsim \exp(-C_0 q_\mathcal{D}^{-2})q_\mathcal{D}^{-d} = O(\exp(-C_0 \cdot k^{2/d})k)$ for some $C_0 > 0$ [Section 12.2 of 48]. To maintain SK's numerical stability as $k, B \to \infty$, $\lambda_{\min}(\Sigma_\varepsilon) = V_{k,\min}$ should not diminish too rapidly. Since $\lambda_{\min}(\Sigma) \geq \max\{\lambda_{\min}(\Sigma_\varepsilon), \lambda_{\min}(\Sigma_M)\} = \lambda_{\min}(\Sigma_\varepsilon) = V_{k,\min}$ and $\|\Sigma^{-1}\| \leq 1/\lambda_{\min}(\Sigma_\varepsilon) = V_{k,\min}^{-1}$, the condition $V_{k,\max}/V_{k,\min} < \infty$ and that on the order of $V_{k,\max}$ stipulated in Theorem 3 ensure the numerical stability of SK while also fulfilling the regularization parameter role played by the noise variances (manifested by $V_{k,\max}$) to achieve the best convergence rate of SK's predictive variance. We will examine the impact of noise variance estimation on the performance of the uniform bound through numerical experiments in the next section.

6 Numerical Experiments

This section compares the performance of the nominal uniform bound assuming the knowledge of the true noise variances (given by (7) in Section 3 and referred to as $\mathbf{CI_u}$), the empirical uniform bound built with noise variances estimated (given at the beginning of Subsection 5.2 and referred to as $\mathbf{CI_{eu}}$), and the Bonferroni correction-based simultaneous confidence interval built with noise variances estimated (given at the end of Subsection 2.1 and referred to as $\mathbf{CI_b}$) through numerical evaluations. We consider the following examples: an M/M/1 queueing model and a dimension-flexible synthetic example.

We start by describing the general experimental setup for numerical evaluations. A simulation experiment is performed with a total budget of B observations to collect at k distinct design points, with n_i observations to

obtain at design point $\mathbf{x}_i$, for $i = 1, 2, \ldots, k$. The k design points are chosen according to either a random design or a fixed design, to be specified for each example. Three budget allocation schemes, i.e., the equal allocation, unequal allocation 1, and unequal allocation 2, are considered. Specifically, the equal budget allocation scheme sets $n_i = \lceil B/k \rceil$, where $\lceil a \rceil$ denotes the least integer not less than a. Assuming that the noise variance function is known, unequal allocation 1 sets $n_i = \left\lceil \frac{V(\mathbf{x}_i)}{\sum_{l=1}^{k} V(\mathbf{x}_l)} B \right\rceil$, while unequal allocation 2 sets $n_i = \left\lceil \frac{\sqrt{V(\mathbf{x}_i)}}{\sum_{l=1}^{k} \sqrt{V(\mathbf{x}_l)}} B \right\rceil$. We note that the goal of unequal allocation 1 is to equalize the variance of sample means across all design points. This approach has been explored in the literature on ranking and selection as well as multi-armed bandit problems [9, 16, 19], aiming to alleviate the impact of heteroscedasticity. Unequal allocation 2 assigns the simulation budget in accordance with the scale function $\rho(\mathbf{x}) := \sqrt{V(\mathbf{x})}$ for $\mathbf{x} \in \mathcal{X}$; it is proposed based on the optimal design density $h^*(\cdot)$ established for controlled nonparametric regression experiments. Specifically, $h^*(x) = \rho(x)/\int_{\mathcal{X}} \rho(u) du$ achieves the asymptotic sharp minimax lower bound for the mean integrated squared error when $x \in \mathcal{X} = [0, 1]$; see [17] and references therein for more details. We consider various experimental settings comprising various combinations of the total budget B, the number of design points k, the design-point sampling scheme to choose k design points, and the budget allocation scheme adopted. The target confidence level is set to 0.95. SK is implemented by adopting two types of kernel functions, respectively, the Gaussian kernel and the Matérn kernel in (15) with smoothness parameter $\nu = 5/2 + d/2$, where d denotes the dimension of the input space $\mathcal{X}$.

For performance evaluation, we repeat the simulation experiment under each experimental setting for $M = 100$ independent macro-replications and calculate the empirical simultaneous coverage probability ($\widehat{\text{SCP}}$) of a given confidence bound defined as

$$\widehat{\text{SCP}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\{f(\mathbf{x}_{0,i}) \in \text{CI}(\mathbf{x}_{0,i}) \text{ for each } i = 1, 2, \ldots, N \text{ on the } m\text{th macro-replication}\},$$

where $\text{CI}(\mathbf{x}_{0,i})$ refers to CI_{u}, CI_{eu}, or CI_{b} obtained at prediction point $\mathbf{x}_{0,i}$, $i = 1, 2, \ldots, N$; and $\mathbf{1}\{\mathcal{A}\}$ is 1 if event $\mathcal{A}$ is true and 0 otherwise. Moreover, we compare the half-widths of CI_{u}, CI_{eu}, and CI_{b} constructed for the N prediction points based on their respective maximum half-widths $H_{\max}$ achieved on each macro-replication.

6.1 M/M/1 Queue

Consider simulating an M/M/1 queue with service rate 1 and arrival rate x with $x \in \mathcal{X} = [0.3, 0.9]$. In this example, the simulation output recorded on each replication at a given design point is the average number of customers in the system from time 0 to T time units. It is well known from queueing theory that the steady-state mean number of customers in the queue is $f(x) = x/(1 - x)$, which is the mean function we intend to estimate. The noise variance function is $V(x) \approx 2x(1 + x)/(T(1 - x)^4)$ for T large [50]. When simulating the M/M/1 queue at each design point, we initialize each replication in steady state and set the run length of each replication T to 1000. Besides the general experimental setup given at the beginning of Section 6, the specific setup for this example is as follows. The value of the grid constant τ is set to ensure that $\sqrt{\beta_k(\tau)}\sigma_{k,B}(x)$ dominates $\gamma_{k,B}(\tau)$ in (7). We consider using a total budget $B \in \{2560, 25600\}$ and varying the number of distinct design points $k \in \{16, 32, 128, 512\}$ when $B = 2560$ and $k \in \{32, 128, 512, 2048\}$ when $B = 25600$. Given each choice of k, we choose k design points according to a uniform design or a grid design, where the former comprises a sample of k independent and uniformly distributed points and the latter k equispaced grid points in $\mathcal{X}$. The three budget allocation schemes mentioned at the beginning of Section 6 are applied given each combination of B, k, and the design-point sampling scheme. A grid of $N = 1000$ equispaced points is selected from $\mathcal{X}$ for prediction.

Summary of results. Tables 1 and 2 show the $\widehat{\text{SCP}}$'s achieved by the empirical uniform bound (CI_{eu}), the nominal uniform bound (CI_{u}), and the Bonferroni correction-based simultaneous confidence intervals (CI_{b}) under different experimental settings when SK is implemented with the Gaussian and Matérn kernels. The following common

observations are made across all experimental settings. First, given a fixed budget B, the $\widehat{\text{SCP}}$'s of CI_{eu} and CI_{u} are typically higher than the target level 0.95, while the $\widehat{\text{SCP}}$'s of CI_{b} are not close to 0.95, except when the number of design points k is relatively small under the grid design. Second, the simultaneous coverage performance of CI_{eu} remains satisfactory as the number of design points k increases until it reaches a point where only a few replications are allocated to some design points, regardless of the kernel function and the design-point sampling scheme adopted; see, e.g., the $\widehat{\text{SCP}}$'s of CI_{eu} obtained under the equal allocation and unequal allocation 1 when $k = 512$ and $B = 2560$. On the other hand, the simultaneous coverage performance of CI_{b} typically deteriorates as k increases given a fixed budget B, regardless of the kernel function adopted. This is evidenced by examining the $\widehat{\text{SCP}}$'s of CI_{b} obtained across all experimental settings shown in Tables 1 and 2. Third, the $\widehat{\text{SCP}}$'s of all bounds increase with the budget B given that the number of design points k is fixed. Lastly, applying unequal budget allocation schemes typically help improve the $\widehat{\text{SCP}}$'s of CI_{b} when the number of design points k is not too large relative to the budget B given. Such an impact on the $\widehat{\text{SCP}}$'s of CI_{eu} and CI_{u} is not as obvious, however.

Table 1. $\widehat{\text{SCP}}$'s of CI_{eu}, CI_{u}, and CI_{b} obtained for the M/M/1 example under the grid design.

B	k	Gaussian									Matérn								
		Equal allocation			Unequal allocation 1			Unequal allocation 2			Equal allocation			Unequal allocation 1			Unequal allocation 2		
		CI_{b}	CI_{eu}	CI_{u}	CI_{b}	CI_{eu}	CI_{u}	CI_{b}	CI_{eu}	CI_{u}	CI_{b}	CI_{eu}	CI_{u}	CI_{b}	CI_{eu}	CI_{u}	CI_{b}	CI_{eu}	CI_{u}
	16	0.80	1	1	0.97	1	1	0.98	1	1	0.76	1	1	1	1	1	0.99	1	1
2560	32	0.68	1	1	0.93	1	1	0.90	1	1	0.65	1	1	0.98	1	1	0.93	1	1
	128	0.46	0.98	0.98	0.52	1	1	0.70	1	1	0.46	0.99	0.99	0.54	1	1	0.75	1	1
	512	0.33	0.92	0.98	0.14	0.88	1	0.57	1	1	0.34	0.93	0.98	0.01	0.28	1	0.64	1	1
	32	0.98	1	1	1	1	1	1	1	1	0.97	1	1	1	1	1	1	1	1
25600	128	0.69	1	1	0.96	1	1	0.93	1	1	0.67	1	1	1	1	1	0.91	1	1
	512	0.54	1	1	0.81	1	1	0.78	1	1	0.45	1	1	0.90	1	1	0.81	1	1
	2048	0.58	1	1	0.33	0.99	1	0.72	1	1	0.44	1	1	0.31	0.99	1	0.72	1	1

Table 2. $\widehat{\text{SCP}}$'s of CI_{eu}, CI_{u}, and CI_{b} obtained for the M/M/1 example under the uniform design.

B	k	Gaussian									Matérn								
		Equal allocation			Unequal allocation 1			Unequal allocation 2			Equal allocation			Unequal allocation 1			Unequal allocation 2		
		CI_{b}	CI_{eu}	CI_{u}	CI_{b}	CI_{eu}	CI_{u}	CI_{b}	CI_{eu}	CI_{u}	CI_{b}	CI_{eu}	CI_{u}	CI_{b}	CI_{eu}	CI_{u}	CI_{b}	CI_{eu}	CI_{u}
	16	0.44	0.92	0.91	0.32	0.88	0.88	0.37	0.89	0.89	0.48	0.94	0.94	0.33	0.91	0.91	0.39	0.93	0.93
2560	32	0.46	0.95	0.95	0.50	0.97	0.98	0.56	0.98	0.98	0.44	0.96	0.96	0.59	0.99	0.99	0.55	0.99	0.99
	128	0.42	0.98	0.99	0.60	1	1	0.55	1	1	0.38	1	0.99	0.62	1	1	0.60	1	1
	512	0.30	0.91	0.98	0.18	0.81	1	0.70	1	1	0.23	0.93	1	0.02	0.36	1	0.67	1	1
	32	0.60	0.94	0.94	0.55	0.98	0.98	0.54	0.99	0.99	0.60	0.96	0.96	0.59	0.99	0.99	0.58	0.99	0.99
25600	128	0.51	1	1	0.83	1	1	0.74	1	1	0.48	1	1	0.85	1	1	0.75	1	1
	512	0.61	1	1	0.82	1	1	0.74	1	1	0.56	1	1	0.81	1	1	0.74	1	1
	2048	0.48	1	1	0.32	0.99	1	0.58	1	1	0.47	1	1	0.30	0.98	1	0.58	1	1

We next examine the maximum half-widths $H_{\max}$ of CI_{eu}, CI_{u}, and CI_{b} obtained under different experimental settings. Figure 1 summarizes the magnitudes of $H_{\max}$ (in a logarithmic scale) of CI_{eu}, CI_{u}, and CI_{b} obtained under the grid design with a total budget of $B = 2560$ when SK is implemented with the Gaussian and Matérn kernels. The corresponding results obtained with $B = 25600$ and those obtained with $B \in \{2560, 25600\}$ under the uniform design are shown in Figures 6—8 in Appendix D.1. We have the following observations from Figures 1 and 6. First, the $H_{\max}$'s of CI_{eu} and CI_{u} are larger than those of CI_{b} under each experimental setting considered, regardless of the kernel function adopted. Second, the magnitudes of $H_{\max}$ of CI_{eu}, CI_{u}, and CI_{b} decrease with the number of design points k when a large total budget $B = 25600$ is applied under all budget allocation schemes.

However, this trend may not hold for a small total budget. Third, for CI_{eu}, CI_u, and CI_b, when a given budget B is allocated to a fixed number of design points, the magnitudes of H_{max} produced under the equal allocation scheme are the highest. It is worth noting that, regarding all bounds, the two unequal budget allocation schemes lead to shorter half-widths than the equal budget allocation scheme without compromising the simultaneous coverage performance, as observed in Table 1. The magnitudes of H_{max} produced under the two unequal allocation schemes are similar, with unequal allocation 2 leading to higher H_{max} values than unequal allocation 1. Lastly, regarding the impact of the design-point sampling scheme, the uniform design tends to lead to more outliers of the H_{max} values compared to the grid design; see Figures 7 and 8 in Appendix D.1 for details. Moreover, under the uniform design, the unequal allocation schemes yield shorter half-widths of all bounds compared to the equal allocation scheme, but the difference between the two unequal allocation schemes is slight.

We summarize the impacts of noise variance estimation, the design-point sampling scheme, and the choice of kernel function observed in the M/M/1 queueing example. First, comparisons based on the $\widehat{SCP}$'s and H_{max}'s of CI_{eu} and CI_u indicate that the performance of the empirical uniform bound is very similar to that of the nominal one, except when the number of replications is too few at some design points, since using the variance estimates can incur additional variability in the SK model fitting. This corroborates the theoretical results provided in Subsection 5.2. Second, with other experimental settings held constant, the performance of all bounds under the grid design which is quasi-uniform typically outperforms that under the uniform design. This aligns with the theoretical results provided in Section 4. Lastly, with other experimental settings held constant, the performance of all bounds built with the Matérn kernel is comparable to that built with the Gaussian kernel for the M/M/1 queueing example. However, the magnitudes of H_{max} of all bounds built with the Matérn kernel tend to be higher than those with the Gaussian kernel; this is because a faster convergence rate of the predictive variance is achieved with the Gaussian kernel than with the Matérn kernel, corroborating Corollaries 1 and 2, along with the comments following the results. Moreover, when the number of replications at some design points is too small, using the Matérn kernel can result in fitting issues that render the SK model unstable and adversely affect simultaneous coverage performance; see the $\widehat{SCP}$'s of CI_{eu} and CI_b obtained under unequal allocation 1 when $B = 2560$ and $k = 512$ shown in Table 1 and the corresponding H_{max}'s shown in Figure 1 (d) in this subsection and Figure 7 (d) in Appendix D.1.

6.2 A Dimension-flexible Example

Consider the following dimension-flexible example adapted from the one studied by [23], where $\mathcal{X} = [-1, 1]^d$ and d denotes the input-space dimension. The simulation output at design point $\mathbf{x} = (x_1, x_2, \ldots, x_d)^\top$ on the jth replication is generated as $\mathcal{Y}_j(\mathbf{x}) = f(\mathbf{x}) + \varepsilon_j(\mathbf{x})$, where $\varepsilon_j(\mathbf{x})$'s are i.i.d. normally distributed with mean zero and variance $V(\mathbf{x})$. The noise variance function is $V(\mathbf{x}) = (2 + \cos(\pi + \sum_{i=1}^d x_i/d))^2$ for any $\mathbf{x} \in \mathcal{X}$. The mean function of interest is given by $f(x) = \sin(9x^2)$ if $d = 1$ and $f(\mathbf{x}) = \sin(9x_1^2) + \sin\left(\left(\sum_{i=2}^d 3x_i/(d-1)\right)^2\right)$ if $d > 1$.

