Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, eds.

INFORMATION CONSISTENCY OF STOCHASTIC KRIGING AND ITS IMPLICATIONS

Yutong Zhang
Xi Chen

Grado Department of Industrial and Systems Engineering
Virginia Tech
1145 Perry Street
Blacksburg, VA 24061, USA

ABSTRACT

In this paper, we introduce the concept of information consistency for Bayesian Gaussian process models and further provide the information consistency results for stochastic kriging (SK). It is found that, to ensure information consistency of SK, the budget allocated should grow in a fashion that is commensurate with the smoothness level of the mean response function to estimate, as the number of design points approaches infinity. Moreover, it is recommended that an experiment design consist of a relatively large number of design points with a few replications at each when given a fixed budget to expend.

1 INTRODUCTION

Complex simulation models of proposed or existing systems are often used to aid system design and analysis. In some situations, however, simulation models can be very expensive to run; in this case, if intense simulation is necessary to evaluate even one scenario and there are many "what if" scenarios to evaluate, simulation cannot deliver the desired answer in a timely manner. To remedy the situation, a metamodel is often used as an accurate, drop-in replacement for the simulation model as if the simulation can be run "on demand" to support real-time decision making. A metamodel is typically built on outputs from simulations run at some selected design points to "map" the performance response surface as a function of the controllable decision variables, or uncontrollable environmental variables. Successful applications of simulation metamodeling have been recorded in many cases (Osorio and Bierlaire 2013; Ouyang et al. 2017; Santos and Santos 2016).

The Gaussian process regression (GPR) or kriging methodology has been one of the most popular metamodeling approaches in various engineering disciplines for approximating the output of deterministic computer experiments (i.e., the same output is produced if the simulation is run twice at the same design point); see, for instance, Santner et al. (2003). One primary reason for GPR models' popularity is that they unite sophisticated and consistent theoretical investigations with computational tractability (Rasmussen and Williams 2006).

Asymptotic properties of GPR models have been investigated from different perspectives over the past few decades. On the one hand, posterior consistency of GPR models has been well studied in various general settings, providing a frequentist's validation of these methods as some kind of updating approaches. That is, as the sample size increases, one expects the posterior distribution of the parameter estimates to concentrate around the true values of the parameters; see, e.g., a comprehensive review of posterior consistency given by Choi and Ramamoorthi (2008).

On the other hand, quite a few research studies have investigated consistency of GPR models from the perspective of prediction. For instance, Vazquez and Bect (2010) investigated the pointwise consistency of the kriging predictor with known mean and covariance functions; for GPR models with a constant noise variance (also known as kriging with nugget effect), Gratiet and Garnier (2015) analyzed the asymptotic

values of the resulting integrated mean squared error (IMSE) and obtained the convergence rates of IMSE for selected kernels as the number of design points tends to infinity. Seeger et al. (2008) proved information consistency of GPR models based on cumulative log loss. Roughly speaking, if the sample size of the observations collected is sufficiently large and the underlying covariance function satisfies some regularity conditions, the prediction based on a GPR model is consistent to the true curve of interest, and the consistency does not depend on the choice of the values of hyper-parameters involved in the covariance function. Moreover, information consistency rates for a wide range of covariance functions were obtained using kernel eigenvalues asymptotics. The interested reader is referred to Shi and Choi (2011) and references therein for details.

Standard GPR models (i.e., kriging and kriging with nugget effect), however, often fall short for the purpose of providing an accurate approximation to the mean response function implied by a stochastic simulation experiment, especially when the simulation model generates random outputs that exhibit strong heteroscedasticity, namely, the simulation variance varies significantly across the design space (Ankenman et al. 2010). The literature on heteroscedastic GPR models for stochastic simulation metamodeling is not sparse (Ankenman et al. 2010; Ng and Yin 2012; van Beers and Kleijnen 2008). The *stochastic kriging* (SK) methodology proposed by Ankenman et al. (2010) has been known as an effective metamodeling tool for approximating a mean response function by correctly accounting for both sampling uncertainty inherent in a stochastic simulation and the response-surface uncertainty. Despite a great amount of research effort dedicated to experiment design and analysis methods for SK metamodeling (Chen and Kim 2014; Chen and Kim 2016; Chen et al. 2012; Chen et al. 2013; Liu and Staum 2010; Wang and Chen 2018), little work has been done to investigate consistency issues of SK and heteroscedastic GPR models alike.

In this paper, we study information consistency of SK with known heteroscedasticity along the lines of Seeger et al. (2008) and Wang and Shi (2014). The results help shed some light on properties of a desirable stochastic simulation experiment design for SK to achieve such a consistency. It is found that the budget used by SK metamodeling should grow in a fashion that is commensurate with the smoothness level of the mean response function of interest, as the number of design points approaches infinity. The remainder of the paper is organized as follows. Section 2 provides a brief review on SK. Section 3 provides the main results. Finally, Section 4 concludes the paper.

2 A BRIEF REVIEW ON STOCHASTIC KRIGING

Proposed by Ankenman et al. (2010), the SK methodology has been known as an effective metamodeling tool for approximating a *mean* response function implied by a stochastic simulation. We provide a brief review on SK in this section before delving into the information consistency issue of interest.

A simulation experiment design for applying SK to approximate a mean response function typically comprises a set of design points (say, k) to run independent simulations and the corresponding numbers of replications to apply at each, i.e., $\{(\mathbf{x}_i, n_i), i = 1,2,\ldots,k\}$. The simulation output obtained at a design point $\mathbf{x}_i \in \mathscr{X} \subset \Re^d$ on the jth simulation replication can be described by the following model:

$$\mathscr{Y}_j(\mathbf{x}_i) = f_0(\mathbf{x}_i) + \varepsilon_j(\mathbf{x}_i), \quad j = 1,2,\ldots,n_i, \tag{1}$$

where $f_0(\cdot)$ is the true unknown mean response function that we intend to estimate, and the simulation errors incurred at $\mathbf{x}_i$ on different simulation replications, $\varepsilon_1(\mathbf{x}_i), \varepsilon_2(\mathbf{x}_i), \ldots$, are assumed to be independent and identically distributed (i.i.d.) normal random variables, $\varepsilon_j(\mathbf{x}_i) \sim \mathscr{N}(\mathbf{0}, \mathsf{V}(\mathbf{x}_i))$, $i = 1,2,\ldots,k$. The simulation error variance $\mathsf{V}(\mathbf{x})$ may depend on $\mathbf{x}$ and $\sup_{\mathbf{x}\in\mathscr{X}} \mathsf{V}(\mathbf{x}) < \infty$. The normality of $\varepsilon_j(\mathbf{x})$ can be anticipated from the fact that in a discrete-event simulation the simulation output $\mathscr{Y}_j(\mathbf{x})$ could be the average of a large number of more basic random variables obtained on the jth simulation replication.

Given the simulation outputs generated, we see from (1) that the average output at $\mathbf{x}_i$ can be written as

$$\bar{\mathscr{Y}}(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathscr{Y}_j(\mathbf{x}_i) = f_0(\mathbf{x}_i) + \bar{\varepsilon}(\mathbf{x}_i), \quad i = 1,2,\ldots,k, \tag{2}$$

where $\bar{\varepsilon}(\mathbf{x}_i) = n_i^{-1}\sum_{j=1}^{n_i}\varepsilon_j(\mathbf{x}_i)$ denotes the average simulation error incurred at $\mathbf{x}_i$. Write $\bar{\boldsymbol{\varepsilon}}$ as a shorthand for the $k\times 1$ vector of average simulation errors $(\bar{\varepsilon}(\mathbf{x}_1),\bar{\varepsilon}(\mathbf{x}_2),\ldots,\bar{\varepsilon}(\mathbf{x}_k))^\top$, and write the $k\times 1$ vector of average outputs as $\bar{\mathscr{Y}} = \left(\bar{\mathscr{Y}}(\mathbf{x}_1),\bar{\mathscr{Y}}(\mathbf{x}_2),\ldots,\bar{\mathscr{Y}}(\mathbf{x}_k)\right)^\top$.