We describe the specific setup for this example next and refer the reader to the beginning of Section 6 for the general experimental setup. We consider two problem cases corresponding to $d = 1$ and $d = 5$. For $d = 1$, we vary the number of design points $k \in \{16, 64, 256, 512\}$ when $B = 2560$ and $k \in \{16, 64, 256, 512, 2048\}$ when $B = 10240$. For $d = 5$, we vary $k \in \{16, 64, 256, 512\}$ when $B = 2560$ and $k \in \{64, 256, 512, 1024, 2048\}$ when $B = 10240$. Given each choice of k, a set of k design points are chosen according to a uniform design or a maximin Latin hypercube design (LHD). The three budget allocation schemes mentioned at the beginning of Section 6 are applied given each combination of B, k, and the design-point sampling scheme. On each macro-replication, we independently generate a set of $N = 2500$ points from $\mathcal{X}$ according to a maximin LHD as the prediction points for performance evaluation.
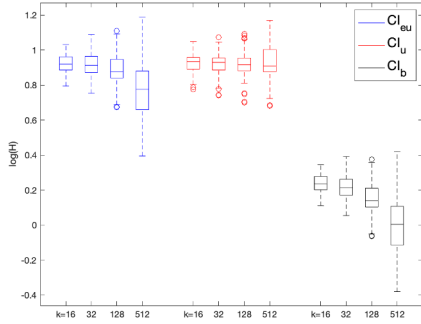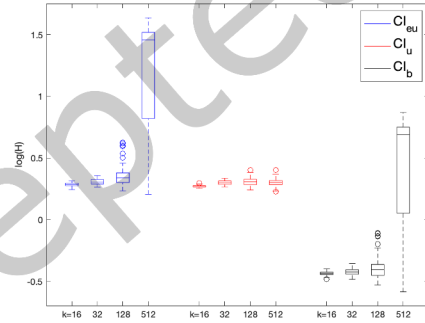
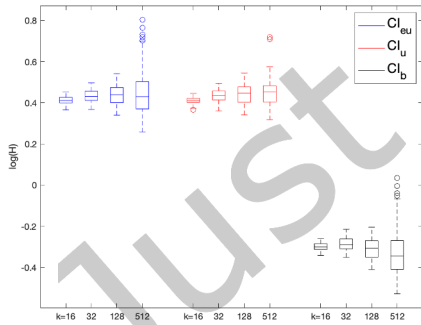(a) Equal allocation, Gaussian
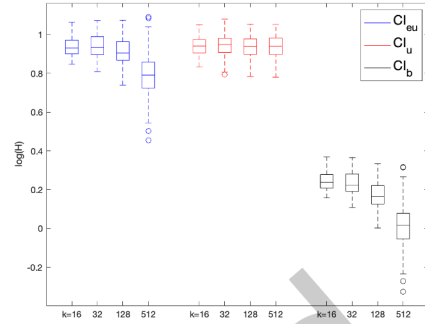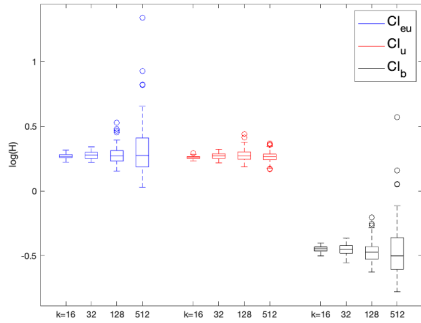
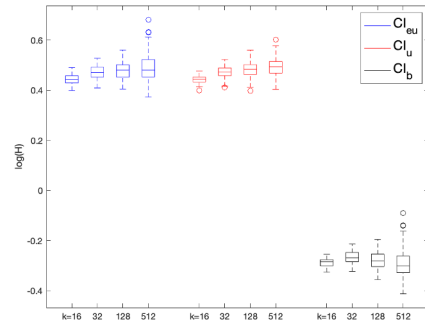(b) Equal allocation, Matérn

(c) Unequal allocation 1, Gaussian

(d) Unequal allocation 1, Matérn

(e) Unequal allocation 2, Gaussian

(f) Unequal allocation 2, Matérn

Fig. 1. Boxplots of $\log(H_{\max})$ of CI_{eu}, CI_u, and CI_b for the M/M/1 example obtained on 100 independent macro-replications under the grid design with $B = 2560$.

Summary of results. We first look into the 1D case. Tables 3 and 4 present the $\widehat{SCP}$'s of the empirical uniform bound (CI_{eu}), the nominal uniform bound (CI_u), and the Bonferroni correction-based simultaneous confidence

intervals (CI_b) under different experimental settings when SK is implemented with the Gaussian and Matérn kernels. Observations akin to those noted in the M/M/1 queueing example are summarized as follows. First, given a fixed budget B, the $\widehat{SCP}$'s of CI_{eu} and CI_u are typically higher than the target level 0.95. Moreover, the performance of CI_{eu} is very similar to that of the nominal one in this 1D case, except when the number of replications is too few at some design points. Second, the $\widehat{SCP}$'s of all bounds increase with the budget B given that the number of design points k is fixed. Third, applying unequal budget allocation schemes typically help improve the $\widehat{SCP}$'s of CI_b when the number of design points k is not too large relative to the budget B given. Lastly, the $\widehat{SCP}$'s of CI_b deteriorate as the number of design points k becomes large, regardless of the budget allocation schemes adopted. We highlight that this 1D case differs from the M/M/1 example in the following aspects. First, the $\widehat{SCP}$'s of CI_b are not significantly lower than those of CI_{eu} and CI_u, so long as the number of design points k is not too large relative to the budget B given. Second, all bounds built with the Matérn kernel achieve comparable or higher $\widehat{SCP}$'s compared to those built with the Gaussian kernel, especially when the number of design points k is small.

Table 3. $\widehat{SCP}$'s of CI_{eu}, CI_u, and CI_b obtained for the 1D case under LHD.

B	k	Gaussian									Matérn								
---	---	Equal allocation			Unequal allocation 1			Unequal allocation 2			Equal allocation			Unequal allocation 1			Unequal allocation 2		
		CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u
	16	0.90	0.92	0.92	0.92	0.95	0.96	0.93	0.94	0.94	1	1	1	1	1	1	1	1	1
2560	64	1	1	1	0.99	1	1	0.99	1	1	1	1	1	1	1	1	1	1	1
	256	0.94	1	1	0.97	1	1	0.94	0.99	1	0.98	1	1	0.98	1	1	0.98	1	1
	512	0.39	0.92	1	0.43	0.94	1	0.58	0.98	1	0.33	0.92	1	0.38	0.92	1	0.53	0.94	1
	16	0.93	0.99	0.99	0.94	0.99	0.99	0.92	0.99	0.99	1	1	1	1	1	1	1	1	1
10240	64	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	256	0.99	1	1	1	1	1	1	1	1	0.99	1	1	1	1	1	1	1	1
	512	1	1	1	1	1	1	0.99	1	1	1	1	1	1	1	1	1	1	1
	2048	0.31	0.99	1	0.34	0.97	1	0.48	0.99	1	0.25	0.93	1	0.21	0.91	1	0.41	0.98	1

Table 4. $\widehat{SCP}$'s of CI_{eu}, CI_u, and CI_b obtained for the 1D case under the uniform design.

B	k	Gaussian									Matérn								
---	---	Equal allocation			Unequal allocation 1			Unequal allocation 2			Equal allocation			Unequal allocation 1			Unequal allocation 2		
		CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u
	16	0.94	0.95	0.94	0.96	0.97	0.93	0.94	0.94	0.94	1	1	1	1	1	1	1	1	1
2560	64	0.91	0.92	0.95	0.99	0.99	0.98	0.96	0.96	0.97	0.99	1	1	1	1	1	1	1	1
	256	0.97	1	0.97	0.97	1	1	0.95	1	1	0.98	1	1	0.99	1	1	0.99	1	1
	512	0.43	0.98	1	0.46	0.92	1	0.58	0.99	1	0.46	0.96	1	0.40	0.88	1	0.65	0.98	1
	16	0.97	0.98	0.97	0.97	0.98	0.98	0.97	0.98	0.98	0.99	1	1	0.99	1	1	0.99	1	1
10240	64	0.99	0.99	0.99	0.99	0.99	0.99	1	1	1	1	1	1	1	1	1	1	1	1
	256	0.98	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	512	0.99	1	1	1	1	1	0.99	1	1	1	1	1	1	1	1	1	1	1
	2048	0.30	0.97	1	0.38	0.98	1	0.56	1	1	0.17	0.93	1	0.29	0.91	1	0.47	0.99	1

We further delve into the maximum half-widths H_{max} of CI_{eu}, CI_u, and CI_b obtained under different experimental settings. To save space, we concentrate on analyzing the results obtained under LHD with the equal allocation scheme. Similar observations apply to the results obtained under LHD with the two unequal budget allocation schemes and under the uniform design with all three budget allocations; see Figures 9 and 10 in Appendix D.2.1 for details. Figures 2 and 3 summarize the magnitudes of H_{max} (in a logarithmic scale) of CI_{eu}, CI_u, and CI_b

obtained under LHD with a total budget of $B = 2560$ and 10240. First, the $H_{\max}$'s of CI_{eu} and CI_u are larger than those of CI_b under each experimental setting considered, regardless of the kernel function adopted. Second, the magnitudes of $H_{\max}$ of CI_{eu}, CI_u, and CI_b decrease with the number of design points k when a large total budget $B = 10240$ is applied. However, this trend is less discernible with a small budget. Third, the magnitudes of $H_{\max}$ of CI_{eu} and CI_u are similar, except when k becomes large relative to the given budget B so that using variance estimates based on a few replications can incur additional variability in SK model fitting, which impacts CI_{eu}; this observation holds regardless of the kernel function adopted. Fourth, applying unequal budget allocation schemes only slightly reduce the half-widths of all three bounds, as observed in Figures 9 (a)—(d) and 10 (a)—(d) in Appendix D.2.1. Lastly, the magnitudes of $H_{\max}$ of all bounds built with the Matérn kernel tend to be higher than those with the Gaussian kernel, which explains the corresponding higher $\widehat{SCP}$'s observed in Table 3. We also highlight some differences in the 1D case from the M/M/1 example. Given a fixed total budget B, especially a small one, using a small number of design points k can lead to unstable SK model fitting when the Gaussian kernel is adopted, resulting in highly variable $H_{\max}$ values for all bounds; see, e.g., the $H_{\max}$'s obtained for $k = 16$ in Figure 2 (a). This is because the true mean function $f(\cdot)$ in this 1D case is nonlinear and its trend changes more frequently than that of the M/M/1 example across the input space, due to its combination of the trigonometric function and higher-order polynomial terms. It is known that the Gaussian kernel is not flexible enough for modeling a function with such features, and using a small number of design points exacerbates this deficiency in capturing the true mean function $f(\cdot)$ in the 1D case.

(a) Equal allocation, Gaussian   (b) Equal allocation, Matérn
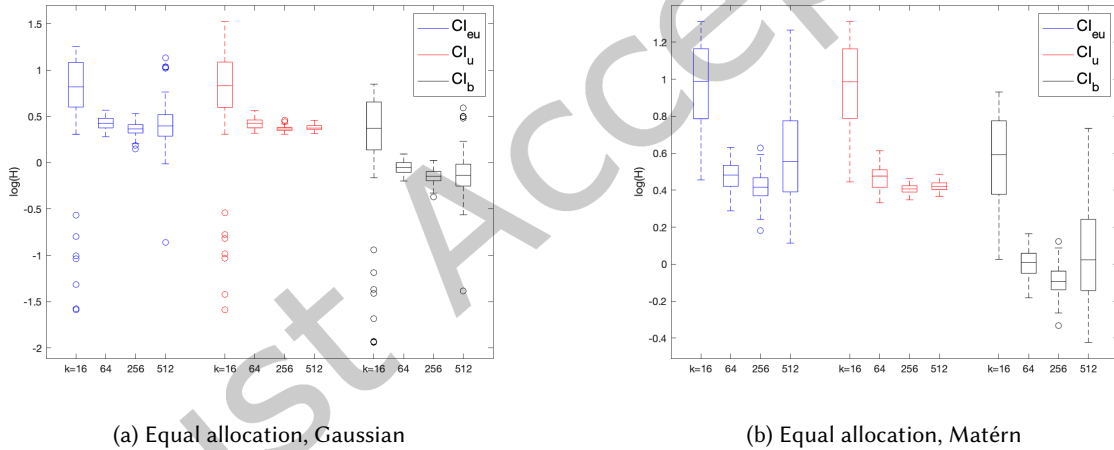
Fig. 2. Boxplots of $\log(H_{\max})$ of CI_{eu}, CI_u, and CI_b for the 1D example obtained on 100 independent macro-replications under LHD with $B = 2560$.

We next investigate the 5D case. Compared to the 1D case, notice that the true mean function $f(\cdot)$ in this 5D case is more complex and changes more rapidly across its input space. Tables 5 and 6 present the $\widehat{SCP}$'s of CI_{eu}, CI_u, and CI_b obtained under various experimental settings. We observe that, similar to the 1D case, the performance of CI_b built with the Gaussian kernel is notably weak, regardless of the budget size B. In contrast, CI_b built with the Matérn kernel delivers relatively robust performance under the same experimental settings. Furthermore, in this 5D case, an increase in the number of design points k is associated with higher $\widehat{SCP}$'s for CI_b especially when the Matérn kernel is adopted, underscoring the need of using a dense design with more design points for capturing the function in high-dimensional problems and enhancing the reliability of the confidence
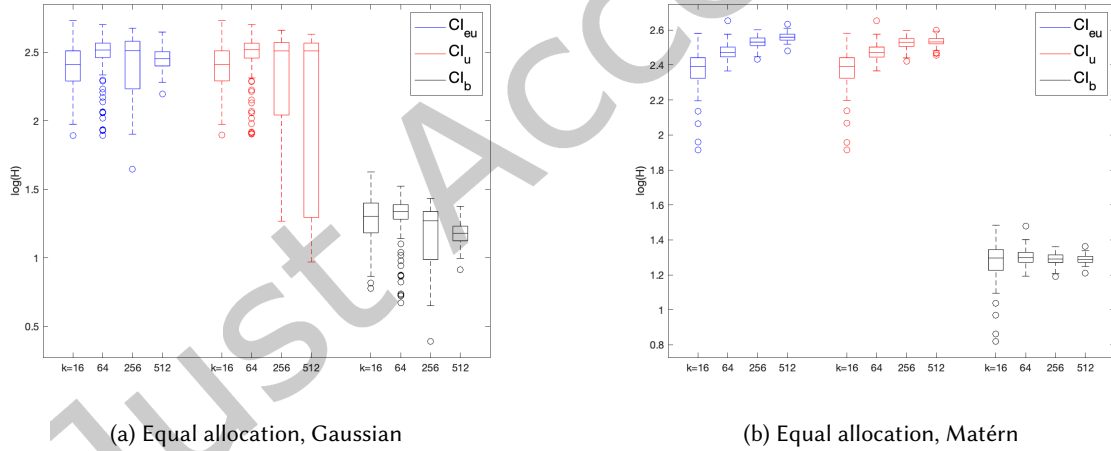
(a) Equal allocation, Gaussian
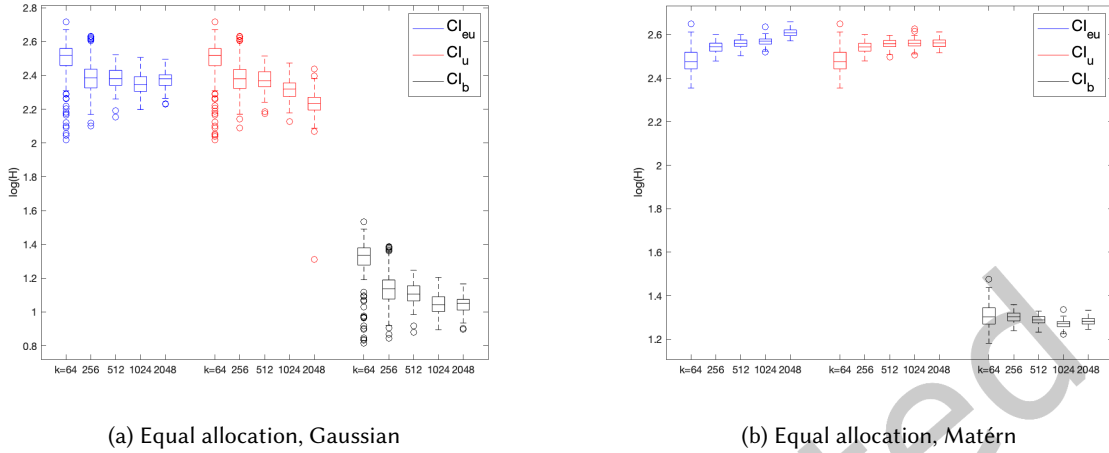