Parallel with the treatment adopted in the standard GPR literature (Rasmussen and Williams 2006; Santner et al. 2003), in SK it is assumed that the underlying mean response function $f_0(\cdot)$ is a zero-mean Gaussian process, denoted by $f_0(\cdot)\sim \mathrm{GP}(\mathbf{0}, K(\cdot,\cdot;\boldsymbol{\theta}))$, where $K(\cdot,\cdot;\boldsymbol{\theta})$ denotes the covariance function or kernel. Specifically, the covariance between the values of f_0 at any two inputs $\mathbf{x},\mathbf{x}'\in\mathscr{X}$ can be modeled as

$$\mathrm{Cov}\left(f_0(\mathbf{x}), f_0(\mathbf{x}')\right) = K(\mathbf{x},\mathbf{x}';\boldsymbol{\theta}), \tag{3}$$

where $\boldsymbol{\theta}$ denotes the vector of hyper-parameters in the covariance function. Commonly used covariance functions include Matérn and squared exponential or Gaussian covariance functions; see Chapter 4 of Rasmussen and Williams (2006) for details. We note that Ankenman et al. (2010) refer to the stipulated stochastic nature of $f_0(\cdot)$ as extrinsic uncertainty, in contrast to the intrinsic uncertainty induced by the simulation error ε that is inherent in a stochastic simulation output; and these two sources of uncertainty are assumed to be independent.

To predict the mean response at any $\mathbf{x}_0\in\mathscr{X}$, SK adopts the following predictor:

$$\mu(\mathbf{x}_0) = \boldsymbol{\Sigma}_f(\mathbf{x}_0,\mathbf{X})^\top\left(\boldsymbol{\Sigma}_{f,k}+\boldsymbol{\Sigma}_{\varepsilon,k}\right)^{-1}\bar{\mathscr{Y}}, \tag{4}$$

and the corresponding predictive variance is given by

$$\sigma^2(\mathbf{x}_0) = \boldsymbol{\Sigma}_f(\mathbf{x}_0,\mathbf{x}_0) - \boldsymbol{\Sigma}_f(\mathbf{x}_0,\mathbf{X})^\top(\boldsymbol{\Sigma}_{f,k}+\boldsymbol{\Sigma}_{\varepsilon,k})^{-1}\boldsymbol{\Sigma}_f(\mathbf{x}_0,\mathbf{X}), \tag{5}$$

where $\mathbf{X} := (\mathbf{x}_1^\top,\mathbf{x}_2^\top,\ldots,\mathbf{x}_k^\top)^\top$ denotes the matrix consisting the k design points. With a slight abuse of notation, we denote $\boldsymbol{\Sigma}_{f,k} := K(\mathbf{X},\mathbf{X};\boldsymbol{\theta})$ as the $k\times k$ covariance matrix across the k design points, with its (i,j)th entry given by $K(\mathbf{x}_i,\mathbf{x}_j;\boldsymbol{\theta})$ for $i,j=1,2,\ldots,k$. The $k\times 1$ vector, $\boldsymbol{\Sigma}_f(\mathbf{x}_0,\mathbf{X})$, contains the covariances between the k design points and the prediction point $\mathbf{x}_0$. The $k\times k$ matrix $\boldsymbol{\Sigma}_{\varepsilon,k}$ is the variance-covariance matrix of $\bar{\boldsymbol{\varepsilon}}$. As the use of common random numbers (CRN) does not necessarily help improve the predictive performance of SK (Chen et al. 2012), in this paper we assume that CRN is not applied in simulation experiments. Hence, $\boldsymbol{\Sigma}_{\varepsilon,k}$ reduces to a $k\times k$ diagonal matrix, i.e., $\boldsymbol{\Sigma}_{\varepsilon,k} = \mathrm{diag}\left(\mathsf{V}(\mathbf{x}_1)/n_1,\mathsf{V}(\mathbf{x}_2)/n_2,\ldots,\mathsf{V}(\mathbf{x}_k)/n_k\right)$.

3 MAIN RESULTS

To start, define $\mathbf{x}_{\le i}$ and $\bar{\mathscr{Y}}_{\le i}$ respectively as $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_i\}$ and $\{\bar{\mathscr{Y}}_1,\bar{\mathscr{Y}}_2,\ldots,\bar{\mathscr{Y}}_i\}$ for $i\ge 1$, where $\bar{\mathscr{Y}}_i$ is a shorthand for $\bar{\mathscr{Y}}(\mathbf{x}_i)$, $i=1,2,\ldots,k$. Similarly, let $\mathbf{x}_{<i}$ and $\bar{\mathscr{Y}}_{<i}$ respectively be $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{i-1}\}$ and $\{\bar{\mathscr{Y}}_1,\bar{\mathscr{Y}}_2,\ldots,\bar{\mathscr{Y}}_{i-1}\}$ for $i\ge 2$. Assume that $f(\cdot)\sim\mathrm{GP}(\mathbf{0},K(\cdot,\cdot;\boldsymbol{\theta}))$. Hence, the stochastic process $f(\cdot)$ induces a measure on the space of continuous functions $\mathscr{F}=\{f(\cdot):\mathscr{X}\mapsto\Re\}$, where recall that $\boldsymbol{\theta}$ is the hyper-parameter vector and we assume that $\boldsymbol{\theta}$ is estimated by $\widehat{\boldsymbol{\theta}}$ via some empirical Bayesian method (Shi and Choi 2011). Also recall that $f_0(\cdot)$ denotes the true underlying mean response function. Denote

$$\begin{aligned} p_{gp}(\bar{\mathscr{Y}}_{\le k}) &= \int_{\mathscr{F}} p(\bar{\mathscr{Y}}_{\le k}|f(\mathbf{x}_{\le k}))\tilde{p}(f(\mathbf{x}_{\le k}))\mathrm{d}f(\mathbf{x}_1)\ldots\mathrm{d}f(\mathbf{x}_k),\\ p_0(\bar{\mathscr{Y}}_{\le k}) &= p(\bar{\mathscr{Y}}_1,\bar{\mathscr{Y}}_2,\ldots,\bar{\mathscr{Y}}_k|f_0(\mathbf{x}_{\le k})) = p(\bar{\mathscr{Y}}_{\le k}|f_0(\mathbf{x}_{\le k})),\end{aligned}$$

where $p_{gp}(\bar{\mathscr{Y}}_{\le k})$ is the predictive distribution of $\bar{\mathscr{Y}}_{\le k}$ given by SK, and $\tilde{p}(f)$ depends on the data collected at design points in $\mathbf{x}_{\le k}$ since the hyper-parameters of $f(\cdot)$ are estimated from the data collected.

We say that SK achieves *information consistency* if $k^{-1}\mathrm{E}_{\mathbf{x}_{\le k}}(D[p_0(\bar{\mathscr{Y}}_{\le k}),p_{gp}(\bar{\mathscr{Y}}_{\le k})])\to 0$ as $k\to\infty$, where $\mathrm{E}_{\mathbf{x}_{\le k}}$ denotes the expectation under the distribution of the design points in $\mathbf{x}_{\le k}$ and $D[p_0(\bar{\mathscr{Y}}_{\le k}),p_{gp}(\bar{\mathscr{Y}}_{\le k})]$ is the Kullback-Leibler (KL) divergence between $p_0(\cdot)$ and $p_{gp}(\cdot)$, i.e., $D[p_0(u),p_{gp}(u)] = \int p_0(u)\log[p_0(u)/p_{gp}(u)]\,\mathrm{d}u$.

Lemma 1 Consider SK as a GP prediction method equipped with a zero-mean GP prior and covariance function $K(\cdot,\cdot;\boldsymbol{\theta})$. Let $\bar{\mathscr{Y}}_j(\mathbf{x}_i)$'s be simulation outputs as described by the output model (1) and the underlying mean response function f_0 be from the reproducing kernel Hilbert space (RKHS) associated with $K(\cdot,\cdot;\boldsymbol{\theta})$. Suppose that $K(\cdot,\cdot;\boldsymbol{\theta})$ is bounded and continuous in $\boldsymbol{\theta}$ and the estimator $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ almost surely as $k \to \infty$. Then for some $M > 0$ and any $\eta > 0$, when k is large enough, it holds that

$$\frac{1}{k}\left(-\log p_{gp}(\bar{\mathscr{Y}}_{\le k}) + \log p_0(\bar{\mathscr{Y}}_{\le k})\right) \le \frac{1}{k}\left(\frac{1}{2}\|f_0\|_K^2 + \frac{1}{2}\log|\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}\boldsymbol{\Sigma}_{f,k} + \mathbf{I}_k| + M\right) + \eta, \tag{6}$$

where recall that $\boldsymbol{\Sigma}_{f,k}$ is the $k \times k$ covariance matrix over the design points in $\mathbf{x}_{\le k}$ stipulated by SK and $\boldsymbol{\Sigma}_{\varepsilon,k} = \mathrm{diag}\,(\mathsf{V}(\mathbf{x}_1)/n_1, \mathsf{V}(\mathbf{x}_2)/n_2, \ldots, \mathsf{V}(\mathbf{x}_k)/n_k)$. Furthermore, $\|f_0\|_K$ is the RKHS norm of f_0 associated with $K(\cdot,\cdot;\boldsymbol{\theta})$, $|\mathbf{A}|$ denotes the determinant of matrix $\mathbf{A}$, and $\mathbf{I}_k$ denotes the $k \times k$ identity matrix.

Proof. Let $\mathscr{H}$ be the RKHS associated with $K(\cdot,\cdot;\boldsymbol{\theta})$ and $\mathscr{H}_k$ the span of $K(\cdot,\mathbf{x}_i;\boldsymbol{\theta})$, i.e., $\mathscr{H}_k = \{f(\cdot) : f(\mathbf{x}) = \sum_{i=1}^k \alpha_i K(\mathbf{x},\mathbf{x}_i;\boldsymbol{\theta}),\ \text{for any } \alpha_i \in \Re\}$. We first assume that $f_0 \in \mathscr{H}_k$, then $f_0(\cdot)$ can be expressed as $f_0(\cdot) = \sum_{i=1}^k \alpha_i K(\cdot,\mathbf{x}_i;\boldsymbol{\theta})$. Denote $\boldsymbol{\alpha} = (\alpha_1,\alpha_2,\ldots,\alpha_k)^\top$. It follows from the properties of RKHS that $\|f_0\|_K^2 = \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{f,k}\boldsymbol{\alpha}$ and $(f_0(\mathbf{x}_1), f_0(\mathbf{x}_2), \ldots, f_0(\mathbf{x}_k))^\top = \boldsymbol{\Sigma}_{f,k}\boldsymbol{\alpha}$.

Let P and $\bar{P}$ be any two measures on $\mathscr{F}$, then it follows from Fenchel-Legendre duality relationship that, for any functional $g(\cdot)$ on $\mathscr{F}$,

$$\mathrm{E}_{\bar{P}}[g(f)] \le \log \mathrm{E}_P[e^{g(f)}] + D(\bar{P},P). \tag{7}$$

In (7), we let

1. $g(f)$ be $\log p(\bar{\mathscr{Y}}_{\le k}|f)$ for any $\bar{\mathscr{Y}}_1, \bar{\mathscr{Y}}_2, \ldots, \bar{\mathscr{Y}}_k \in \Re$ and $f \in \mathscr{F}$;
2. P be the probability measure induced by $\mathrm{GP}(\mathbf{0}, K(\cdot,\cdot;\widehat{\boldsymbol{\theta}}))$, hence its finite dimensional distribution at $z_1, z_2, \ldots, z_k$ is $\tilde{p}(z_1,z_2,\ldots,z_k) = \mathscr{N}(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}_{f,k})$, and

$$\mathrm{E}_P[e^{g(f)}] = \mathrm{E}_P[p(\bar{\mathscr{Y}}_{\le k}|f)] = \int_{\mathscr{F}} p(\bar{\mathscr{Y}}_{\le k}|f)\tilde{p}(f(\mathbf{x}_{\le k}))\mathrm{d}f(\mathbf{x}_1)\ldots\mathrm{d}f(\mathbf{x}_k) = p_{gp}(\bar{\mathscr{Y}}_{\le k}); \tag{8}$$

   where $\widehat{\boldsymbol{\Sigma}}_{f,k}$ is defined in the same way as $\boldsymbol{\Sigma}_{f,k}$ but with $\boldsymbol{\theta}$ being replaced by $\widehat{\boldsymbol{\theta}}$.
3. $\bar{P}$ be the posterior distribution of $f(\cdot)$ on $\mathscr{F}$ which has a prior distribution $\mathrm{GP}(\mathbf{0}, K(\cdot,\cdot;\boldsymbol{\theta}))$ and normal likelihood $\prod_{i=1}^k \mathscr{N}(\bar{\mathscr{Y}}_i; f(\mathbf{x}_i), \mathsf{V}(\mathbf{x}_i)/n_i)$, where $\bar{\mathscr{Y}} = (\bar{\mathscr{Y}}_1, \bar{\mathscr{Y}}_2, \ldots, \bar{\mathscr{Y}}_k)^\top = (\boldsymbol{\Sigma}_{f,k} + \boldsymbol{\Sigma}_{\varepsilon,k})\boldsymbol{\alpha}$. And $\bar{\mathscr{Y}}$ is a vector of average outputs observed at $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. Hence, $\bar{P}(f) = p(f|\bar{\mathscr{Y}}_{\le k}, \mathbf{x}_{\le k})$ is a probability measure on $\mathscr{F}$. By the GPR property, the posterior of $(f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_k))^\top$ is

$$\begin{aligned}
\bar{p}\,(f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_k)) &:= p\left(f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_k)|\bar{\mathscr{Y}}_{\le k}, \mathbf{x}_{\le k}\right) \\
&= \mathscr{N}\left(\boldsymbol{\Sigma}_{f,k}(\boldsymbol{\Sigma}_{f,k} + \boldsymbol{\Sigma}_{\varepsilon,k})^{-1}\bar{\mathscr{Y}}, \boldsymbol{\Sigma}_{f,k}(\boldsymbol{\Sigma}_{f,k} + \boldsymbol{\Sigma}_{\varepsilon,k})^{-1}\boldsymbol{\Sigma}_{\varepsilon,k}\right), \\
&= \mathscr{N}(\boldsymbol{\Sigma}_{f,k}\boldsymbol{\alpha}, \boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}),
\end{aligned}$$

   where $\mathbf{B} = \boldsymbol{\Sigma}_{\varepsilon,k}^{-1}\boldsymbol{\Sigma}_{f,k} + \mathbf{I}_k$.

Now, on the one hand, it follows that

$$\begin{aligned}
D(\bar{P},P) &= \int_{\mathscr{F}} \log\left(\frac{\mathrm{d}\bar{P}}{\mathrm{d}P}\right)\mathrm{d}\bar{P} \\
&= \int_{\Re^k} \bar{p}\left(f(\mathbf{x}_1),f(\mathbf{x}_2),\ldots,f(\mathbf{x}_k)\right)\log\left[\frac{\bar{p}\left(f(\mathbf{x}_1),f(\mathbf{x}_2),\ldots,f(\mathbf{x}_k)\right)}{\tilde{p}\left(f(\mathbf{x}_1),f(\mathbf{x}_2),\ldots,f(\mathbf{x}_k)\right)}\right]\mathrm{d}f(\mathbf{x}_1)\ldots\mathrm{d}f(\mathbf{x}_k) \\
&= \frac{1}{2}\left(\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}) + (\boldsymbol{\Sigma}_{f,k}\boldsymbol{\alpha})^\top\boldsymbol{\Sigma}_{f,k}^{-1}(\boldsymbol{\Sigma}_{f,k}\boldsymbol{\alpha}) - k - \log|\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}| + \log|\widehat{\boldsymbol{\Sigma}}_{f,k}|\right) \\
&= \frac{1}{2}\left(-\log|\widehat{\boldsymbol{\Sigma}}_{f,k}\boldsymbol{\Sigma}_{f,k}| + \log|\mathbf{B}| + \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}) + (\boldsymbol{\Sigma}_{f,k}\boldsymbol{\alpha})^\top\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}(\boldsymbol{\Sigma}_{f,k}\boldsymbol{\alpha}) - k\right) \\
&= \frac{1}{2}\left(-\log|\widehat{\boldsymbol{\Sigma}}_{f,k}\boldsymbol{\Sigma}_{f,k}| + \log|\mathbf{B}| + \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}) + \|f_0\|_K^2 + \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{f,k}(\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k} - \mathbf{I}_k)\boldsymbol{\alpha} - k\right),
\end{aligned}$$

where $\mathrm{tr}(\mathbf{A})$ denotes the trace of matrix $\mathbf{A}$. On the other hand,

$$\mathrm{E}_{\bar{P}}[g(f)] = \mathrm{E}_{\bar{P}}[\log p(\bar{\mathscr{Y}}_{\le k}|f)] = \sum_{i=1}^{k}\mathrm{E}_{\bar{P}}[\log p\left(\bar{\mathscr{Y}}_i|f(\mathbf{x}_i)\right)], \tag{9}$$

since

$$\log p\left(\bar{\mathscr{Y}}_i|f(\mathbf{x}_i)\right) = -\frac{1}{2}\cdot\log\left(2\pi\frac{\mathsf{V}(\mathbf{x}_i)}{n_i}\right) - \frac{1}{2}\cdot\frac{\left(\bar{\mathscr{Y}}_i - f(\mathbf{x}_i)\right)^2}{\mathsf{V}(\mathbf{x}_i)/n_i}.$$