(b) Equal allocation, Matérn
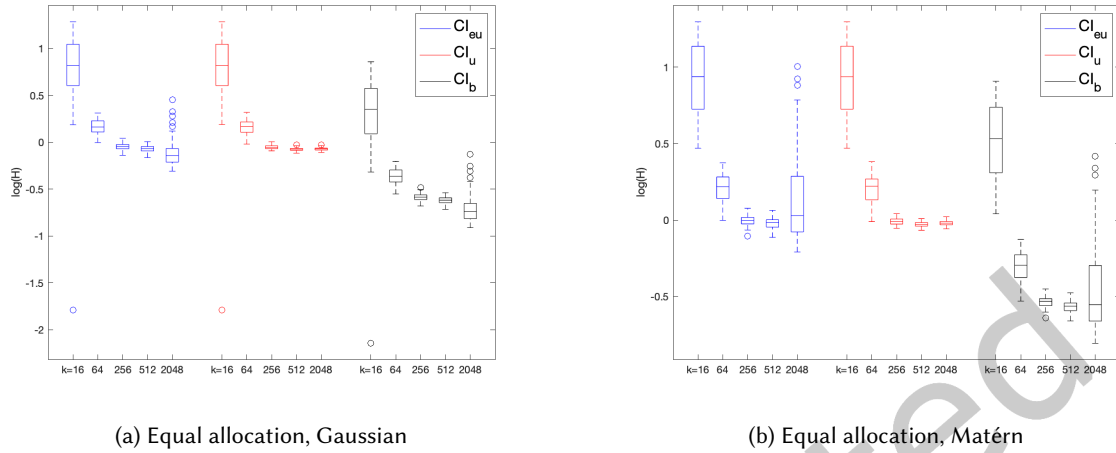
Fig. 3. Boxplots of $\log(H_{\max})$ of CI_{eu}, CI_u, and CI_b for the 1D example obtained on 100 independent macro-replications under LHD with $B = 10240$.

bound built. Regarding CI_{eu} and CI_u, the Matérn kernel also facilitates consistently strong simultaneous coverage performance across different experimental settings, highlighting its suitability for modeling highly nonlinear and rapidly changing response surfaces in high-dimensional problems. In contrast, due to the limitations of the Gaussian kernel in modeling the mean surface $f(\cdot)$ in this 5D case, the $\widehat{SCP}$'s of CI_{eu} and CI_u suffer especially when the number of design points k is small.

Table 5. $\widehat{SCP}$'s of CI_{eu}, CI_u, and CI_b obtained for the 5D case under LHD.

		Gaussian									Matérn								
B	k	Equal allocation			Unequal allocation 1			Unequal allocation 2			Equal allocation			Unequal allocation 1			Unequal allocation 2		
		CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u
2560	16	0	0.51	0.49	0	0.51	0.50	0	0.52	0.51	0.10	1	1	0.07	1	1	0.08	1	1
	64	0	1	1	0	1	1	0	0.99	0.99	0.39	1	1	0.44	1	1	0.36	1	1
	256	0.03	1	1	0.01	1	1	0	1	1	0.79	1	1	0.70	1	1	0.78	1	1
	512	0.17	1	0.97	0.17	1	0.99	0.16	1	0.98	0.80	1	1	0.78	1	1	0.78	1	1
10240	64	0	1	1	0	0.99	0.99	0	1	1	0.32	1	1	0.32	1	1	0.34	1	1
	256	0.01	1	1	0.01	1	1	0.01	1	1	0.65	1	1	0.70	1	1	0.64	1	1
	512	0.08	1	1	0.05	1	1	0.10	1	1	0.90	1	1	0.90	1	1	0.84	1	1
	1024	0.17	1	1	0.19	1	1	0.19	1	1	0.92	1	1	0.89	1	1	0.92	1	1
	2048	0.24	1	1	0.24	1	1	0.22	1	0.99	0.96	1	1	0.97	1	1	0.99	1	1

Figures 4 and 5 summarize the magnitudes of $H_{\max}$ (in a logarithmic scale) of CI_{eu}, CI_u, and CI_b, employing LHD with a total budget $B = 2560$ and 10240. For simplicity, the discussion below focuses on the results obtained under LHD with the equal allocation scheme. Similar observations apply to the results obtained under LHD with the two unequal budget allocation schemes and under the uniform design with all three budget allocations, which are detailed in Figures 11 and 12 provided in Appendix D.2.2. First and foremost, the variation exhibited by the $H_{\max}$ values of all bounds built with the two kernels demonstrates the notable difference between the two kernels in this 5D case as the number of design points k increases, especially when the given budget B is small. Specifically, the Gaussian kernel leads to increased instability in modeling as the number of design points k grows,

Table 6. $\widehat{SCP}$'s of CI_{eu}, CI_u, and CI_b obtained for the 5D case under the uniform design.

B	k	Gaussian									Matérn								
		Equal allocation			Unequal allocation 1			Unequal allocation 2			Equal allocation			Unequal allocation 1			Unequal allocation 2		
		CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u	CI_b	CI_{eu}	CI_u
2560	16	0	0.69	0.71	0	0.74	0.72	0	0.73	0.73	0.14	1	1	0.12	1	1	0.12	1	1
	64	0	0.99	1	0	1	1	0	1	1	0.33	1	1	0.22	1	1	0.33	1	1
	256	0	1	1	0	1	1	0	1	1	0.65	1	1	0.60	1	1	0.66	1	1
	512	0.24	1	0.99	0.21	1	0.96	0.23	1	0.99	0.80	1	1	0.77	1	1	0.72	1	1
10240	64	0	1	1	0	1	1	0	1	1	0.19	1	1	0.19	1	1	0.20	1	1
	256	0	1	1	0.01	1	1	0	1	1	0.57	1	1	0.58	1	1	0.58	1	1
	512	0.09	1	1	0.11	1	1	0.09	1	1	0.83	1	1	0.86	1	1	0.76	1	1
	1024	0.24	1	1	0.18	1	1	0.23	1	1	0.92	1	1	0.92	1	1	0.95	1	1
	2048	0.28	1	0.99	0.23	1	0.99	0.27	1	0.99	0.95	1	1	0.95	1	1	0.92	1	1

contrasting with the Matérn kernel which maintains consistent modeling stability across the range of k values considered; compare Figure 4 (a) and (b). Furthermore, given a large budget B, the H_{max}'s of all bounds built using the Gaussian kernel show a decreasing trend as the number of design points k increases. However, such a trend seems absent when the Matérn kernel is adopted. This observation explains the corresponding low $\widehat{SCP}$'s of CI_b observed in Tables 5 and 6. An explanation is that the predictive variance (hence the confidence bound width) associated with the Gaussian kernel diminishes faster compared with the Matérn kernel, as indicated by Corollaries 1 and 2 and the comments immediately following them.

(a) Equal allocation, Gaussian

(b) Equal allocation, Matérn

Fig. 4. Boxplots of $\log(H_{max})$ of CI_{eu}, CI_u, and CI_b for the 5D example obtained on 100 independent macro-replications under LHD with $B = 2560$.

Finally, we summarize key insights obtained from the numerical evaluations. When applying SK for meta-modeling purposes, selecting an appropriate kernel suitable for modeling the function of interest is crucial. The Gaussian kernel performs well in queueing applications like the M/M/1 example considered here [53]. In contrast, only the Matérn kernel with smoothness parameter $v = 5/2 + d/2$ was considered by [23] from where the dimension-flexible example originated, which also showed its competence in our evaluations. Therefore, choosing an appropriate kernel function is example-dependent and should take into account the features of the function of

(a) Equal allocation, Gaussian

(b) Equal allocation, Matérn

Fig. 5. Boxplots of $\log(H_{\max})$ of CI_{eu}, CI_{u}, and CI_{b} for the 5D example obtained on 100 independent macro-replications under LHD with $B = 10240$.

interest. The performance of Bonferroni correction-based simultaneous confidence intervals is sensitive to this choice, and they can be unreliable even when built by adopting an appropriate experimental design. In contrast, the uniform error bound is more robust, especially in high-dimensional problems. Second, fixed designs can outperform random designs in facilitating stability in SK model fitting and simultaneous coverage performance. Lastly, unequal allocation schemes can enhance the performance of confidence bounds, but their impact is limited.

7 Conclusion

In this paper, we proposed a method for constructing a uniform error bound with a prescribed confidence level for the SK predictor. The investigation of the asymptotic properties of the proposed uniform error bound delves into analyzing the convergence rate of SK's predictive variance under the supremum norm. The work sheds light on suitable design-point sampling schemes and budget allocation schemes for SK metamodeling to ensure desirable large-sample predictive performance. Furthermore, we conducted theoretical and numerical investigations into the impact of noise variance estimation on the empirical uniform error bound's performance. Our comprehensive numerical evaluations corroborated the theoretical findings and demonstrated the competence of the uniform error bound in implementation under various experimental settings.

In analyzing the convergence rates of SK's predictive variance under the supremum norm, we derived an upper bound for SK's predictive variance, characterized this upper bound as the solution to a regularized kernel based approximation (or equivalently, a kernel ridge regression, KRR) problem with noiseless observations of kernel function values, and identified the maximum noise variance across design points as playing a crucial role in effectively serving as the regularization parameter in the KRR problem. The conditions on the regularization parameter and experimental designs to ensure the solution to the KRR problem achieves the fastest convergence rate under the supremum norm then translate into conditions on budget allocation and design-point selection for SK's predictive variance under the supremum norm and hence the proposed uniform error bound to achieve a desirable convergence rate. It is worth noting that, for uncertainty quantification of GPR modeling subject to the impact homoscedastic noise, Wang (2021) examined the reliability of the pointwise GP confidence interval under model misspecification [44]. The study also recognized the role played by the ratio of the noise variance to the

spatial variance as the regularization parameter in GPR, considering that its value can decrease with the increase in the number of design points. However, it is crucial to highlight that the value of this regularization parameter is externally imposed without a connection to budget allocation and hence no insights were derived. A caveat to note is that the results derived in this work relies on the assumption that the mean function of interest can be accurately modeled by a GP with the specified kernel function. Our numerical examples demonstrated that selecting an appropriate kernel function is example-dependent and should take into account the features of the function of interest.

There are several promising avenues for future research. Firstly, despite its desirable simultaneous coverage guarantee and relative robustness to kernel choice and design-point sampling schemes, the uniform bound often exhibits wider interval widths compared to the Bonferroni bound, resulting in reduced sharpness in specific applications. One potential approach to enhancing the uniform bound's performance is to reduce $\beta(\tau)$ and $\gamma_{k,B}(\tau)$ in (7) by optimizing the choice of the grid constant τ and employing suitable experimental designs. Secondly, investigating the reliability of the uniform error bound under model misspecification is a priority. Moreover, extensions to other cases, such as f being a sample of a non-zero mean GP, a nonstationary GP, or unsuitable to be modeled as a sample of a GP, are promising directions to explore. Thirdly, studying the effects of GPR hyper-parameter estimation on the uniform error bound's accuracy warrants a detailed examination. Moreover, given that substituting sample variances with smoothed variance estimates enhances SK model fitting and predictive performance [45, 46], using smoothed variance estimates is anticipated to improve the simultaneous coverage performance of the uniform bound. We intend to investigate this impact, building upon the discussion in Subsection 5.2. Lastly, developing a numerically stable and computationally efficient uniform error bound for large-scale GPR applications under heteroscedasticity presents an intriguing challenge.

Acknowledgments

References

[1] A. Alaa and M. van der Schaar. 2017. Bayesian inference of individualized treatment effects using multi-task Gaussian processes. In *The Proceedings of the 31st Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 3427–3435.

[2] B. E. Ankenman, B. L. Nelson, and J. Staum. 2010. Stochastic kriging for simulation metamodeling. *Operations Research* 58 (2010), 371–382.

[3] R. Arcangéli, M. C. López de Silanes, and J. J. Torrens. 2012. Extension of sampling inequalities to Sobolev semi-norms of fractional order and derivative data. *Numer. Math.* 121 (2012), 587–608.

[4] P. Bickel and M. Rosenblatt. 1973. On some global measures of deviations of density function stimates. *The Annals of Statistics* 31 (1973), 1852–1884.

[5] L. Cai, L. Gu, Q. Wang, and S. Wang. 2021. Simultaneous confidence bands for nonparametric regression with missing covariate data. *Annals of the Institute of Statistical Mathematics* 73 (2021), 1249–1279.

[6] L. Cai, R. Liu, S. Wang, and L. Yang. 2019. Simultaneous confidence bands for mean and variance functions based on deterministic design. *Statistica Sinica* 29 (2019), 505–525.

[7] T. Cai, M. Low, and Z. Ma. 2014. Adaptive confidence bands for nonparametric regression functions. *J. Amer. Statist. Assoc.* 109 (2014), 1054–1070.

[8] G. Cao, L. Wang, Y. Li, and L. Yang. 2016. Oracle-efficient confidence envelopes for covariance functions in dense functional data. *Statistica Sinica* 26 (2016), 359–383.

[9] A. Carpentier, R. Munos, and A. Antos. 2015. Adaptive strategy for stratified Monte Carlo sampling. *Journal of Machine Learning Research* 16 (2015), 2231–2271.

[10] X. Chen, B. E. Ankenman, and B. L. Nelson. 2012. The effects of common random numbers on stochastic kriging metamodels. *ACM Transactions on Modeling and Computer Simulation* 22, Article 7 (2012), 20 pages.

[11] X. Chen and K.-K. Kim. 2014. Stochastic kriging with biased sample estimates. *ACM Transactions on Modeling and Computer Simulation* 24, Article 8 (2014), 23 pages.

[12] X. Chen and Q. Zhou. 2017. Sequential design strategies for mean response surface metamodeling via stochastic kriging with adaptive exploration and exploitation. *European Journal of Operational Research* 262 (2017), 575–585.

[13] S. R. Chowdhury and A. Gopalan. 2017. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning*. JMLR.org, 844–853.

[14] K. De Brabanter, J. De Brabanter, and J. A. K. Suykens. 2011. Approximate confidence and prediction intervals for least squares support vector regression. *IEEE Transactions on Neural Networks* 22 (2011), 110–120.

[15] G. Dellino, J. P. C. Kleijnen, and C. Meloni. 2012. Robust optimization in simulation: Taguchi and Krige combined. *INFORMS Journal on Computing* 24 (2012), 471–484.

[16] A. B. Dieker and S.-H. Kim. 2021. Efficient fully sequential indifference-zone procedures using properties of multidimensional Brownian motion exiting a sphere. https://arxiv.org/abs/2104.08784

[17] S. Efromovich. 2007. Sequential design and estimation in heteroscedastic nonparametric regression. *Sequential Analysis* 26 (2007), 3–25.

[18] R. Eubank and P. Speckman. 1993. Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.* 88 (1993), 1287–1301.

[19] P. I. Frazier. 2014. A fully sequential elimination procedure for indifference-zone ranking and selection with tight Bounds on probability of correct selection. *Operations Research* 62 (2014), 926–942.

[20] Z. Geng, F. Yang, X. Chen, and N. Wu. 2015. Gaussian process based modeling and experimental design for sensor calibration in drifting environments. *Sensors and Actuators B: Chemical* 216 (2015), 321–331.

[21] L. Gu and L. Yang. 2015. Oracally efficient estimation for single-index link function with simultaneous confidence Band. *Electronic Journal of Statistics* 9 (2015), 1540–1561.

[22] G. Johnston. 1982. Probabilities of maximal deviations for nonparametric regression function estimates. *Journal of Multivariate Analysis* 12 (1982), 402–414.

[23] B. Kamiński. 2015. A method for the updating of stochastic kriging metamodels. *European Journal of Operational Research* 247 (2015), 859–866.