By Taylor's expansion, we expand $\log p\left(\bar{\mathscr{Y}}_i|f(\mathbf{x}_i)\right)$ at $f_0(\mathbf{x}_i)$,

$$\begin{aligned}
\log p\left(\bar{\mathscr{Y}}_i|f(\mathbf{x}_i)\right) &= \log p\left(\bar{\mathscr{Y}}_i|f_0(\mathbf{x}_i)\right) + \frac{\mathrm{d}\left(\log p\left(\bar{\mathscr{Y}}_i|f_0(\mathbf{x}_i)\right)\right)}{\mathrm{d}f(\mathbf{x}_i)}\left(f(\mathbf{x}_i) - f_0(\mathbf{x}_i)\right) \\
&\quad + \frac{1}{2}\frac{\mathrm{d}^2\left(\log p\left(\bar{\mathscr{Y}}_i|\tilde{f}(\mathbf{x}_i)\right)\right)}{\mathrm{d}^2 f(\mathbf{x}_i)}\left(f(\mathbf{x}_i) - f_0(\mathbf{x}_i)\right)^2,
\end{aligned}$$

where $\tilde{f}(\mathbf{x}_i) = f_0(\mathbf{x}_i) + \lambda\left(f(\mathbf{x}_i) - f_0(\mathbf{x}_i)\right)$ for some $\lambda \in (0,1)$ and

$$\frac{\mathrm{d}(\log p\left(\bar{\mathscr{Y}}_i|f(\mathbf{x}_i)\right))}{\mathrm{d}f(\mathbf{x}_i)} = \frac{\left(\bar{\mathscr{Y}}_i - f(\mathbf{x}_i)\right)}{\mathsf{V}(\mathbf{x}_i)/n_i}, \quad \frac{\mathrm{d}^2(\log p\left(\bar{\mathscr{Y}}_i|f(\mathbf{x}_i)\right))}{\mathrm{d}^2 f(\mathbf{x}_i)} = -\frac{n_i}{\mathsf{V}(\mathbf{x}_i)}.$$

Hence, we have

$$\begin{aligned}
\mathrm{E}_{\bar{P}}[\log p\left(\bar{\mathscr{Y}}_i|f(\mathbf{x}_i)\right)] &= \log p\left(\bar{\mathscr{Y}}_i|f_0(\mathbf{x}_i)\right) + \frac{\mathrm{d}\left(\log p\left(\bar{\mathscr{Y}}_i|f_0(\mathbf{x}_i)\right)\right)}{\mathrm{d}f(\mathbf{x}_i)}\mathrm{E}_{\bar{P}}\left[f(\mathbf{x}_i) - f_0(\mathbf{x}_i)\right] \\
&\quad - \frac{n_i}{2\mathsf{V}(\mathbf{x}_i)}\mathrm{E}_{\bar{P}}\left[\left(f(\mathbf{x}_i) - f_0(\mathbf{x}_i)\right)^2\right],
\end{aligned} \tag{10}$$

where $\mathrm{E}_{\bar{P}}\left[f(\mathbf{x}_i) - f_0(\mathbf{x}_i)\right] = 0$ and $\mathrm{E}_{\bar{P}}\left[\left(f(\mathbf{x}_i) - f_0(\mathbf{x}_i)\right)^2\right] = (\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1})_{ii}$, i.e., the (i,i)th entry of the matrix $\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}$, $i = 1,2,\ldots,k$.

It follows from (9) and (10) that

$$\begin{aligned}
\mathrm{E}_{\bar{P}}[g(f)] &= \mathrm{E}_{\bar{P}}[\log p(\bar{\mathscr{Y}}_{\le k}|f)] = \sum_{i=1}^{k}\mathrm{E}_{\bar{P}}[\log p\left(\bar{\mathscr{Y}}_i|f(\mathbf{x}_i)\right)] \\
&= \sum_{i=1}^{k}\log p\left(\bar{\mathscr{Y}}_i|f_0(\mathbf{x}_i)\right) - \frac{1}{2}\sum_{i=1}^{k}\frac{(\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1})_{ii}}{\mathsf{V}(\mathbf{x}_i)/n_i} = \sum_{i=1}^{k}\log p\left(\bar{\mathscr{Y}}_i|f_0(\mathbf{x}_i)\right) - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}) \\
&= \log p_0(\bar{\mathscr{Y}}_{\le k}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}).
\end{aligned} \tag{11}$$

Finally, we have from (7)–(11) that

$$\begin{aligned}
-\log p_{gp}(\bar{\mathscr{Y}}_{\le k}) + \log p_0(\bar{\mathscr{Y}}_{\le k}) &= -\log \mathrm{E}_P[e^{g(f)}] + \mathrm{E}_{\bar{P}}[g(f)] + \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}) \\
&\le D(\bar{P},P) + \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}) \\
&= \frac{1}{2}\Big(-\log|\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k}| + \log|\mathbf{B}| + \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}) + \|f_0\|_K^2 \\
&\quad + \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{f,k}(\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k} - \mathbf{I}_k)\boldsymbol{\alpha} - k + \mathrm{tr}(\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}\boldsymbol{\Sigma}_{\varepsilon,k}^{-1})\Big). \qquad (12)
\end{aligned}$$

When k is large enough, for any $\eta > 0$, we have there exists $\eta' > 0$ such that $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}) \le \mathrm{tr}\left((\mathbf{I}_k + \eta'\boldsymbol{\Sigma}_{f,k})\mathbf{B}^{-1}\right)$ satisfying $\mathrm{tr}(\eta'\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}) \le \eta \cdot k$. Hence,

$$\begin{aligned}
\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}) + \mathrm{tr}(\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}) &\le \mathrm{tr}\left((\mathbf{I}_k + \eta'\boldsymbol{\Sigma}_{f,k})\mathbf{B}^{-1} + \boldsymbol{\Sigma}_{\varepsilon,k}^{-1}\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}\right) \\
&= \mathrm{tr}\left((\mathbf{I}_k + \boldsymbol{\Sigma}_{\varepsilon,k}^{-1}\boldsymbol{\Sigma}_{f,k})\mathbf{B}^{-1} + \eta'\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}\right) \\
&= k + \mathrm{tr}(\eta'\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}) \le k + \eta \cdot k. \qquad (13)
\end{aligned}$$

Hence, from (12) and (13) we have

$$\begin{aligned}
&-\log p_{gp}(\bar{\mathscr{Y}}_{\le k}) + \log p_0(\bar{\mathscr{Y}}_{\le k}) \\
&\le \frac{1}{2}\Big(-\log|\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k}| + \log|\mathbf{B}| + k + \mathrm{tr}(\eta'\boldsymbol{\Sigma}_{f,k}\mathbf{B}^{-1}) + \|f_0\|_K^2 + \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{f,k}(\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k} - \mathbf{I}_k)\boldsymbol{\alpha} - k\Big) \\
&= \frac{1}{2}\Big(-\log|\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k}| + \log|\mathbf{B}| + \eta \cdot k + \|f_0\|_K^2 + \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{f,k}(\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k} - \mathbf{I}_k)\boldsymbol{\alpha}\Big).
\end{aligned}$$

Now since the covariance function is bounded and continuous in $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ almost surely, $\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k} - \mathbf{I}_k \to 0$ as $k \to \infty$. There exists a positive constant M such that for k large enough, $-\log|\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k}| < M$ and $\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{f,k}(\widehat{\boldsymbol{\Sigma}}_{f,k}^{-1}\boldsymbol{\Sigma}_{f,k} - \mathbf{I}_k)\boldsymbol{\alpha} < M$. Therefore, we have

$$-\log p_{gp}(\bar{\mathscr{Y}}_{\le k}) \le -\log p_0(\bar{\mathscr{Y}}_{\le k}) + \frac{1}{2}\|f_0\|_K^2 + M + \frac{1}{2}\log|\mathbf{B}| + \eta \cdot k$$

for any $f_0 \in \mathscr{H}_k$. Equivalently, we have

$$-\frac{1}{k}\log p_{gp}(\bar{\mathscr{Y}}_{\le k}) \le -\frac{1}{k}\log p_0(\bar{\mathscr{Y}}_{\le k}) + \frac{1}{k}\left(\frac{1}{2}\|f_0\|_K^2 + M + \frac{1}{2}\log|\mathbf{B}|\right) + \eta, \qquad (14)$$

for any $f_0 \in \mathscr{H}_k$.