[24] I. Keilegom and G. Claeskens. 2003. Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics* 31 (2003), 1852–1884.

[25] J. P. C. Kleijnen. 2015. *Design and analysis of simulation experiments* (2nd ed.). Springer, New York.

[26] J. P. C. Kleijnen and W. C. M. van Beers. 2022. Statistical tests for cross-validation of Kriging models. *INFORMS Journal on Computing* 34 (2022), 607–621.

[27] B. Laurent and P. Massart. 2000. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics* 28 (2000), 1302–1338.

[28] L. Le Gratiet and J. Garnier. 2015. Asymptotic analysis of the learning curve for Gaussian process regression. *Machine Learning* 98 (2015), 407–433.

[29] A. Lederer and J. Umlauft. 2019. Uniform error bounds for Gaussian process regression with application to safe control. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (Eds.). Curran Associates Inc., Vancouver, BC, Canada, 659–669.

[30] Z. Liu and M. Li. 2020. Non-asymptotic analysis in kernel ridge regression. https://arxiv.org/pdf/2006.01350v1.pdf

[31] Z. Liu and M. Li. 2023. On the estimation of derivatives using plug-in kernel ridge regression estimators. *Journal of Machine Learning Research* 24 (2023), 1–37.

[32] X. Lu, A. Rudi, E. Borgonovo, and L. Rosasco. 2020. Faster kriging: Facing high-dimensional simulators. *Operations Research* 68 (2020), 233–249.

[33] C. Ma, R. Pathak, and M. J. Wainwright. 2023. Optimally tackling covariate shift in RKHS-based nonparametric regression. *The Annals of Statistics* 51 (2023), 738–761.

[34] S. H. Ng and J. Yin. 2012. Bayesian kriging analysis and design for stochastic simulations. *ACM Transactions on Modeling and Computer Simulation* 22, Article 17 (2012), 26 pages.

[35] C. E. Rasmussen and C. K. I. Williams. 2006. *Gaussian processes for machine learning*. The MIT Press, Cambridge, MA, USA.

[36] C. Rieger and B. Zwicknagl. 2010. Sampling inequalities for infinitely smooth functions, with applications to interpolation and machine learning. *Advances in Computational Mathematics* 32 (2010), 103–129.

[37] T. J. Santner, B. J. Williams, and W. I. Notz. 2003. *The design and analysis of computer experiments*. Springer, New York.

[38] R. Schüssler and M. Trede. 2016. Constructing minimum-width confidence bands. *Economics Letters* 145 (2016), 182–185.

[39] M. W. Seeger, S. M. Kakade, and D. P. Foster. 2008. Information consistency of nonparametric Gaussian process methods. *IEEE Transactions on Information Theory* 54 (2008), 2376–2382.

[40] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. 2012. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory* 58 (2012), 3250–3265.

[41] A. L. Teckentrup. 2020. Convergence of Gaussian process regression with estimated hyper-parameters and applications in Bayesian inverse Problems. *SIAM/ASA Journal on Uncertainty Quantification* 8 (2020), 1310–1337.

[42] M. J. Wainwright. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, Cambridge.

[43] J. Wang, F. Cheng, and L. Yang. 2013. Smooth simultaneous confidence bands for cumulative distribution functions. *Journal of Nonparametric Statistics* 25 (2013), 395–407.

[44] W. Wang. 2021. On the inference of applying Gaussian process modeling to a deterministic function. *Electronic Journal of Statistics* 15 (2021), 5014–5066.

[45] W. Wang and X. Chen. 2016. The effects of estimation of heteroscedasticity on stochastic kriging. In *Proceedings of the 2016 Winter Simulation Conference*, T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick (Eds.). IEEE, Piscataway, New Jersey, 326–337.

[46] W. Wang and X. Chen. 2018. An adaptive two-stage dual metamodeling approach for stochastic simulation experiments. *IISE Transactions* 50 (2018), 820–836.

[47] W. Wang, R. Tuo, and C. F. J. Wu. 2020. On prediction properties of kriging: Uniform error bounds and robustness. *J. Amer. Statist. Assoc.* 115 (2020), 920–930.

[48] H. Wendland. 2005. *Scattered data approximation*. Cambridge University Press, Cambridge, UK.

[49] H. Wendland and C. Rieger. 2005. Approximate interpolation with applications to selecting smoothing parameters. *Numer. Math.* 101 (2005), 729–748.

[50] W. Whitt. 1989. Planning queueing simulations. *Management Science* 35 (1989), 1341–1366.

[51] G. Wynne, F.-X. Briol, and M. Girolami. 2021. Convergence guarantees for Gaussian process means with misspecified likelihoods and smoothness. *Journal of Machine Learning Research* 22 (2021), 1–40.

[52] G. Xie and X. Chen. 2020. Uniform error bounds for stochastic kriging. In *Proceedings of the 2020 Winter Simulation Conference*, K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing (Eds.). IEEE, Piscataway, New Jersey, 361–372.

[53] W. Xie, B. Nelson, and J. Staum. 2010. The Influence of Correlation Functions on Stochastic Kriging Metamodels. In *Proceedings of the 2010 Winter Simulation Conference*, B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan (Eds.). IEEE, Piscataway, New Jersey, 1067–1078.

[54] Y. Yun, A. Bhattacharya, and D. Pati. 2017. Frequentist coverage and sup-norm convergence rate in Gaussian process regression. https://arxiv.org/abs/1708.04753v1

[55] Y. Zhang and X. Chen. 2021. Information consistency of stochastic kriging and its implications. In *Proceedings of the 2021 Winter Simulation Conference*, S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper (Eds.). IEEE, Piscataway, New Jersey, 1–12.

[56] Y. Zhang, J. Duchi, and M. Wainwright. 2015. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research* 16 (2015), 3299–3340.

A Proofs in Section 5

A.1 Proof of Proposition 5

PROOF. We start by presenting some definitions and results essential for the subsequent development.

Definition 3 (DEFINITION 2.7 ON PAGE 26 OF [42]). *A random variable X with mean $\mu = \mathrm{E}[X]$ is sub-exponential if there are non-negative parameters (v, α) such that*

$$\mathrm{E}[e^{t(X-\mu)}] \le e^{\frac{v^2 t^2}{2}}, \qquad \text{for all } |t| < \frac{1}{\alpha}.$$

Lemma 1 (CONCENTRATION INEQUALITY FOR CHI-SQUARED VARIABLES, PAGE 29 OF WAINWRIGHT [42]). *A chi-squared $\mathcal{X}^2$ random variable with n degrees of freedom, denoted by $Y \sim \mathcal{X}_n^2$, is sub-exponential with parameters $(v, \alpha) = (2\sqrt{n}, 4)$, and the two-sided tail bound follows as*

$$\mathrm{P}\left(\left|\frac{1}{n}Y - 1\right| \ge t\right) \le 2e^{-nt^2/8}, \qquad \text{for all } t \in (0, 1).$$

We are now ready to prove Proposition 5. At any given design point $\mathbf{x}_i$ ($i = 1, 2, \ldots, k$), consider the random variable $(n_i - 1)\widehat{V}(\mathbf{x}_i)/V(\mathbf{x}_i)$. Since the noise terms incurred at $\mathbf{x}_i$, $\varepsilon_j(\mathbf{x}_i)$'s, are i.i.d. normal with mean zero and variance $V(\mathbf{x}_i)$, it follows that $(n_i - 1)\widehat{V}(\mathbf{x}_i)/V(\mathbf{x}_i) \sim \mathcal{X}_{n_i-1}^2$, a $\mathcal{X}^2$ random variable with $(n_i - 1)$ degrees of freedom.

By Lemma 1, we have

$$P\left(\left|\frac{\widehat{V}(\mathbf{x}_i)}{V(\mathbf{x}_i)} - 1\right| \geq t\right) \leq 2e^{-(n_i-1)t^2/8}, \qquad \text{for all } t \in (0,1). \tag{A.17}$$

Define events $\mathcal{A}_i := \left\{\left|\frac{\widehat{V}(\mathbf{x}_i)-V(\mathbf{x}_i)}{n_i}\right| \leq t \cdot \frac{V(\mathbf{x}_i)}{n_i}\right\}$ and $\mathcal{B}_i := \left\{\left|\frac{\widehat{V}(\mathbf{x}_i)-V(\mathbf{x}_i)}{n_i}\right| \leq t \cdot V_{k,\max}\right\}$ for $i = 1, 2, \ldots, k$. In light of (A.17) and $V_{k,\max} = \max_{1 \leq i \leq k} \frac{V(\mathbf{x}_i)}{n_i}$, we have for each $\mathbf{x}_i$ given,

$$P\left(\mathcal{B}_i^c\right) \leq P\left(\mathcal{A}_i^c\right) \leq 2e^{-(n_i-1)t^2/8}, \qquad \text{for all } t \in (0,1). \tag{A.18}$$

In particular, it follows from (A.18) that, for all $t \in (0,1)$,

$$P\left(\max_{1 \leq i \leq k}\left|\frac{\widehat{V}(\mathbf{x}_i)-V(\mathbf{x}_i)}{n_i}\right| \geq t \cdot V_{k,\max}\right) = P\left(\cup_{i=1}^k \mathcal{B}_i^c\right) \leq 2\sum_{i=1}^k e^{-(n_i-1)t^2/8} \leq 2ke^{-(n_k^*-1)t^2/8}. \tag{A.19}$$

Set $e^{-\eta_{k,B}} = 2ke^{-(n_k^*-1)t^2/8}$ for any $t \in (0,1)$. Correspondingly, a given $\eta_{k,B} > 0$ corresponds to $t = \sqrt{\frac{8(\eta_{k,B}+\log(2k))}{(n_k^*-1)}}$. Then (A.19) can be equivalently written as

$$P\left(\max_{1 \leq i \leq k}\left|\frac{\widehat{V}(\mathbf{x}_i)-V(\mathbf{x}_i)}{n_i}\right| \geq \sqrt{\frac{8(\eta_{k,B}+\log(2k))}{(n_k^*-1)}} \cdot V_{k,\max}\right) \leq e^{-\eta_{k,B}}. \tag{A.20}$$

Now, if we set $\eta_{k,B} = \log\left((\pi_k\alpha)^{-1}\right)$ so that $\sum_{k=1}^{\infty} \pi_k = 1/3$, then applying the union bounds over all $k \geq 1$ yields

$$\max_{1 \leq i \leq k}\left|\frac{\widehat{V}(\mathbf{x}_i)-V(\mathbf{x}_i)}{n_i}\right| \leq \sqrt{\frac{8(\eta_{k,B}+\log(2k))}{(n_k^*-1)}} \cdot V_{k,\max}, \quad \forall k \geq 1$$

holds true with probability of at least $1 - \alpha/3$.

$\square$

A.2 Proof of Theorem 3

Proof. Define $\Sigma := \Sigma_M + \Sigma_\varepsilon$ and $\widehat{\Sigma} := \Sigma_M + \widehat{\Sigma}_\varepsilon$. And let $\Delta\Sigma_\varepsilon := \widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon$. Recall the definitions of $\widehat{LB}(\mathbf{x})$ and $\widehat{UB}(\mathbf{x})$ of the empirical uniform bound. We have

$$\left|\widehat{LB}(\mathbf{x}) - LB(\mathbf{x})\right| = \left|\widehat{UB}(\mathbf{x}) - UB(\mathbf{x})\right| \leq \Delta\mu_{k,B}(\mathbf{x}) + \sqrt{\beta(\tau)}\Delta\sigma_{k,B}(\mathbf{x}) + \Delta\gamma_{k,B}(\tau), \tag{A.21}$$

where for all $\mathbf{x} \in \mathcal{X}$,

$$\Delta\mu_{k,B}(\mathbf{x}) := \left|\widehat{\mu}_{k,B}(\mathbf{x}) - \mu_{k,B}(\mathbf{x})\right| = \left|\Sigma_M(\mathbf{x}_0, \mathbf{X})^\top\left(\widehat{\Sigma}^{-1} - \Sigma^{-1}\right)\bar{\mathcal{Y}}\right|,$$

$$\Delta\sigma_{k,B}(\mathbf{x}) := \sqrt{\Delta\sigma_{k,B}^2(\mathbf{x})},$$

$$\Delta\sigma_{k,B}^2(\mathbf{x}) := \left|\widehat{\sigma}_{k,B}^2(\mathbf{x}) - \sigma_{k,B}^2(\mathbf{x})\right| = \left|\Sigma_M(\mathbf{x}_0, \mathbf{X})^\top\left(\widehat{\Sigma}^{-1} - \Sigma^{-1}\right)\Sigma_M(\mathbf{x}_0, \mathbf{X})\right|,$$

$$\Delta\gamma_{k,B}(\tau) := \Delta L_{\mu_{k,B}}\tau + \sqrt{\beta(\tau)}\Delta\omega_{\sigma_{k,B}}(\tau), \tag{A.22}$$

with $\Delta L_{\mu_{k,B}} := \left|\widehat{L}_{\mu_{k,B}} - L_{\mu_{k,B}}\right|$ and $\Delta\omega_{\sigma_{k,B}}(\tau) := \left(2\tau L_\Sigma k\left\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\right\|\max_{\mathbf{x},\mathbf{x}' \in \mathcal{X}}\Sigma_M(\mathbf{x}, \mathbf{x}')\right)^{\frac{1}{2}}$.

Recall from (4) that

$$\omega_{\sigma_{k,B}}(\tau) = \sqrt{2\tau L_\Sigma \left(1 + k\,\|\Sigma^{-1}\| \max_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \Sigma_M(\mathbf{x},\mathbf{x}')\right)},$$

and $\widehat{\omega}_{\sigma_{k,B}}(\tau)$ is obtained by replacing Σ with $\widehat{\Sigma}$ in the equation above. It follows from $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for any $a, b \ge 0$ that $\left|\widehat{\omega}_{\sigma_{k,B}}(\tau) - \omega_{\sigma_{k,B}}(\tau)\right| \le \Delta\omega_{\sigma_{k,B}}(\tau)$.

The first part of the proof is identical to that of Theorem 2 but by replacing $\sum_{k=1}^{\infty} \pi_k = 1/2$ with $\sum_{k=1}^{\infty} \pi_k = 1/3$ in (8) and (10). The second part of the proof focuses on showing that $\left|\widehat{LB}(\mathbf{x}) - LB(\mathbf{x})\right| = \left|\widehat{UB}(\mathbf{x}) - UB(\mathbf{x})\right| = o\left(\sqrt{\beta(\tau)}\sigma_{k,B}(\mathbf{x}) + \gamma_{k,B}(\tau)\right)$ for all $\mathbf{x} \in \mathcal{X}$ despite the impact of noise variance estimation.

From Proposition 5, we have with probability at least $1 - \alpha/3$,

$$\max_{1\le i\le k} \left|\frac{\widehat{V}(\mathbf{x}_i) - V(\mathbf{x}_i)}{n_i}\right| \lesssim \left(\frac{\log k}{n_k^*}\right)^{\frac{1}{2}} \cdot V_{k,\max}. \tag{A.23}$$

By the Woodbury matrix identity, we have

$$\widehat{\Sigma}^{-1} - \Sigma^{-1} = \Sigma^{-1}\left(\mathbf{I} + \Delta\Sigma_\varepsilon\Sigma^{-1}\right)^{-1}\Delta\Sigma_\varepsilon\Sigma^{-1}. \tag{A.24}$$

It follows that

$$\left\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\right\| \le \left\|\Sigma^{-1}\right\|\left\|\left(\mathbf{I} + \Delta\Sigma_\varepsilon\Sigma^{-1}\right)^{-1}\right\|\left\|\Delta\Sigma_\varepsilon\Sigma^{-1}\right\| \le \left\|\Sigma^{-1}\right\|\left\|\Delta\Sigma_\varepsilon\Sigma^{-1}\right\|, \tag{A.25}$$

where, thanks to Weyl's Theorem, we have used

$$\left\|\left(\mathbf{I} + \Delta\Sigma_\varepsilon\Sigma^{-1}\right)^{-1}\right\| = \lambda_{\max}\left(\left(\mathbf{I} + \Delta\Sigma_\varepsilon\Sigma^{-1}\right)^{-1}\right) = \frac{1}{\lambda_{\min}\left(\mathbf{I} + \Delta\Sigma_\varepsilon\Sigma^{-1}\right)} \le 1, \tag{A.26}$$

and $\lambda_{\min}\left(\mathbf{I} + \Delta\Sigma_\varepsilon\Sigma^{-1}\right) \le 1$ since $\Delta\Sigma_\varepsilon\Sigma^{-1}$ is positive definite.

First, we analyze the terms in $\Delta\gamma_{k,B}(\tau)$ defined in (A.22). It follows that, with probability of at least $1 - \alpha/3$,

$$\begin{aligned}
\Delta L_{\mu_{k,B}}\tau &= L_\Sigma\sqrt{k}\left\|\left(\widehat{\Sigma}^{-1} - \Sigma^{-1}\right)\bar{\mathcal{Y}}\right\|\tau \\
&\overset{(i)}{\le} L_\Sigma\sqrt{k}\left\|\Sigma^{-1}\right\|\left\|\Delta\Sigma_\varepsilon\Sigma^{-1}\bar{\mathcal{Y}}\right\|\tau \\
&\le L_\Sigma\sqrt{k}\left\|\Sigma^{-1}\right\|\left\|\Delta\Sigma_\varepsilon\right\|\left\|\Sigma^{-1}\bar{\mathcal{Y}}\right\|\tau \\
&= L_\Sigma\sqrt{k}\left\|\Sigma^{-1}\bar{\mathcal{Y}}\right\|\tau\left\|\Sigma^{-1}\right\|\left\|\Delta\Sigma_\varepsilon\right\| \\
&\overset{(ii)}{\lesssim} L_{\mu_{k,B}}\tau \cdot V_{k,\min}^{-1} \cdot \left(\frac{\log k}{n_k^*}\right)^{\frac{1}{2}} \cdot V_{k,\max} \\
&\overset{(iii)}{\asymp} L_{\mu_{k,B}}\tau\left(\frac{\log k}{n_k^*}\right)^{\frac{1}{2}} = o(L_{\mu_{k,B}}\tau)
\end{aligned} \tag{A.27}$$

if $\log k = o(n_k^*)$ as $k, B \to \infty$. In (A.27), step (i) on the right-hand side follows from (A.24) and (A.26), step (ii) follows from (4), (A.23), and that $\|\Sigma^{-1}\| \le 1/\lambda_{\min}(\Sigma_\varepsilon) = V_{k,\min}^{-1}$, and step (iii) follows since $V_{k,\max}/V_{k,\min} < \infty$.

Next, we have that with probability of at least $1 - \alpha/3$

$$
\sqrt{\beta(\tau)}\Delta\omega_{\sigma_{k,B}}(\tau) = \sqrt{\beta(\tau)}\left(2\tau L_{\Sigma}k\left\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\right\|\max_{\mathbf{x},\mathbf{x}'\in\mathcal{X}}\Sigma_{\mathsf{M}}(\mathbf{x},\mathbf{x}')\right)^{\frac{1}{2}}
$$

$$
\overset{(i)}{\leq} \sqrt{\beta(\tau)}\left(2\tau L_{\Sigma}k\left\|\Sigma^{-1}\right\|\left\|\Delta\Sigma_{\varepsilon}\right\|\left\|\Sigma^{-1}\right\|\max_{\mathbf{x},\mathbf{x}'\in\mathcal{X}}\Sigma_{\mathsf{M}}(\mathbf{x},\mathbf{x}')\right)^{\frac{1}{2}}
$$

$$
\overset{(ii)}{\leq} \left(\beta(\tau)\cdot 2\tau\cdot L_{\Sigma}\cdot k\cdot\max_{\mathbf{x},\mathbf{x}'\in\mathcal{X}}\Sigma_{\mathsf{M}}(\mathbf{x},\mathbf{x}')\mathsf{V}_{k,\min}^{-1}\right)^{\frac{1}{2}}\mathsf{V}_{k,\min}^{-\frac{1}{2}}\left(\left(\frac{\log k}{n_k^*}\right)^{\frac{1}{2}}\cdot\mathsf{V}_{k,\max}\right)^{\frac{1}{2}}
$$

$$
\overset{(iii)}{\asymp} \sqrt{\beta(\tau)}\omega_{\sigma_{k,B}}(\tau)\left(\frac{\log k}{n_k^*}\right)^{\frac{1}{4}} = o\left(\sqrt{\beta(\tau)}\omega_{\sigma_{k,B}}(\tau)\right) \tag{A.28}
$$

if $\log k = o(n_k^*)$ as $k, B \to \infty$. In (A.28), step (i) on the right-hand side follows from (A.25), step (ii) follows (A.23) and that $\|\Sigma^{-1}\| \leq \mathsf{V}_{k,\min}^{-1}$, and step (iii) follows from (11) and that $\mathsf{V}_{k,\max}/\mathsf{V}_{k,\min} < \infty$.

Moreover, due to the union bound, we have with probability of at least $1 - 2\alpha/3$,

$$
\Delta\mu_{k,B}(\mathbf{x}) = \left|\Sigma_{\mathsf{M}}(\mathbf{x},\mathbf{X})^{\top}\left(\widehat{\Sigma}^{-1} - \Sigma^{-1}\right)\bar{\mathcal{Y}}\right|
$$

$$
\overset{(i)}{\leq} \|\Sigma_{\mathsf{M}}(\mathbf{x},\mathbf{X})\|\left\|\Sigma^{-1}\right\|\left\|\Delta\Sigma_{\varepsilon}\right\|\left\|\Sigma^{-1}\bar{\mathcal{Y}}\right\|
$$

$$
\overset{(ii)}{\lesssim} \sqrt{k}\mathsf{V}_{k,\min}^{-1}\left(\frac{\log k}{n_k^*}\right)^{\frac{1}{2}}\cdot\mathsf{V}_{k,\max}\sqrt{k}\mathsf{V}_{k,\min}^{-\frac{1}{2}}
$$

$$
\overset{(iii)}{\asymp} k\left(\frac{\log k}{n_k^*}\right)^{\frac{1}{2}}\mathsf{V}_{k,\min}^{-\frac{1}{2}} = o\left(\sqrt{\beta(\tau)}\sigma_{k,B}(\mathbf{x})\right) \quad \forall\mathbf{x}\in\mathcal{X}, \tag{A.29}
$$

if (1) in the finite smoothness case, $k^{\frac{6\nu+d}{2d}} = o(n_k^*)$, thanks to Proposition 2; and (2) in the infinite smoothness case, $k^{\frac{2d-1}{d}}(\log k)\exp\left((C + \bar{C})k^{\frac{1}{d}}\right) = o(n_k^*)$, thanks to Proposition 4, where $C, \bar{C} > 0$ are as specified in the proof of Proposition 4. In (A.29), step (i) on the right-hand side follows from (A.24) and (A.26), step (ii) follows from (A.23), (10) with $\eta_{k,B} = \log\left(1/(\pi_k\alpha)\right)$ and $\sum_{k=1}^{\infty}\pi_k = 1/2$ being replaced by $\sum_{k=1}^{\infty}\pi_k = 1/3$, and that $\|\Sigma^{-1}\| \leq \mathsf{V}_{k,\min}^{-1}$, and step (iii) follows since $\mathsf{V}_{k,\max}/\mathsf{V}_{k,\min} < \infty$.

Finally, it holds with probability of $1 - \alpha/3$ that

$$
\Delta\sigma_{k,B}^2(\mathbf{x}) = \left|\Sigma_{\mathsf{M}}(\mathbf{x},\mathbf{X})^{\top}\left(\widehat{\Sigma}^{-1} - \Sigma^{-1}\right)\Sigma_{\mathsf{M}}(\mathbf{x},\mathbf{X})\right|
$$

$$
\overset{(i)}{\leq} \|\Sigma_{\mathsf{M}}(\mathbf{x},\mathbf{X})\|\left\|\Sigma^{-1}\right\|\left\|\Delta\Sigma_{\varepsilon}\Sigma^{-1}\right\|\|\Sigma_{\mathsf{M}}(\mathbf{x},\mathbf{X})\|
$$

$$
\leq \|\Sigma_{\mathsf{M}}(\mathbf{x},\mathbf{X})\|\left\|\Sigma^{-1}\right\|\|\Delta\Sigma_{\varepsilon}\|\left\|\Sigma^{-1}\right\|\|\Sigma_{\mathsf{M}}(\mathbf{x},\mathbf{X})\|
$$

$$
\overset{(ii)}{\lesssim} \sqrt{k}\mathsf{V}_{k,\min}^{-1}\left(\frac{\log k}{n_k^*}\right)^{\frac{1}{2}}\cdot\mathsf{V}_{k,\max}\cdot\mathsf{V}_{k,\min}^{-1}\sqrt{k}
$$

$$
\overset{(iii)}{\asymp} k\mathsf{V}_{k,\min}^{-1}\left(\frac{\log k}{n_k^*}\right)^{\frac{1}{2}} = o(\sigma_{k,B}^2(\mathbf{x})), \quad \forall\mathbf{x}\in\mathcal{X}, \tag{A.30}
$$

if (1) in the finite smoothness case, $k^{\frac{6\nu-d}{d}} (\log k) = o(n_k^*)$, thanks to Proposition 2; and (2) in the infinite smoothness case, $k^2 (\log k) \exp(2(\bar{C} + C)k^{\frac{1}{d}}) = o(n_k^*)$, thanks to Proposition 4, where $C, \bar{C} > 0$ are as specified in the proof of Proposition 4. In both finite and infinite smoothness cases, it follows from (A.30) $\sqrt{\beta(\tau)}\Delta\sigma_{k,B}(\mathbf{x}) = o\left(\sqrt{\beta(\tau)}\sigma_{k,B}(\mathbf{x})\right)$ for all $\mathbf{x} \in \mathcal{X}$. In (A.30), step (i) on the right-hand side follows from (A.25), step (ii) follows from (A.23) and that $\|\Sigma^{-1}\| \le \mathsf{V}_{k,\min}^{-1}$, and step (iii) follows since $\mathsf{V}_{k,\max}/\mathsf{V}_{k,\min} < \infty$.

Due to the union bound, combining (A.21), (A.27), (A.28), (A.29), and (A.30) completes the proof. □

B Reproducing Kernel Hilbert Space

We provide some necessary notation and preliminaries to reproducing kernel Hilbert space (RKHS) in this section. Denote the marginal distribution of design points on $\mathcal{X}$ by $\mathbb{P}$ and let $L_2(\mathcal{X})$ be the L_2 space with respect to $\mathbb{P}$. For $f, g : \mathcal{X} \to \mathbb{R}$, let $\langle f, g \rangle_2 = \left(\int_{\mathcal{X}} f(\mathbf{x})g(\mathbf{x}) \, d\mathbb{P}(\mathbf{x})\right)^{1/2}$ denote the inner product.

Consider a continuous and positive definite kernel function K that satisfies the Hilbert-Schmidt condition:

$$\int_{\mathcal{X} \times \mathcal{X}} K^2(\mathbf{x}, \mathbf{x}')d\mathbb{P}(\mathbf{x})d\mathbb{P}(\mathbf{x}') < \infty.$$

For any $f \in L_2(\mathcal{X})$, we define an integral operator $T_K : L_2(\mathcal{X}) \to \mathcal{H}$ via

$$T_K f(\mathbf{x}) := \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')d\mathbb{P}(\mathbf{x}'), \quad \mathbf{x} \in \mathcal{X}.$$

Mercer's Theorem [Section 4.3 of 35] states that there exists countable pairs of eigenvalues and eigenfunctions $(\mu_i, \psi_i)_{i \in \mathbb{N}} \subset (0, \infty) \times L_2(\mathcal{X})$ of T_K such that

$$T_K \psi_i = \mu_i \psi_i, \quad \forall i \in \mathbb{N},$$

where $\{\psi_i\}_{i=1}^{\infty}$ form an orthonormal basis of $L_2(\mathcal{X})$ and $\mu_1 \ge \mu_2 \ge \cdots$ are non-negative with $\lim_{i \to \infty} \mu_i = 0$. Furthermore, the kernel function K has the following expansion:

$$K(\mathbf{x}, \mathbf{x}') = \Sigma_{i=1}^{\infty} \mu_i \psi_i(\mathbf{x})\psi_i(\mathbf{x}'), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X},$$

where the convergence of the series on the right-hand side is absolute and uniform over any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. The reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with kernel K can be defined as

$$\mathcal{H} := \left\{ f \in L_2(\mathcal{X}) : \|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{\infty} \frac{f_i^2}{\mu_i} < \infty, f_i = \langle f, \psi_i \rangle_2 \right\},$$

which is equipped with the inner product $\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} f_i g_i / \mu_i$ for any $f = \sum_{i=1}^{\infty} f_i \psi_i$ and $g = \sum_{i=1}^{\infty} g_i \psi_i$ in $\mathcal{H}$.

Now, consider an estimator of a function $f_0 \in \mathcal{H}$ given by

$$f_\lambda = (T_K + \lambda \mathrm{id})^{-1} T_K f_0,$$

where id is the identify operator. It can be shown that f_λ is the solution to the following optimization problem [54]:

$$f_\lambda := \underset{f \in \mathcal{H}}{\mathrm{argmin}} \left\{ \|f - f_0\|_2^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}, \tag{A.31}$$

which is the population counterpart of the following KRR problem:

$$\widehat{f_\lambda} := \underset{f \in \mathcal{H}}{\mathrm{argmin}} \left\{ \frac{1}{k} \sum_{i=1}^{k} (f(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}.$$

The estimator f_λ in (A.31) can be obtained through the integral operator $T_{\tilde{K}}$ as $f_\lambda = T_{\tilde{K}} f_0$, where $\tilde{K}$ is the equivalent kernel for K: $\tilde{K}$ has the same eigenfunctions as K but its eigenvalues are changed to $\nu_i = \mu_i/(\mu_i + \lambda)$ for $i \in \mathbb{N}$, i.e., $\tilde{K}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \nu_i \psi_i(\mathbf{x}) \psi_i(\mathbf{x}')$. Let $\widetilde{\mathcal{H}}$ be the RKHS associated with kernel $\tilde{K}$. We have $\widetilde{\mathcal{H}}$ equivalent to $\mathcal{H}$ as a functional space, but with a different inner product:

$$\langle f, g \rangle_{\widetilde{H}} = \langle f, g \rangle_2 + \lambda \langle f, g \rangle_H,$$

and the corresponding RKHS norm is $\| \cdot \|_{\widetilde{H}}$. For more details on equivalent kernels, we refer the interested reader to [54] and Chapter 7 of [35].

C   Random Design Setting

This subsection focuses on the analysis of predictive variance of SK in the random design setting, i.e., the design points are randomly drawn from the input space $\mathcal{X}$ according to some distribution $\mathbb{P}$. Our investigation builds on the literature on the non-asymptotic error analysis for kernel ridge regression estimators [30, 54].