By taking the infimum on the right-hand side of (14) over f_0 and applying the Representer theorem (See, e.g., Lemma 2 of Seeger et al. (2008)), we have

$$\frac{1}{k}\left(-\log p_{gp}(\bar{\mathscr{Y}}_{\le k}) + \log p_0(\bar{\mathscr{Y}}_{\le k})\right) \le \frac{1}{k}\left(\frac{1}{2}\|f_0\|_K^2 + M + \frac{1}{2}\log|\mathbf{B}|\right) + \eta,$$

for all $f_0 \in \mathscr{H}$. □

Lemma 1 provides a regret bound for SK-based prediction, competing against experts from the RKHS associated with the covariance function K of the GP stipulated by SK. The bound depends on the squared RKHS norm $\|f_0\|_K^2$, $\log|\mathbf{B}| = \log|\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}\boldsymbol{\Sigma}_{f,k} + \mathbf{I}_k|$ (referred to as the regret term, which also depends on K), and the simulation experiment design adopted. We next examine the regret term $\log|\mathbf{B}|$ using the Mercer eigenexpansion of the covariance function K. To lighten notation, below we omit the hyper-parameter vector $\boldsymbol{\theta}$ when referring to K. First, Let us recall Mercer's theorem; see further details from, e.g., Chatterji et al. (2019).

Theorem 1 (Mercer's theorem) Let $\mathscr{X} \subset \Re^d$ be compact and ν be a finite Borel measure with support $\mathscr{X}$. Suppose K is a continuous square integrable positive definite kernel on $\mathscr{X}$, and define a positive definite operator $\mathscr{T}_K : L_2(\mathscr{X};\nu) \mapsto L_2(\mathscr{X};\nu)$ by $(\mathscr{T}_K f)(\cdot) := \int_{\mathscr{X}} K(\cdot,\mathbf{x}) f(\mathbf{x}) d\nu$. Then there exists a sequence of eigenfunctions $\{\phi_m\}_{m\in\mathbb{N}}$ that forms an orthonormal basis of $L_2(\mathscr{X};\nu)$ consisting of eigenfunctions of $\mathscr{T}_K$, and an associated sequence of non-negative eigenvalues $\{\lambda_m\}_{m\in\mathbb{N}}$ such that $\mathscr{T}_K(\phi_m) = \lambda_m\phi_m$ for $m \in \mathbb{N}$. Moreover, K can be represented as $K(\mathbf{x},\mathbf{x}') = \sum_{m=1}^{\infty} \lambda_m \phi_m(\mathbf{x})\phi_m(\mathbf{x}')$, $\forall \mathbf{x},\mathbf{x}' \in \mathscr{X}$.

Assumption 1 Let K be a Mercer kernel satisfying Theorem 1.

1. The λ_m's are in a decreasing order, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$.
2. $\forall \mathbf{x},\mathbf{x}' \in \mathscr{X}, |K(\mathbf{x},\mathbf{x}')| \leq \bar{K}$, for some $\bar{K} > 0$.
3. $\forall m \geq 1, \forall \mathbf{x} \in \mathscr{X}, |\phi_m(\mathbf{x})| \leq \bar{\phi}$, for some $\bar{\phi} > 0$.

Lemma 2 (Lemma 1 in Vakili et al. (2021)) For all positive definite matrices $\mathbf{P} \in \Re^{n\times n}$, $\log\det(\mathbf{P}) \leq n\log(\mathrm{tr}(\mathbf{P})/n)$.

Theorem 2 Consider an SK metamodel with a covariance function K satisfying Assumption 1. For any $D \geq 1$,

$$\log|\mathbf{B}| \leq D\log\left(1 + \frac{1}{D}\bar{K}\sum_{i=1}^{k} t_i\right) + \delta_D \sum_{i=1}^{k} t_i, \tag{15}$$

where $\delta_D = \sum_{m=D+1}^{\infty} \lambda_m \bar{\phi}^2$ and $t_i = n_i/\mathsf{V}(\mathbf{x}_i)$, $i = 1,2,\ldots,k$.

Proof. To upper bound $\log|\mathbf{B}|$, we first consider a projection on a D-dimensional feature space $\boldsymbol{\phi}_D(\cdot) = (\phi_1(\cdot),\phi_2(\cdot),\ldots,\phi_D(\cdot))^\top$ spanned by the first D features (corresponding to the D largest eigenvalues of K). Specifically, define $K_p(\mathbf{x},\mathbf{x}') = \sum_{m=1}^{D} \lambda_m\phi_m(\mathbf{x})\phi_m(\mathbf{x}')$ and $K_o(\mathbf{x},\mathbf{x}') = K(\mathbf{x},\mathbf{x}') - K_p(\mathbf{x},\mathbf{x}')$ for any $\mathbf{x},\mathbf{x}' \in \mathscr{X}$. It follows immediately from Assumption 1 that $K_o(\mathbf{x},\mathbf{x}') \leq \delta_D$ for any $\mathbf{x},\mathbf{x}' \in \mathscr{X}$. Similarly, we have $\boldsymbol{\Sigma}_{f,k} = K(\mathbf{X},\mathbf{X}) = K_p(\mathbf{X},\mathbf{X}) + K_o(\mathbf{X},\mathbf{X})$, where $K_p(\mathbf{X},\mathbf{X})$ denotes the $k\times k$ matrix with its (i,j)th entry given by $K_p(\mathbf{x}_i,\mathbf{x}_j)$ $(i,j = 1,2,\ldots,k)$, and $K_o(\mathbf{X},\mathbf{X})$ denotes the corresponding orthogonal part. Therefore,

$$\begin{aligned}
\log|\mathbf{B}| &= \log|\mathbf{I}_k + \boldsymbol{\Sigma}_{\varepsilon,k}^{-1}\boldsymbol{\Sigma}_{f,k}| = \log|\mathbf{I}_k + \boldsymbol{\Sigma}_{\varepsilon,k}^{-1}(K_p(\mathbf{X},\mathbf{X}) + K_o(\mathbf{X},\mathbf{X}))| \\
&= \log|(\mathbf{I}_k + \boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_p(\mathbf{X},\mathbf{X}))(\mathbf{I}_k + (\mathbf{I}_k + \boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_p(\mathbf{X},\mathbf{X}))^{-1}\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_o(\mathbf{X},\mathbf{X}))| \\
&= \underbrace{\log|\mathbf{I}_k + \boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_p(\mathbf{X},\mathbf{X})|}_{:=\log(|\mathbf{B}|)_a} + \underbrace{\log|\mathbf{I}_k + (\mathbf{I}_k + \boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_p(\mathbf{X},\mathbf{X}))^{-1}\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_o(\mathbf{X},\mathbf{X})|}_{:=\log(|\mathbf{B}|)_b}.
\end{aligned} \tag{16}$$

We first consider the term $\log(|\mathbf{B}|)_a$. By eigen-decomposition, $K_p(\mathbf{X},\mathbf{X}) = \Phi_{k,D}\boldsymbol{\Lambda}\Phi_{k,D}^\top$, where $\Phi_{k,D} = (\boldsymbol{\phi}_D(\mathbf{x}_1),\boldsymbol{\phi}_D(\mathbf{x}_2),\ldots,\boldsymbol{\phi}_D(\mathbf{x}_k))^\top$ is a $k\times D$ matrix, whose ith row is given by the feature vector $\boldsymbol{\phi}_D^\top(\mathbf{x}_i)$, $i = 1,2,\ldots,k$; and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_D)$ is a $D\times D$ diagonal matrix. Denote $\mathbf{A} = \boldsymbol{\Sigma}_{\varepsilon,k}^{-1}\Phi_{k,D}\boldsymbol{\Lambda}^{\frac{1}{2}}$ and $\mathbf{C} = \boldsymbol{\Lambda}^{\frac{1}{2}}\Phi_{k,D}^\top$. Then, $\mathbf{AC} = \boldsymbol{\Sigma}_{\varepsilon,k}^{-1}\Phi_{k,D}\boldsymbol{\Lambda}\Phi_{k,D}^\top = \boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_p(\mathbf{X},\mathbf{X})$, where $\mathbf{AC}$ is a $k\times k$ matrix with $\mathrm{tr}(\mathbf{AC})$ being finite. By Weinstein–Aronszajin identity (Pozrikidis 2014), we have $|\mathbf{I}_k + \boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_p(\mathbf{X},\mathbf{X})| = |\mathbf{I}_D + \mathbf{G}_k|$, where $\mathbf{G}_k = \mathbf{CA} = \boldsymbol{\Lambda}^{\frac{1}{2}}\Phi_{k,D}^\top\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}\Phi_{k,D}\boldsymbol{\Lambda}^{\frac{1}{2}}$.