Consider the predictive variance $\sigma_{k,B}^2(\mathbf{x}_0)$ given in (3). We have

$$\sigma_{k,B}^2(\mathbf{x}_0) \le \Sigma_M(\mathbf{x}_0, \mathbf{x}_0) - \Sigma_M(\mathbf{x}_0, \mathbf{X})^\top \left(\Sigma_M(\mathbf{X}, \mathbf{X}) + V_{k,\max} \mathbf{I}_k \right)^{-1} \Sigma_M(\mathbf{x}_0, \mathbf{X}), \tag{A.32}$$

where recall that $V_{k,\max} = \max_{1 \le i \le k} V(\mathbf{x}_i)/n_i$. *Define $\lambda := V_{k,\max}/(\xi^2 k)$ in this subsection.* We can write the upper bound for $\sigma_{k,B}^2(\mathbf{x}_0)$ given by the right-hand side of (A.32) as $\xi^2 \left(K_{\mathbf{x}_0}(\mathbf{x}_0) - \widehat{K}_{\mathbf{x}_0}(\mathbf{x}_0) \right)$, where $K_{\mathbf{x}_0}(\mathbf{x}') := K(\mathbf{x}_0, \mathbf{x}')$ for any $\mathbf{x}' \in \mathcal{X}$ and recall from Section 2.1 that $K(\mathbf{x}, \mathbf{x}') = \Sigma_M(\mathbf{x}, \mathbf{x}')/\xi^2$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ is the kernel function; in addition,

$$\widehat{K}_{\mathbf{x}_0}(\mathbf{x}') = K(\mathbf{x}', \mathbf{X})^\top \left(K(\mathbf{X}, \mathbf{X}) + k \lambda \mathbf{I}_k \right)^{-1} K(\mathbf{X}, \mathbf{x}_0). \tag{A.33}$$

Inspired by [30] and [54], we realize that, for any given prediction point $\mathbf{x}_0 \in \mathcal{X}$, $\widehat{K}_{\mathbf{x}_0}(\cdot)$ is the solution to the following kernel ridge regression (KRR) problem with noiseless observations of $K_{\mathbf{x}_0}$ at the design points:

$$\widehat{K}_{\mathbf{x}_0} := \operatorname*{argmin}_{g \in \mathcal{H}} \left[\frac{1}{k} \sum_{i=1}^{k} \left(K_{\mathbf{x}_0}(\mathbf{x}_i) - g(\mathbf{x}_i) \right)^2 + \lambda \|g\|_{\mathcal{H}}^2 \right], \tag{A.34}$$

where $\mathcal{H}$ denotes the RKHS associated with kernel K, $\| \cdot \|_{\mathcal{H}}$ denotes the norm equipped with $\mathcal{H}$, and λ represents the regularization parameter. Hence, we can investigate the order of the upper bound $\xi^2 \left(K_{\mathbf{x}_0}(\mathbf{x}_0) - \widehat{K}_{\mathbf{x}_0}(\mathbf{x}_0) \right)$ on $\sigma_{k,B}^2(\mathbf{x}_0)$ through studying the bias of the noiseless KRR estimator $\widehat{K}_{\mathbf{x}_0}$.

C.1   Main Results

We are now in a position to state the assumption on the eigenfunctions of the kernel function K and the main result that will be useful for establishing the convergence rate of the predictive variance under the supremum norm in the random design setting.

Assumption 1. *There exists a constant $C_\psi > 0$ such that $\|\psi_i\|_\infty \le C_\psi$ for all $i \in \mathbb{N}$.*

Define $\tilde{\kappa}^2 := \sup_{\mathbf{x} \in \mathcal{X}} \tilde{K}(\mathbf{x}, \mathbf{x})$ and

$$\gamma(\lambda) := \sum_{i=1}^{\infty} \frac{\mu_i}{\lambda + \mu_i}, \tag{A.35}$$

which is often referred to as the effective dimensionality of kernel K with respect to $L_2(\mathcal{X})$ [56]. It is follows directly from Assumption 1 that $\tilde{\kappa}^2 \le C_\psi^2 \sum_{i=1}^{\infty} \nu_i \lesssim \gamma(\lambda)$.

Theorem 4. *Under Assumption 1, it holds with probability at least $1 - \delta$ with any $\delta \in (0, 1/3)$ that*

$$\|K_{\mathbf{x}_0} - \widehat{K}_{\mathbf{x}_0}\|_\infty \le \frac{C(k, \tilde{\kappa})}{1 - C(k, \tilde{\kappa})} \lambda C_\psi^2 \gamma(\lambda), \tag{A.36}$$

where

$$C(k, \tilde{\kappa}) = \tilde{\kappa}^2 \frac{\sqrt{2\log(\delta^{-1})}}{\sqrt{k}} \left(4 + \frac{4\tilde{\kappa}\sqrt{2\log(\delta^{-1})}}{3\sqrt{k}} \right).$$

In particular, taking $\delta = k^{-10}$ yields

$$C(k, \tilde{\kappa}) = \tilde{\kappa}^2 \sqrt{20} \sqrt{\frac{\log k}{k}} \left(4 + \frac{4\tilde{\kappa}\sqrt{20}}{3} \sqrt{\frac{\log k}{k}} \right);$$

and choosing λ such that $\tilde{\kappa}^2 = o(\sqrt{k/\log k})$ simplifies (A.36) *to*

$$\|K_{\mathbf{x}_0} - \widehat{K}_{\mathbf{x}_0}\|_\infty \le 2\lambda C_\psi^2 \gamma(\lambda) \lesssim \lambda \gamma(\lambda).$$

Theorem 4 follows from applying Lemma 4 and Theorems 1 in [30] to the noiseless KRR estimator $\widehat{K}_{\mathbf{x}_0}$ given in (A.33). For the sake of brevity, we omit the proof here. Notice that setting $\delta = k^{-10}$ simplifies the bound in (A.36) and facilitates further analysis. Such a choice is quite common in the nonparametric regression literature; see, e.g., [31], [33], and [54]. Theorem 4 enables us to further investigate the convergence rate of $\|K_{\mathbf{x}_0} - \widehat{K}_{\mathbf{x}_0}\|_\infty$ for commonly used kernel functions.

Polynomially decaying kernels. Kernels with polynomially decaying eigenvalues include those that underlie the Sobolev spaces with different orders of smoothness. One popular kernel function belonging to this category is the Matérn kernel when $\mathbb{P}$ is the uniform distribution on $\mathcal{X}$ [40].

Assumption 2. *The eigenvalues of K, $\{\mu_j\}_{j=1}^\infty$, satisfy $\mu_j \asymp j^{-2\nu/d}$ for some $\nu > d/2$.*

Corollary 3. *Under Assumptions 1 and 2, it holds with probability at least $1 - k^{-10}$ that*

$$\|K_{\mathbf{x}_0} - \widehat{K}_{\mathbf{x}_0}\|_\infty \lesssim \lambda^{\frac{2\nu-d}{2\nu}}, \tag{A.37}$$

where λ is chosen such that $\tilde{\kappa}^2 = o(\sqrt{k/\log k})$ holds. In particular, for any $0 < \lambda' \le \lambda$, it holds with probability at least $1 - k^{-10}$ for all $\mathbf{x}_0 \in \mathcal{X}$ that

$$K_{\mathbf{x}_0}(\mathbf{x}_0) - \widehat{K}_{\mathbf{x}_0}(\mathbf{x}_0; \lambda') \le K_{\mathbf{x}_0}(\mathbf{x}_0) - \widehat{K}_{\mathbf{x}_0}(\mathbf{x}_0; \lambda) \lesssim \lambda^{\frac{2\nu-d}{2\nu}}.$$

Proof. Under Assumption 2, by selecting $j' = \lfloor \lambda^{-d/(2\nu)} \rfloor$ where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x, we have from (A.35) that

$$\gamma(\lambda) = \sum_{j=1}^\infty \frac{1}{1 + j^{2\nu/d}\lambda} = \sum_{j=1}^{j'} \frac{1}{1 + j^{2\nu/d}\lambda} + \sum_{j=j'+1}^\infty \frac{1}{1 + j^{2\nu/d}\lambda} \lesssim \lambda^{-\frac{d}{2\nu}} + \sum_{j=j'+1}^\infty \frac{1}{1 + j^{2\nu/d}\lambda}. \tag{A.38}$$

The last term in (A.38) can be further upper bounded as follows:

$$\sum_{j=j'+1}^\infty \frac{1}{1 + j^{2\nu/d}\lambda} \le \frac{1}{\lambda} \int_{\lambda^{-d/(2\nu)}}^\infty \frac{1}{u^{2\nu/d}} du = \lambda^{-1} \lambda^{\frac{2\nu-d}{2\nu}} \left(\frac{d}{2\nu - d} \right) \asymp \lambda^{-\frac{d}{2\nu}}. \tag{A.39}$$

It follows from (A.38) and (A.39) that $\gamma(\lambda) \lesssim \lambda^{-d/(2\nu)}$. By choosing λ such that that $\tilde{\kappa}^2 \lesssim \gamma(\lambda) = o(\sqrt{k/\log k})$, we have the upper bound in (A.37) follow from Theorem 4 immediately. The last statement follows since $K_{\mathbf{x}_0}(\mathbf{x}_0) - \widehat{K}_{\mathbf{x}_0}(\mathbf{x}_0; \lambda)$ is a nonincreasing function of λ. □

With respect to Corollary 3, one viable choice of λ is $\lambda \asymp k^{-p}(\log k)^b$, where $p > 0$ and $b \in \mathbb{R}$. We can show that, if $p = v/d$ and $b > v/d$, then $\tilde{\kappa}^2 \lesssim \gamma(\lambda) \lesssim \left(k^{-p}(\log k)^b\right)^{-\frac{d}{2v}} \lesssim k^{\frac{1}{2}}(\log k)^{-\frac{d}{2v}b} = o(\sqrt{k/\log k})$. In this case, it follows that $\|K_{\mathbf{x}_0} - \widehat{K}_{\mathbf{x}_0}\|_\infty \lesssim k^{-\frac{2v-d}{2d}}(\log k)^{\frac{2v-d}{2v}b}$. Hence, with a high probability, for any $\mathbf{x}_0 \in \mathcal{X}$, $\sigma_{k,B}(\mathbf{x}_0) \lesssim k^{-\frac{2v-d}{4d}}(\log k)^{\frac{2v-d}{4v}b}$. Recall the definition $\lambda = V_{k,\max}/(\xi^2 k)$; it is not difficult to allocate the simulation budget to fulfill this choice of λ. In case the budget allocation adopted makes λ decay to zero at a faster rate than the aforementioned case, Corollary 3 makes it clear that the corresponding upper bound on $\|K_{\mathbf{x}_0} - \widehat{K}_{\mathbf{x}_0}\|_\infty$ and hence the upper bound on the order of $\sigma_{k,B}(\cdot)$ still hold; however, it can be loose and does not imply the optimal convergence rate of $\sigma_{k,B}(\cdot)$.

We can further extend Theorem 2 and study the order of the uniform error bound in the random design setting as $k, B \to \infty$, when polynomially decaying kernels are adopted. We see from (8) that both $\gamma_{k,B}(\tau)$ and $\sqrt{\beta_k(\tau)}\sup_{\mathbf{x} \in \mathcal{X}}\sigma_{k,B}(\mathbf{x})$ must converge to zero as $k, B \to \infty$ to guarantee a vanishing approximation error at $\forall \mathbf{x} \in \mathcal{X}$. Given the aforemeioned choice of $\lambda \asymp k^{-p}(\log k)^b$ with $p = v/d$ and $b > v/d$, we have that $\lim_{k,B \to \infty} \gamma_{k,B}(\tau) = 0$ can be achieved if the grid constant as a function of k, $\tau(k)$, decreases sufficiently fast. Recall from (12) that $\gamma_{k,B}(\tau) \lesssim \left(\left(\beta_k(\tau) \cdot \tau(k) \cdot k V_{k,\min}^{-1}\right)^{\frac{1}{2}} + \tau(k) + \tau(k) \cdot k V_{k,\min}^{-\frac{1}{2}}\right)$. Assuming that $V_{k,\max}/V_{k,\min} < \infty$, we have $V_{k,\min} \asymp V_{k,\max} \asymp k^{1-p}(\log k)^b$. Setting $\tau(k) = O(k^{-a})$ as $k, B \to \infty$ with $a = 2v/d$ ensures that $\lim_{k,B \to \infty} \gamma_{k,B}(\tau) = 0$ and moreover $\gamma_{k,B}(\tau) = o\left(\sqrt{\beta_k(\tau)}\sup_{\mathbf{x} \in \mathcal{X}}\sigma_{k,B}(\mathbf{x})\right)$ as $k, B \to \infty$. The latter follows from the fact that $\beta_k(\tau) = O(\log k)$ due to (9) and $\sigma_{k,B}(\mathbf{x}) = O\left(k^{-\frac{2v-d}{4d}}(\log k)^{\frac{2v-d}{4v}b}\right)$ at $\forall \mathbf{x} \in \mathcal{X}$ thanks to Corollary 3. Therefore, $\gamma_{k,B}(\tau) + \sqrt{\beta_k(\tau)}\sup_{\mathbf{x} \in \mathcal{X}}\sigma_{k,B}(\mathbf{x}) \lesssim \sqrt{\beta_k(\tau)}\sup_{\mathbf{x} \in \mathcal{X}}\sigma_{k,B}(\mathbf{x}) = O\left(k^{-\frac{2v-d}{4d}}(\log k)^{\frac{2v-d}{4v}b+\frac{1}{2}}\right)$ as $k, B \to \infty$.

Exponentially decaying kernels. Consider kernels having exponentially decaying eigenvalues. Roughly speaking, infinitely smooth stationary kernels belong to this category. One important kernel function that falls into this category is the squared exponential or Gaussian kernel when $\mathbb{P}$ is the uniform distribution or normal on $\mathcal{X}$ [39, 40]: $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\|\mathbf{x} - \mathbf{x}'\|^2/\theta\right)$, where $\theta > 0$ denotes the lengthscale parameter.

Assumption 3. *The eigenvalues of K, $\{\mu_j\}_{j=1}^\infty$, satisfy $\mu_j \asymp b^{j^{\frac{1}{d}}}$ for some $b \in (0, 1)$.*

Corollary 4. *Under Assumptions 1 and 3, it holds with probability at least $1 - k^{-10}$ that*

$$\|K_{\mathbf{x}_0} - \widehat{K}_{\mathbf{x}_0}\|_\infty \lesssim \lambda \left(\log \lambda^{-1}\right)^d, \tag{A.40}$$

where λ is chosen such that $\tilde{\kappa}^2 = o(\sqrt{k/\log k})$ holds. In particular, for any $0 < \lambda' \leq \lambda$, it holds with probability at least $1 - k^{-10}$ for all $\mathbf{x}_0 \in \mathcal{X}$ that

$$K_{\mathbf{x}_0}(\mathbf{x}_0) - \widehat{K}_{\mathbf{x}_0}(\mathbf{x}_0; \lambda') \leq K_{\mathbf{x}_0}(\mathbf{x}_0) - \widehat{K}_{\mathbf{x}_0}(\mathbf{x}_0; \lambda) \lesssim \lambda \left(\log \lambda^{-1}\right)^d.$$

Proof. Under Assumption 3, by setting $j^* = \left\lfloor \left(\frac{\log \lambda^{-1}}{\log b^{-1}}\right)^d \right\rfloor$, we have from (A.35) that

$$
\begin{aligned}
\gamma(\lambda) &= \sum_{j=1}^{j^*} \frac{1}{1 + \lambda \mu_j^{-1}} + \sum_{j=j^*+1}^{\infty} \frac{1}{1 + \lambda \mu_j^{-1}} \\
&\lesssim \left(\frac{\log \lambda^{-1}}{\log b^{-1}}\right)^d + \lambda^{-1} \sum_{j=j^*+1}^{\infty} b^{j^{\frac{1}{d}}} \\
&\lesssim \left(\log \lambda^{-1}\right)^d + \lambda^{-1} \int_{\left(\frac{\log \lambda^{-1}}{\log b^{-1}}\right)^d}^{\infty} b^{x^{\frac{1}{d}}} \, \mathrm{d}x.
\end{aligned}
\tag{A.41}
$$

Let $t = \zeta x^{1/d}$, where $\zeta = -\log b$. Then $b^{x^{1/d}} = e^{-t}$, $\mathrm{d}x = (d\zeta^{-d})t^{d-1}\mathrm{d}t$, and the range of t is $[\log \lambda^{-1}, \infty)$. Hence, the second term on the right-hand side of (A.41) can be upper bounded as follows:

$$
\begin{aligned}
\lambda^{-1} \int_{\left(\frac{\log \lambda^{-1}}{\log b^{-1}}\right)^d}^{\infty} b^{x^{\frac{1}{d}}} \, \mathrm{d}x &= \lambda^{-1} \int_{\log \lambda^{-1}}^{\infty} e^{-t}(d\zeta^{-d})t^{d-1}\mathrm{d}t \\
&= \lambda^{-1} d\zeta^{-d} \int_{\log \lambda^{-1}}^{\infty} e^{-t} t^{d-1}\mathrm{d}t \\
&= \lambda^{-1} d\zeta^{-d} \Gamma\left(d, \log \lambda^{-1}\right),
\end{aligned}
\tag{A.42}
$$

where $\Gamma(s, u) = \int_u^{\infty} t^{s-1}e^{-t}\mathrm{d}t$ denotes the upper incomplete gamma function; in our setting, $s = d$ and $u = \log \lambda^{-1}$. It is also known that $\Gamma(s, u) \to u^{s-1}e^{-u}$ as $u \to \infty$. Therefore, (A.42) becomes

$$
\lambda^{-1} \int_{\left(\frac{\log \lambda^{-1}}{\log b^{-1}}\right)^d}^{\infty} b^{x^{\frac{1}{d}}} \, \mathrm{d}x \asymp \lambda^{-1} d\zeta^{-d} (\log \lambda^{-1})^{d-1}\lambda \asymp (\log \lambda^{-1})^{d-1}.
\tag{A.43}
$$