Furthermore,

$$\begin{aligned}\operatorname{tr}(\mathbf{I}_D+\mathbf{G}_k) &= D+\operatorname{tr}\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\Phi_{k,D}^{\top}\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}\Phi_{k,D}\boldsymbol{\Lambda}^{\frac{1}{2}}\right)\\ &= D+\sum_{i=1}^{k}t_i\operatorname{tr}\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\phi}_D(\mathbf{x}_i)\boldsymbol{\phi}_D^{\top}(\mathbf{x}_i)\boldsymbol{\Lambda}^{\frac{1}{2}}\right)\\ &= D+\sum_{i=1}^{k}t_i\sum_{m=1}^{D}\lambda_m\phi_m^2(\mathbf{x}_i)\le D+\bar{K}\sum_{i=1}^{k}t_i,\end{aligned} \tag{17}$$

where the third step follows from the fact that $\operatorname{tr}\left(\mathbf{A}\mathbf{A}^{\top}\right)=\operatorname{tr}\left(\mathbf{A}^{\top}\mathbf{A}\right)$ by setting $\mathbf{A}=\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\phi}_D(\mathbf{x}_i)$ and the last step holds because $\sum_{m=1}^{D}\lambda_m\phi_m^2(\mathbf{x})=K_p(\mathbf{x},\mathbf{x}')\le\bar{K}$ for any $\mathbf{x}\in\mathscr{X}$, thanks to Assumption 1. By Lemma 2 and (17), we can bound $\log(|\mathbf{B}|)_a$ as follows:

$$\log|\mathbf{I}_k+\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_p(\mathbf{X},\mathbf{X})|\le D\log\left(1+\frac{1}{D}\bar{K}\sum_{i=1}^{k}t_i\right). \tag{18}$$

The term $\log(|\mathbf{B}|)_b$ can be bounded in a similar fashion as done for $\log(|\mathbf{B}|)_a$. Specifically, we first note that

$$\operatorname{tr}((\mathbf{I}_k+\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_p(\mathbf{X},\mathbf{X}))^{-1}\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_o(\mathbf{X},\mathbf{X}))\le\operatorname{tr}(\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_o(\mathbf{X},\mathbf{X}))\bar{\lambda}\le\operatorname{tr}(\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_o(\mathbf{X},\mathbf{X})), \tag{19}$$

where the first inequality follows from applying the result in Fang et al. (1994) that $\operatorname{tr}(\mathbf{P}_1\mathbf{P}_2)\le\bar{\lambda}\operatorname{tr}(\mathbf{P}_2)$, with $\bar{\lambda}$ denoting the maximum eigenvalue of $\mathbf{P}_1$ and noting that $\bar{\lambda}\le 1$ for $\mathbf{P}_1=(\mathbf{I}_k+\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_p(\mathbf{X},\mathbf{X}))^{-1}$. Since for any $\mathbf{x},\mathbf{x}'\in\mathscr{X}$, $K_o(\mathbf{x},\mathbf{x}';\boldsymbol{\theta})\le\delta_D$, we further have

$$\operatorname{tr}(\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_o(\mathbf{X},\mathbf{X}))\le\delta_D\sum_{i=1}^{k}t_i. \tag{20}$$

Then, by Lemma 2, (19) and (20), we can bound $\log(|\mathbf{B}|)_b$ as follows:

$$\log|\mathbf{I}_k+(\mathbf{I}_k+\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_p(\mathbf{X},\mathbf{X}))^{-1}(\boldsymbol{\Sigma}_{\varepsilon,k}^{-1}K_o(\mathbf{X},\mathbf{X}))|\le k\log\left(1+\frac{1}{k}\delta_D\sum_{i=1}^{k}t_i\right)\le\delta_D\sum_{i=1}^{k}t_i, \tag{21}$$

where the second inequality follows since $\log(1+x)\le x$ for any $x\ge 0$. Finally, combining (16), (18), and (21) yields

$$\log|\mathbf{B}|\le D\log\left(1+\frac{1}{D}\bar{K}\sum_{i=1}^{k}t_i\right)+\delta_D\sum_{i=1}^{k}t_i.$$

□

Under some further assumptions on the eigendecay profile of the covariance function or kernel K, we can obtain specific orders of $\log|\mathbf{B}|$.

Definition 1 Consider the sequence of eigenvalues of a kernel K satisfying Assumption 1.

1. For some $C_p>0,\beta_p>1$, K is said to be a (C_p,β_p) polynomial eigendecay kernel, if for any $m\in\mathbb{N}$, $\lambda_m\le C_p m^{-\beta_p}$.
2. For some $C_{e,1},C_{e,2},\beta_e>0$, K is said to be a $(C_{e,1},C_{e,2},\beta_e)$ exponential eigendecay kernel, if for any $m\in\mathbb{N}$, $\lambda_m\le C_{e,1}\exp(-C_{e,2}m^{\beta_e})$.

Corollary 3 Denote $T_k = \sum_{i=1}^{k} t_i = \sum_{i=1}^{k} n_i/\mathsf{V}(\mathbf{x}_i)$ and consider the regret term, $\log|\mathbf{B}|$, as defined in Lemma 1 and studied in Theorem 2.

1. If K has a polynomial eigendecay, $\log|\mathbf{B}| \leq 2\left[\left(C_p\bar{\phi}^2 T_k\right)^{\frac{1}{\beta_p}} \log^{-\frac{1}{\beta_p}}(1+\bar{K}T_k)+1\right]\log(1+\bar{K}T_k)$. That is, $\log|\mathbf{B}| = \mathscr{O}\left(T_k^{\frac{1}{\beta_p}} \log^{1-\frac{1}{\beta_p}} T_k\right)$.
2. If K has an exponential eigendecay, $\log|\mathbf{B}| \leq 2\left[\left(2C_{e,2}^{-1}\left(\log T_k + C_{\beta_e}\right)\right)^{\frac{1}{\beta_e}}+1\right]\log(1+\bar{K}T_k)$. That is, $\log|\mathbf{B}| = \mathscr{O}\left(\log^{1+\frac{1}{\beta_e}} T_k\right)$.

Proof. The proof is in the same vein as that of Corollary 1 in Vakili et al. (2021). Under the polynomial eigendecay condition, we have

$$\delta_D = \sum_{m=D+1}^{\infty} \lambda_m \bar{\phi}^2 \leq \sum_{m=D+1}^{\infty} C_p m^{-\beta_p}\bar{\phi}^2 \leq \int_D^{\infty} C_p z^{-\beta_p}\bar{\phi}^2 dz = C_p D^{1-\beta_p}\bar{\phi}^2.$$

Then, D is selected to be $\lceil (C_p\bar{\phi}^2 T_k)^{\frac{1}{\beta_p}} \log^{-\frac{1}{\beta_p}}(1+\bar{K}T_k)\rceil$ to ensure $\delta_D T_k \leq \log(1+\bar{K}T_k)$ by directly solving the inequality. Thus, in light of Lemma 2, we have $\log|\mathbf{B}| \leq 2\left[\left(C_p\bar{\phi}^2 T_k\right)^{\frac{1}{\beta_p}} \log^{-\frac{1}{\beta_p}}(1+\bar{K}T_k)+1\right]\log(1+\bar{K}T_k)$.