Hence, $\gamma(\lambda) \lesssim (\log \lambda^{-1})^d$ follows from (A.41) and (A.43). By choosing λ such that that $\tilde{\kappa}^2 \lesssim \gamma(\lambda) = o(\sqrt{k/\log k})$, we have the upper bound in (A.40) follow from Theorem 4 immediately. The last statement follows since $K_{\mathbf{x}_0}(\mathbf{x}_0) - \widehat{K}_{\mathbf{x}_0}(\mathbf{x}_0; \lambda)$ is a nonincreasing function of λ. □

Some comments follow from Corollary 4. One viable choice of λ is $\lambda \asymp \exp(-k^{\frac{1}{p}})$ for some $p > 0$. By choosing $p > 2d$, it follows that $\tilde{\kappa}^2 \lesssim \gamma(\lambda) \lesssim k^{\frac{d}{p}} = o(\sqrt{k/\log k})$. The high probability upper bound given by (A.40) in this case follows as $\|K_{\mathbf{x}_0} - \widehat{K}_{\mathbf{x}_0}\|_\infty \lesssim \exp(-k^{\frac{1}{p}})k^{\frac{d}{p}}$. Therefore, it follows that with a high probability, for any $\mathbf{x}_0 \in \mathcal{X}$, $\sigma_{k,B}(\mathbf{x}_0) \lesssim \exp(-\frac{1}{2}k^{\frac{1}{p}})k^{\frac{d}{2p}}$. Recall that $\lambda := \mathsf{V}_{k,\max}/(\xi^2 k)$; we see that it is possible to allocate the computational budget to fulfill this choice of λ. In case the budget allocation adopted makes λ diminish at a faster rate than the case discussed above, the corresponding upper bound on $\|K_{\mathbf{x}_0} - \widehat{K}_{\mathbf{x}_0}\|_\infty$ and hence the upper bound on the order of $\sigma_{k,B}(\cdot)$ still hold; however, in this case the upper bound is not as tight.

Finally, we further extend Theorem 2 by discussing the order of the uniform error bound in the random design setting as $k, B \to \infty$, when exponentially decaying kernels are adopted. We see from (8) that both $\gamma_{k,B}(\tau)$ and $\sqrt{\beta_k(\tau)} \sup_{\mathbf{x} \in \mathcal{X}} \sigma_{k,B}(\mathbf{x})$ must converge to zero as $k, B \to \infty$ to guarantee a vanishing approximation error at $\forall \mathbf{x} \in \mathcal{X}$. Given the aforemeioned choice of $\lambda \asymp \exp(-k^{\frac{1}{p}})$ with $p > 2d$, we have that $\lim_{k,B\to\infty} \gamma_{k,B}(\tau) = 0$ can be achieved if the grid constant as a function of k, $\tau(k)$, decreases sufficiently fast. Recall from (12) that $\gamma_{k,B}(\tau) \lesssim \left(\left(\beta_k(\tau) \cdot \tau(k) \cdot k\mathsf{V}_{k,\min}^{-1}\right)^{\frac{1}{2}} + \tau(k) + \tau(k) \cdot k\mathsf{V}_{k,\min}^{-\frac{1}{2}}\right)$. Assuming that $\mathsf{V}_{k,\max}/\mathsf{V}_{k,\min} < \infty$, we have $\mathsf{V}_{k,\min} \asymp$

$V_{k,\max} \asymp k \exp(-k^{\frac{1}{p}})$. Setting $\tau(k) = O\left(\exp\left(-ak^{\frac{1}{p}}\right)\right)$ where $a \geq 2$ ensures that $\lim_{k,B\to\infty} \gamma_{k,B}(\tau) = 0$ and moreover $\gamma_{k,B}(\tau) = o\left(\sqrt{\beta_k(\tau)} \sup_{\mathbf{x}\in\mathcal{X}} \sigma_{k,B}(\mathbf{x})\right)$ as $k, B \to \infty$. The latter follows from the fact that $\beta_k(\tau) = O(k^{\frac{1}{p}})$ due to (9) and $\sigma_{k,B}(\mathbf{x}) = O\left(\exp(-\frac{1}{2}k^{\frac{1}{p}})k^{\frac{d}{2p}}\right)$ at $\forall \mathbf{x} \in \mathcal{X}$ thanks to Corollary 4. Therefore, $\gamma_{k,B}(\tau) + \sqrt{\beta_k(\tau)} \sup_{\mathbf{x}\in\mathcal{X}} \sigma_{k,B}(\mathbf{x}) \lesssim \sqrt{\beta_k(\tau)} \sup_{\mathbf{x}\in\mathcal{X}} \sigma_{k,B}(\mathbf{x}) = O\left(\exp(-\frac{1}{2}k^{\frac{1}{p}})k^{\frac{1+d}{2p}}\right)$ as $k, B \to \infty$.

D   Additional Figures for Section 6

D.1   Figures for the M/M/1 Queueing Example

(a) Equal allocation, Gaussian

(b) Equal allocation, Matérn

(c) Unequal allocation 1, Gaussian

(d) Unequal allocation 1, Matérn

(e) Unequal allocation 2, Gaussian

(f) Unequal allocation 2, Matérn
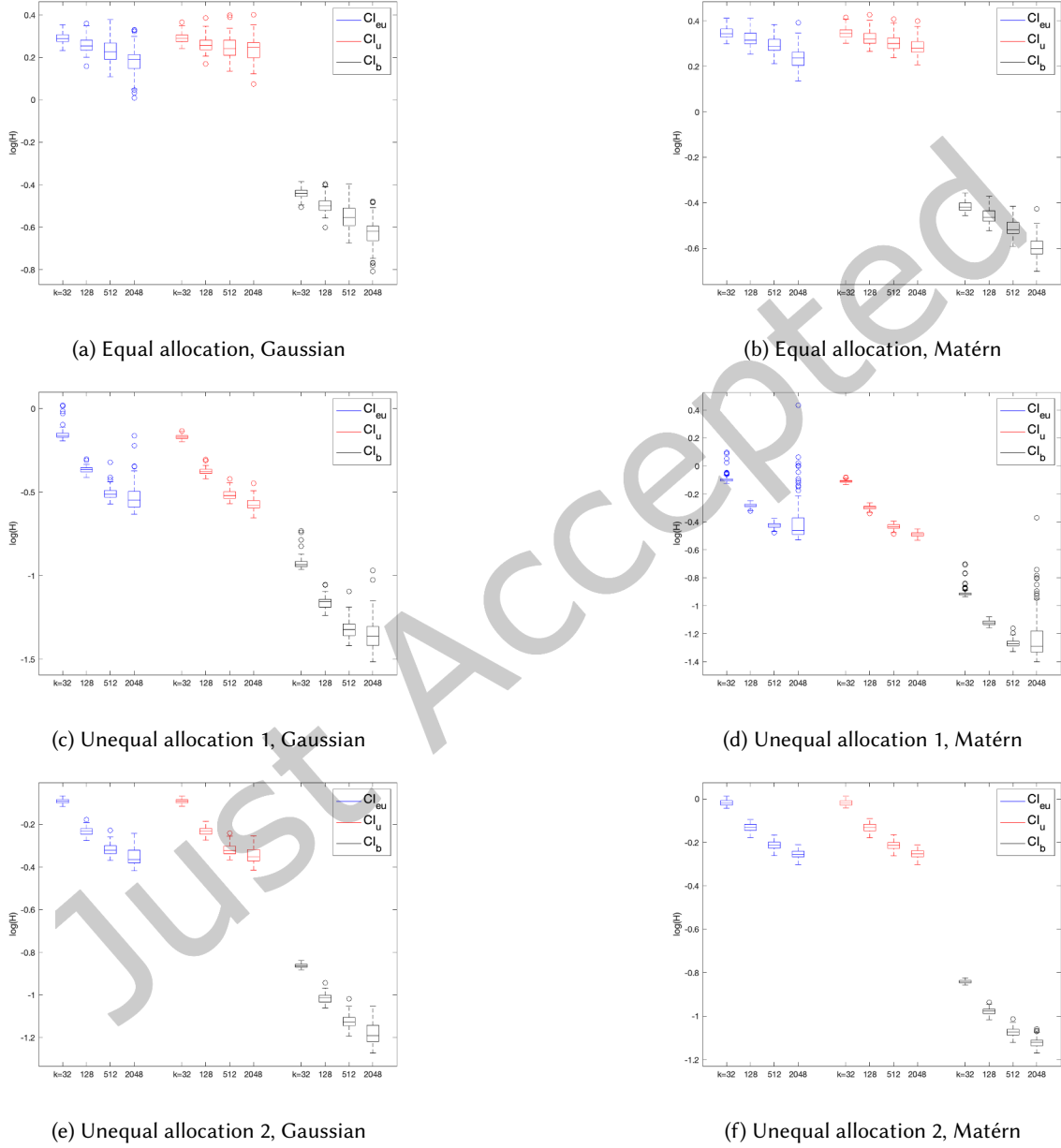
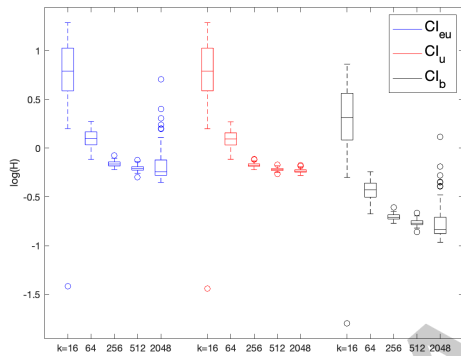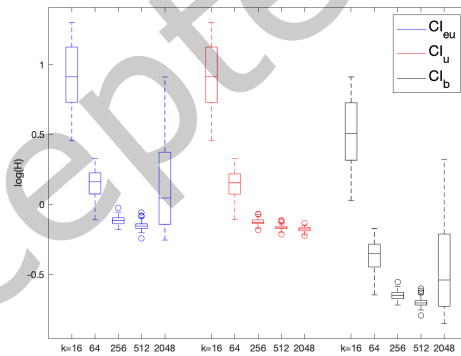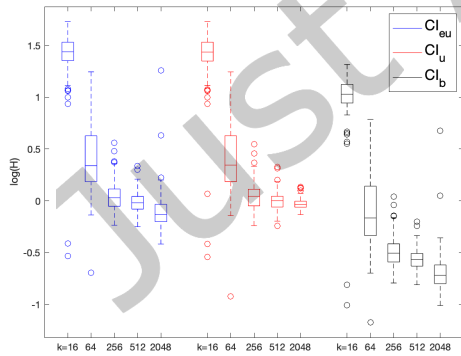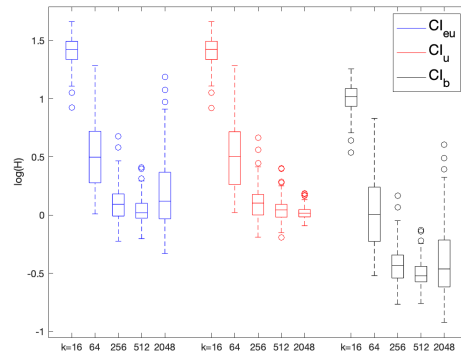Fig. 6.   Boxplots of $\log(H_{\max})$ of CI_{eu}, CI_u, and CI_b for the M/M/1 example obtained on 100 independent macro-replications under the grid design with $B = 25600$.
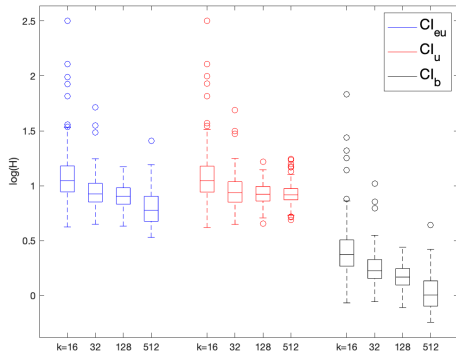
(a) Equal allocation, Gaussian
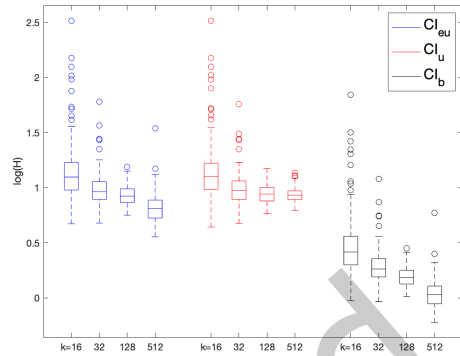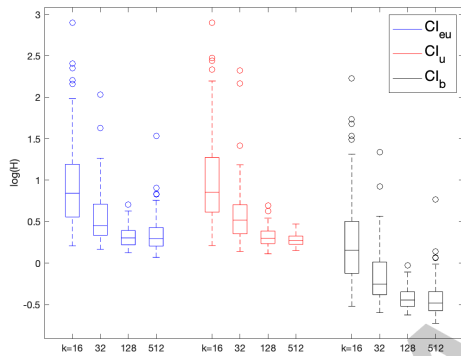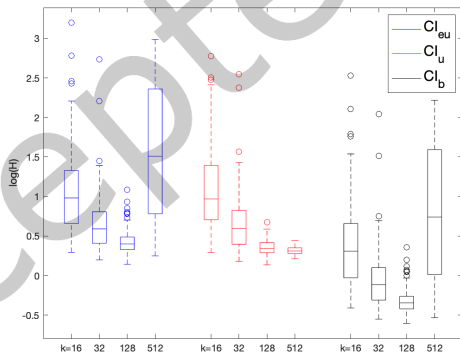
(b) Equal allocation, Matérn

(c) Unequal allocation 1, Gaussian

(d) Unequal allocation 1, Matérn
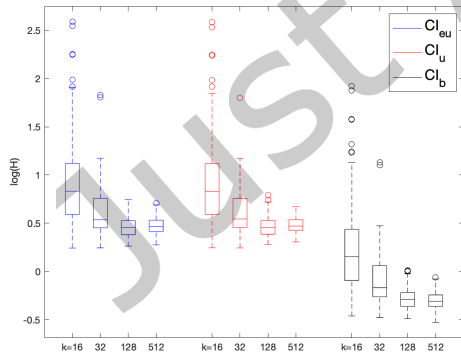
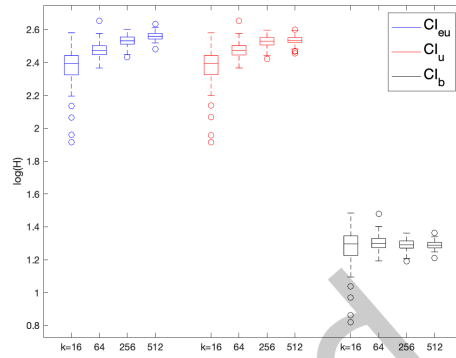(e) Unequal allocation 2, Gaussian
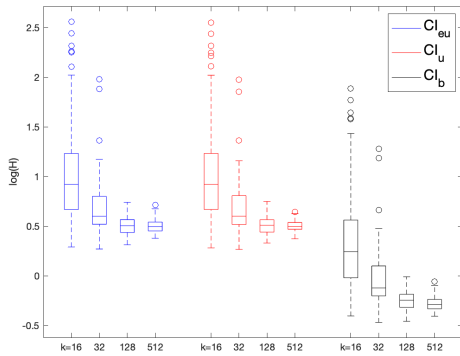
(f) Unequal allocation 2, Matérn