Under the exponential eigendacay condition, δ_D can be upper bounded as follows:

$$\delta_D = \sum_{m=D+1}^{\infty} \lambda_m \bar{\phi}^2 \leq \sum_{m=D+1}^{\infty} C_{e,1}\exp(-C_{e,2}m^{\beta_e})\bar{\phi}^2 \leq \int_D^{\infty} C_{e,1}\exp(-C_{e,2}z^{\beta_e})\bar{\phi}^2 dz. \tag{22}$$

We further consider two cases: $\beta_e = 1$ and $\beta_e \neq 1$. The selection of D in both cases intends to ensure that $\delta_D T_k \leq \log(1+\bar{K}T_k)$ by setting the left-hand side to 1, which is dominated by the right-hand side when T_k is reasonably large. When $\beta_e = 1$, it can be easily shown by (22) that $\delta_D \leq C_{e,1}C_{e,2}^{-1}\exp(-C_{e,2}D)\bar{\phi}^2$. Thus, D is selected to be $\left\lceil C_{e,2}^{-1}\log\left(C_{e,1}\bar{\phi}^2 T_k C_{e,2}^{-1}\right)\right\rceil$. When $\beta_e \neq 1$, δ_D can be bounded by $2C_{e,1}(C_{e,2}\beta_e)^{-1}\left(2C_{e,2}^{-1}\left(\beta_e^{-1}-1\right)\right)^{\frac{1}{\beta_e}-1}\exp\left(-\left(\beta_e^{-1}-1\right)\right)\exp\left(-C_{e,2}D^{\beta_e}/2\right)\bar{\phi}^2$ following similar steps as given in the proof of Corollary 1 in Vakili et al. (2021). It follows that D can be selected to be $\left\lceil\left(2C_{e,2}^{-1}\left[\log T_k + \log\left(2C_{e,1}\bar{\phi}^2(\beta_e C_{e,2})^{-1}\right)+\left(\beta_e^{-1}-1\right)\left(\log\left[2C_{e,2}^{-1}\left(\beta_e^{-1}-1\right)\right]-1\right)\right]\right)^{\frac{1}{\beta_e}}\right\rceil$. Hence, in light of Lemma 2 and the analysis above, we have $\log|\mathbf{B}| \leq 2\left[\left(2C_{e,2}^{-1}\left(\log T_k + C_{\beta_e}\right)\right)^{\frac{1}{\beta_e}}+1\right]\log(1+\bar{K}T_k)$, where $C_{\beta_e} = \log(C_{e,1}\bar{\phi}^2 C_{e,2}^{-1})$ when $\beta_e = 1$ and $C_{\beta_e} = \log(2C_{e,1}\bar{\phi}^2(\beta_e C_{e,2})^{-1}) + (\beta_e^{-1}-1)(\log[2C_{e,2}^{-1}(\beta_e^{-1}-1)]-1)$ when $\beta_e \neq 1$. □

Remark 1 Corollary 3 enables one to obtain specific orders of $\log|\mathbf{B}|$ for kernels commonly used in practice. Recall that d denotes the input-space dimensionality.

1. As a special case of polynomial eigendecay kernels, a Matétrn kernel has $\lambda_m = \mathscr{O}(m^{-\frac{2\nu+d}{d}})$ for any $m \in \mathbb{N}$, where $\nu > 1/2$ denotes the smooth parameter. In this case, $\log|\mathbf{B}| = \mathscr{O}\left(T_k^{\frac{d}{2\nu+d}} \log^{\frac{2\nu}{2\nu+d}} T_k\right)$.
2. As a special case of exponential eigendecay kernels, a squared exponential kernel has $\lambda_m = \mathscr{O}\left(\exp\left(-m^{\frac{1}{d}}\right)\right)$ for any $m \in \mathbb{N}$. In this case, $\log|\mathbf{B}| = \mathscr{O}\left(\log^{d+1} T_k\right)$.

In light of Corollary 3, we further stipulate some mild assumptions on the budget allocation of SK so that $\log|\mathbf{B}|$ remains reasonably small with the increase of k for kernels with different eigendecay profiles.

Assumption 2 The total simulation budget $B_k = \sum_{i=1}^{k} n_i$ satisfies $B_k = \mathscr{O}(k^b)$ for some $b > 0$.

Assumption 3 The total simulation budget satisfying $B_k = \mathscr{O}(k^b)$, where $0 < b < \beta_p$ holds if K is a polynomial eigendecay kernel.

Corollary 4 Under Assumption 2, $\log|\mathbf{B}| = o(k)$ for exponential eigendecay kernels. Under Assumption 3, $\log|\mathbf{B}| = o(k)$ for polynomial eigendecay kernels.

Proof. Recall its definition in Corollary 3, $T_k = \sum_{i=1}^{k} t_i = \sum_{i=1}^{n} n_i/\mathsf{V}(\mathbf{x}_i)$. We have $T_k \leq \sum_{i=1}^{k} n_i/\underline{\mathsf{V}} = B_k/\underline{\mathsf{V}}$, where $\underline{\mathsf{V}} = \inf_{\mathbf{x}\in\mathscr{X}} \mathsf{V}(\mathbf{x}) > 0$. Under Assumption 2 (respectively, Assumption 3), $T_k = \mathscr{O}(k^b/\underline{\mathsf{V}}) = \mathscr{O}(k^b)$.

In light of Corollary 3, we further have, for a polynomial eigendecay kernel, that

$$\log|\mathbf{B}| = \mathscr{O}\left(T_k^{\frac{1}{\beta_p}} \log^{1-\frac{1}{\beta_p}} T_k\right) = \mathscr{O}\left(k^{\frac{b}{\beta_p}} \log^{1-\frac{1}{\beta_p}}\left(k^b\right)\right) = \mathscr{O}\left(k^{\frac{b}{\beta_p}} \log^{1-\frac{1}{\beta_p}}(k)\right) = o(k),$$

where the last step follows from Assumption 3.

For an exponential eigendecay kernel, we have

$$\log|\mathbf{B}| = \mathscr{O}\left(\log^{\frac{1}{\beta_e}+1} T_k\right) = \mathscr{O}\left(\log^{\frac{1}{\beta_e}+1}(k^b)\right) = \mathscr{O}\left(\log^{\frac{1}{\beta_e}+1}(k)\right) = o(k).$$

□

In light of Lemma 1 and Corollary 4, we arrive at the following result regarding information consistency of SK.

Theorem 5 Under an SK metamodel as described in Section 2, the conditions given in Lemma 1, and Assumptions 1 to 3, $k^{-1}\mathrm{E}_{\mathbf{x}_{\leq k}}\left(D[p_0(\bar{\mathscr{Y}}_{\leq k}), p_{gp}(\bar{\mathscr{Y}}_{\leq k})]\right) \to 0$ as $k \to \infty$.

Proof. It follows from the definition of information consistency that

$$\begin{aligned} D\left[p_0(\bar{\mathscr{Y}}_{\leq k}), p_{gp}(\bar{\mathscr{Y}}_{\leq k})\right] &= \int_{\mathscr{R}^k} p_0(\bar{\mathscr{Y}}_{\leq k}) \log\left(\frac{p_0(\bar{\mathscr{Y}}_{\leq k})}{p_{gp}(\bar{\mathscr{Y}}_{\leq k})}\right) \mathrm{d}\bar{\mathscr{Y}}_1 \ldots \mathrm{d}\bar{\mathscr{Y}}_k \\ &= \int_{\mathscr{R}^k} p_0(\bar{\mathscr{Y}}_{\leq k}) \left(\log p_0(\bar{\mathscr{Y}}_{\leq k}) - \log p_{gp}(\bar{\mathscr{Y}}_{\leq k})\right) \mathrm{d}\bar{\mathscr{Y}}_1 \ldots \mathrm{d}\bar{\mathscr{Y}}_k. \end{aligned} \tag{23}$$

From Corollary 4, we have $\log|\mathbf{B}| = o(k)$ for polynomial eigendecay kernels and exponential eigendecay kernels. Furthermore, since $f_0 \in \mathscr{H}$, the RKHS associated with K, $\|f_0\|_K < \infty$. Hence, it follows from Lemma 1 that

$$\frac{1}{k}\mathrm{E}_{\mathbf{x}_{\leq k}}\left(D\left[p_0(\bar{\mathscr{Y}}_{\leq k}), p_{gp}(\bar{\mathscr{Y}}_{\leq k})\right]\right) \leq \frac{1}{2k}\|f_0\|_K^2 + \frac{1}{2k}\log|\mathbf{B}| + \frac{M}{k} \to 0, \qquad \text{as } k \to \infty.$$

□

Remark 2 Corollary 4 provides some sufficient conditions on budget allocation that ensure information consistency of SK as the number of design points k and the total simulation budget allocated B_k approach infinity. For estimating a sufficiently smooth mean response function f_0, i.e., f_0 lying in an RKHS of an exponential eigendecay kernel K, there is no particular requirement on the budget allocation to achieve information consistency so long as $B_k, k \to \infty$. Interestingly, for estimating a mean response function f_0 that is not that smooth, i.e., f_0 belonging to an RKHS of a polynomial eigendecay kernel K, the budget allocated to the k design points should avoid growing too quickly as $k \to \infty$. In particular, for f_0 contained in the RKHS of a Matétrn kernel, we see that the budget $B_k = O(k^b)$ must satisfy that $b < \beta_p = (2\nu + d)/d$ for SK to achieve information consistency. The smoother the function f_0 (i.e., the greater ν) is, the less stringent the condition is on the budget allocated B_k. A converse effect holds for the input-space dimensionality d on B_k. Therefore, we recommend setting k relatively large and using relatively few replications at each given a fixed budget to expend. This echoes with the suggestion given by Wang and Chen (2018) to use a "dense and shallow" design for SK.