Fig. 7. Boxplots of $\log(H_{\max})$ of CI_{eu}, CI_u, and CI_b for the M/M/1 example obtained on 100 independent macro-replications under the uniform design with $B = 2560$.
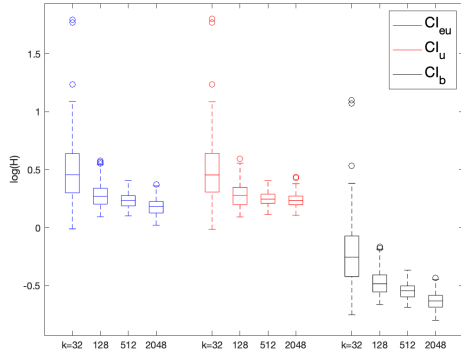
(a) Equal allocation, Gaussian
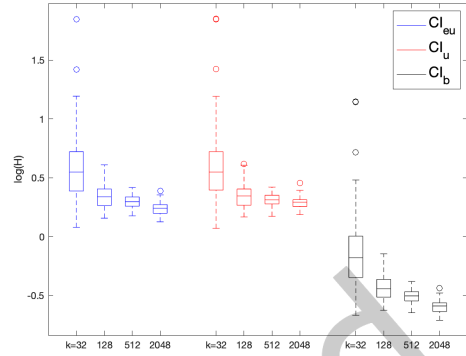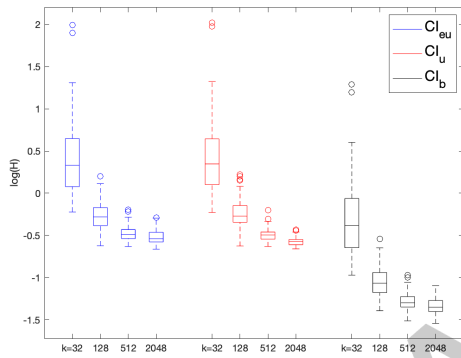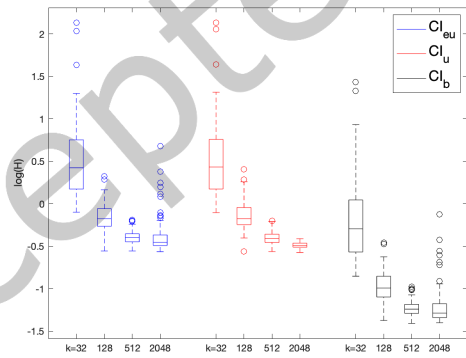
(b) Equal allocation, Matérn

(c) Unequal allocation 1, Gaussian

(d) Unequal allocation 1, Matérn
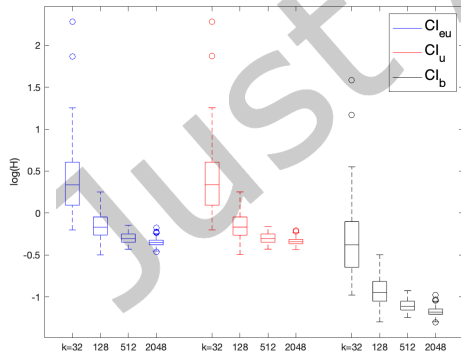
(e) Unequal allocation 2, Gaussian
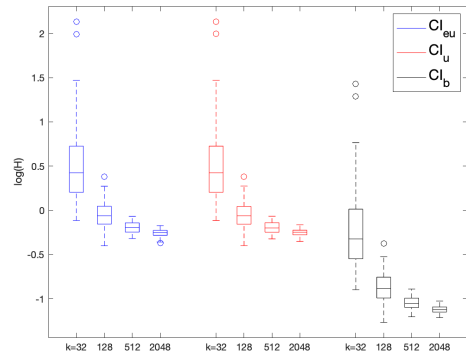
(f) Unequal allocation 2, Matérn

Fig. 8. Boxplots of $\log(H_{\max})$ of $\mathrm{CI_{eu}}$, $\mathrm{CI_u}$, and $\mathrm{CI_b}$ for the M/M/1 example obtained on 100 independent macro-replications under the uniform design with $B = 25600$.

D.2 Figures for Dimension-flexible Example

D.2.1 Additional Figures for the 1D Case. Figures 9 and 10 summarize the magnitudes of $H_{\max}$ (in a logarithmic scale) of CI_{eu}, CI_u, and CI_b obtained with a total budget of $B = 2560$ and 10240, respectively. These figures illustrate that, regardless of the kernel function adopted, different budget allocation schemes and design-point sampling schemes (i.e., LHD and uniform) tend to yield similar patterns in $H_{\max}$ magnitudes.
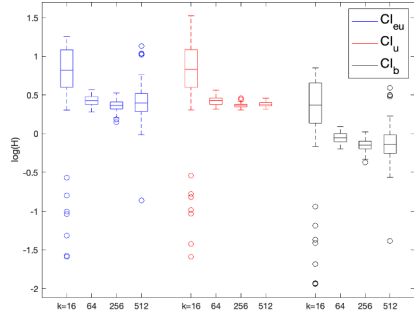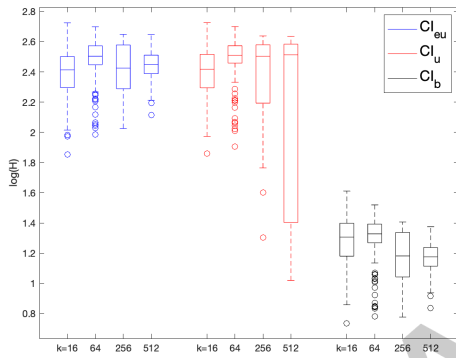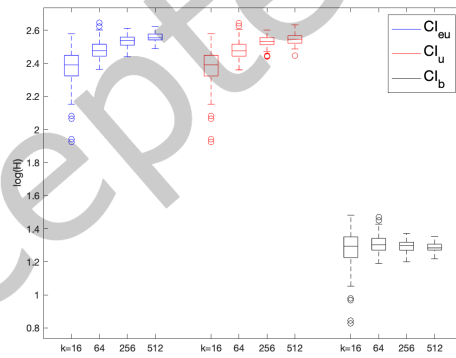
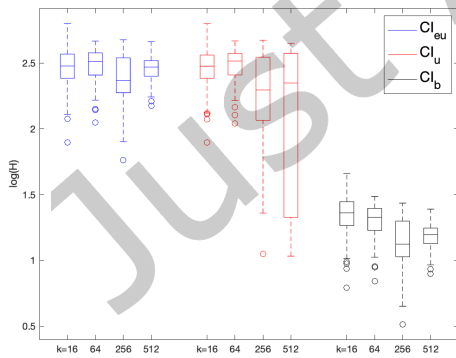(a) LHD + equal allocation, Gaussian
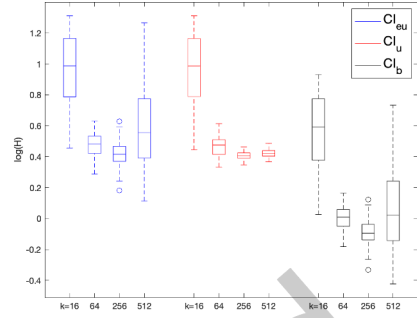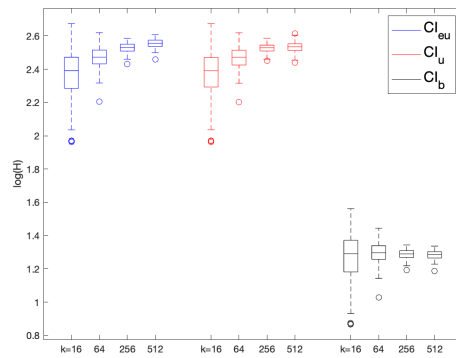
(b) LHD + equal allocation, Matérn
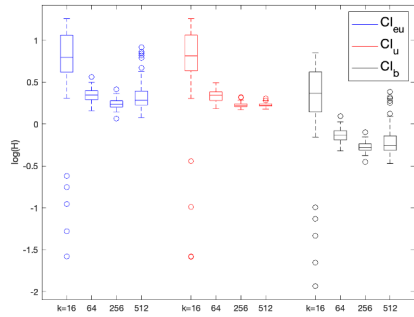
(c) LHD + unequal allocation 1, Gaussian
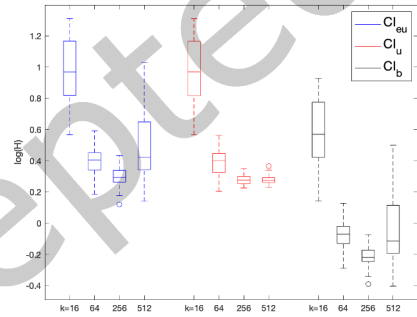
(d) LHD + unequal allocation 1, Matérn
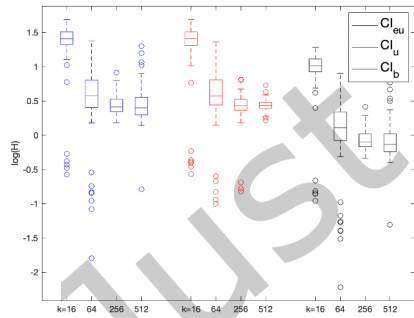
(e) Uniform + equal allocation, Gaussian
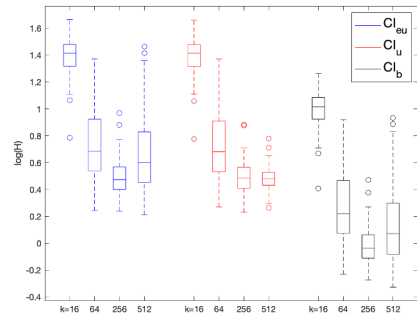
(f) Uniform + equal allocation, Matérn

Fig. 9. Boxplots of $\log(H_{\max})$ of CI_{eu}, CI_u, and CI_b for the 1D example obtained on 100 independent macro-replications with $B = 2560$.
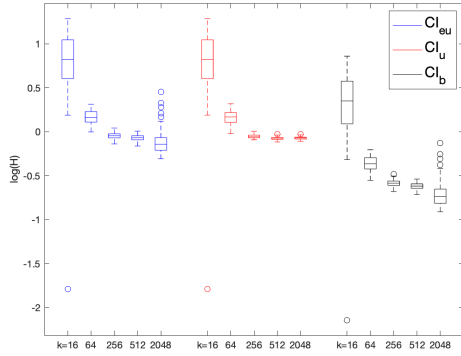
(a) LHD + equal allocation, Gaussian
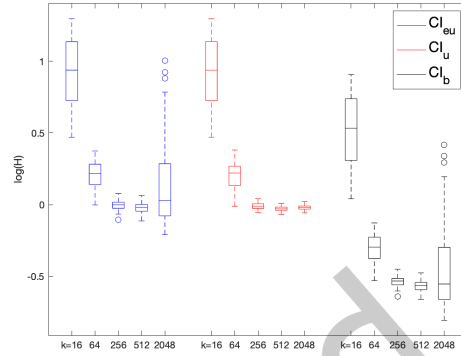
(b) LHD + equal allocation , Matérn

(c) LHD + unequal allocation 1, Gaussian
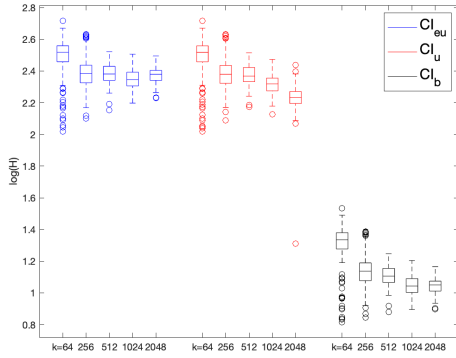
(d) LHD + unequal allocation 1, Matérn
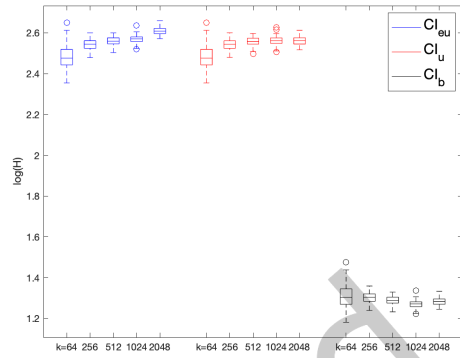
(e) Uniform + equal allocation, Gaussian

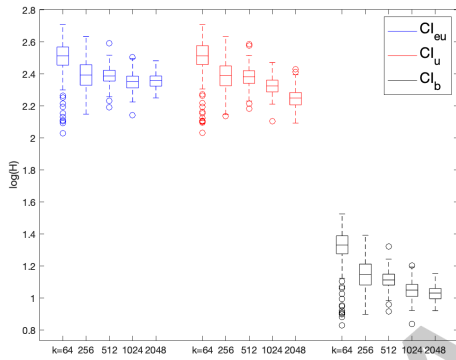(f) Uniform + equal allocation, Matérn

Fig. 10. Boxplots of $\log(H_{\max})$ of CI_{eu}, CI_u, and CI_b for the 1D example obtained on 100 independent macro-replications with $B = 10240$.

D.2.2 Additional Figures for the 5D Case. Figures 11 and 12 summarize the magnitudes of $H_{\max}$ (in a logarithmic scale) of CI_{eu}, CI_u, and CI_b obtained with a total budget of $B = 2560$ and 10240, respectively. These figures illustrate that, regardless of the kernel adopted, different budget allocation schemes and design-point sampling schemes (i.e., LHD and uniform) tend to yield similar patterns in $H_{\max}$ magnitudes.
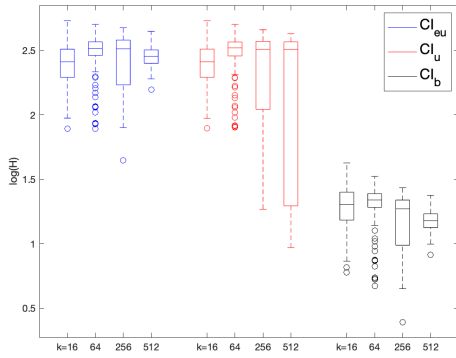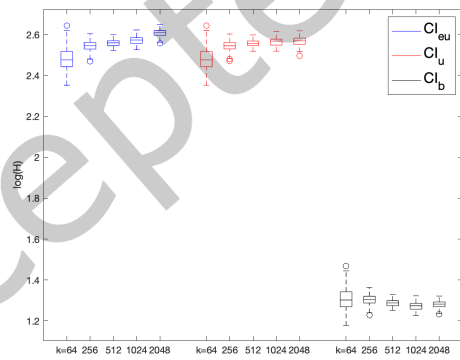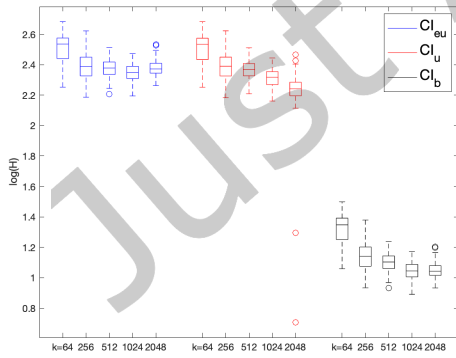
(a) LHD + equal allocation, Gaussian

(b) LHD + equal allocation, Matérn

(c) LHD + unequal allocation 1, Gaussian

(d) LHD + unequal allocation 1, Matérn

(e) Uniform + equal allocation, Gaussian

(f) Uniform + equal allocation, Matérn

Fig. 11. Boxplots of $\log(H_{\max})$ of CI_{eu}, CI_{u}, and CI_{b} for the 5D example obtained on 100 independent macro-replications with $B = 2560$.

(a) LHD + equal allocation, Gaussian
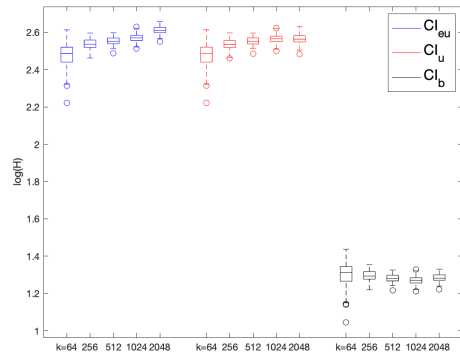
(b) LHD + equal allocation, Matérn

(c) LHD + unequal allocation 1, Gaussian

(d) LHD + unequal allocation 1, Matérn

(e) Uniform + equal allocation, Gaussian

(f) Uniform + equal allocation, Matérn

Fig. 12. Boxplots of $\log(H_{\max})$ of CI_{eu}, CI_u, and CI_b for the 5D example obtained on 100 independent macro-replications with $B = 10240$.