We remark on Theorem 5 with some interpretations. From Theorem 5, the KL divergence between the two distribution functions for $\bar{\mathscr{Y}}_{\leq k}|\mathbf{x}_{\leq k}$ from the true and the assumed models becomes zero, asymptotically. For the sake of brevity, we note without showing details that

$$p_{gp}(\bar{\mathscr{Y}}_{\leq k}) := p_{\widehat{\theta}}(\bar{\mathscr{Y}}_{\leq k}|\mathbf{x}_{\leq k}) = \prod_{i=1}^{k} p_{\widehat{\theta}}(\bar{\mathscr{Y}}_i|\mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i}), \tag{24}$$

where

$$p_{\widehat{\theta}}(\bar{\mathscr{Y}}_i|\mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i}) = \int_{\mathscr{F}} p(\bar{\mathscr{Y}}_i|f, \mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i}) \mathrm{d}p_{\widehat{\theta}}(f|\mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i}), \quad \mathrm{d}p_{\widehat{\theta}}(f|\mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i}) = \frac{p(\bar{\mathscr{Y}}_{<i}|f, \mathbf{x}_{<i}) \mathrm{d}p_{\widehat{\theta}}(f)}{\int_{\mathscr{F}} p(\bar{\mathscr{Y}}_{<i}|f', \mathbf{x}_{<i}) \mathrm{d}p_{\widehat{\theta}}(f')}.$$

Similarly, under the true model, it holds that

$$p_0(\bar{\mathscr{Y}}_{\leq k}) := p(\bar{\mathscr{Y}}_{\leq k}|f_0, \mathbf{x}_{\leq k}) = \prod_{i=1}^{k} p(\bar{\mathscr{Y}}_i|f_0, \mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i}). \tag{25}$$

Paralleling the developments given in Seeger et al. (2008) and Wang and Shi (2014), we refer to $p(\bar{\mathscr{Y}}_{\leq i}|f_0, \mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i})$ and $p_{\widehat{\theta}}(\bar{\mathscr{Y}}_{\leq i}|\mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i})$ as Bayesian prediction strategies. We can see from (23) to (25) that

$$D\left[p_0(\bar{\mathscr{Y}}_{\leq k}), p_{gp}(\bar{\mathscr{Y}}_{\leq k})\right] = \int_{\mathscr{R}^k} \sum_{i=1}^{k} Q(\bar{\mathscr{Y}}_i|\mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i}) p_0(\bar{\mathscr{Y}}_{\leq k}) \mathrm{d}\bar{\mathscr{Y}}_{\leq k},$$

where $Q(\bar{\mathscr{Y}}_i|\mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i}) = \log p(\bar{\mathscr{Y}}_{\leq i}|f_0, \mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i}) - \log p_{\widehat{\theta}}(\bar{\mathscr{Y}}_i|\mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i})$ is a loss function for $i \geq 2$ (with $Q(\bar{\mathscr{Y}}_i|\mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i}) = 0$ for $i = 1$), and $\sum_{i=1}^{k} Q(\bar{\mathscr{Y}}_i|\mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i})$ is referred to as cumulative log loss. Similar to the results as given in Seeger et al. (2008) and Wang and Shi (2014), Theorem 5 here can be interpreted as the average of cumulative log loss $k^{-1}\sum_{i=1}^{k} Q(\bar{\mathscr{Y}}_i|\mathbf{x}_{\leq i}, \bar{\mathscr{Y}}_{<i})$ approaching zero asymptotically. Hence, we have related the information consistency of SK to sequential prediction under cumulative log loss.

4 CONCLUSIONS

In this paper, we proved information consistency results for SK metamodels. We showed that information consistency of SK depends on not only the covariance function of the GP prior but also the budget allocation adopted in the stochastic simulation experiment. Our investigation has focused on the case where the true mean response function f_0 is contained in the RKHS of the GP prior adopted by SK, hence one potential extension is to elaborate these results for SK in a more accurate way by studying small ball probabilities and entropy calculations in the GP priors as performed in van der Vaart and van Zanten (2011) for standard GPR models. Moreover, in our analysis we have assumed that the simulation noise variances $\mathsf{V}(\mathbf{x}_i)$'s are known and hence the noise variance-covariance matrix $\mathbf{\Sigma}_{\varepsilon,k}$ is given. Another potential direction is to extend the analysis to the case where estimation of simulation noise variances is required.

ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation under Grants No. CMMI-1846663 and IIS-1849300.

REFERENCES

Ankenman, B. E., B. L. Nelson, and J. Staum. 2010. "Stochastic Kriging for Simulation Metamodeling". *Operations Research* 58:371–382.

Chatterji, N., A. Pacchiano, and P. Bartlett. 2019. "Online Learning with Kernel Losses". In *Proceedings of the 36th International Conference on Machine Learning*, edited by K. Chaudhuri and R. Salakhutdinov, 971–980. Long Beach, CA.

Chen, X., 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* 22:7/1–7/20.

Chen, X., B. E. Ankenman, and B. L. Nelson. 2013. "Enhancing Stochastic Kriging Metamodels with Gradient Estimators". *Operations Research* 61:512–528.

Chen, X., and K.-K. Kim. 2014. "Stochastic Kriging with Biased Sample Estimates". *ACM Transactions on Modeling and Computer Simulation* 24:8/1–8/23.

Chen, X., and K.-K. Kim. 2016. "Efficient VaR and CVaR Measurement via Stochastic Kriging". *INFORMS Journal on Computing* 28:629–644.

Choi, T., and R. V. Ramamoorthi. 2008. "Remarks on consistency of posterior distributions". In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanata K. Ghosh*, Volume 3, 170–186. Beachwood, OH: Institute of Mathematics Statistics.

Fang, Y., K. Loparo, and X. Feng. 1994. "Inequalities for the Trace of Matrix Product". *IEEE Transactions on Automatic Control* 39:2489–2490.

Gratiet, L. L., and J. Garnier. 2015. "Asymptotic Analysis of the Learning Curve for Gaussian process regression". *Journal of Machine Learning* 98:407–433.

Liu, M., and J. Staum. 2010. "Stochastic Kriging for Efficient Nested Simulation of Expected Shortfall". *Journal of Risk* 12:3–27.

Ng, S. H., and J. Yin. 2012. "Bayesian Kriging Analysis and Design for Stochastic Simulations". *ACM Transactions on Modeling and Computer Simulation* 22:1–26.

Osorio, C., and M. Bierlaire. 2013. "A Simulation-based Optimization Framework for Urban Transportation Problems". *Operations Research* 61:1333–1345.

Ouyang, L., Y. Ma, J. Wang, and Y. Tu. 2017. "A New Loss Function for Multi-response Optimization with Model Parameter Uncertainty and Implementation Errors". *European Journal of Operational Research* 258:552–563.

Pozrikidis, C. 2014. *An Introduction to Grids, Graphs, and Networks*. Oxford University Press.

Rasmussen, C. E., and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. Massachusetts: the MIT Press.

Santner, T. J., B. J. Williams, and W. I. Notz. 2003. *The Design and Analysis of Computer Experiments*. New York: Springer.

Santos, M. I., and P. M. Santos. 2016. "Switching Regression Metamodels in Stochastic Simulation". *European Journal of Operational Research* 251:142–147.

Seeger, M. W., S. M. Kakade, and D. P. Foster. 2008. "Information Consistency of Nonparametric Gaussian Process Methods". *IEEE Transactions on Information Theory* 54:2376–2382.

Shi, J., and T. Choi. 2011. *Gaussian Process Regression Analysis for Functional Data*. Taylor & Francis Group, LLC.

Vakili, S., K. Khezeli, and V. Picheny. 2021. "On Information Gain and Regret Bounds in Gaussian Process Bandits". In *Proceedings of 2021 International Conference on Artificial Intelligence and Statistics*, edited by A. Banerjee and K. Fukumizu, 82–90.

van Beers, W. C. M., and J. P. C. Kleijnen. 2008. "Customized Sequential Designs for Random Simulation Experiments: Kriging Metamodeling and Bootstrapping". *European Journal of Operational Research* 186:1099–1113.

van der Vaart, A., and H. van Zanten. 2011. "Information Rates of Nonparametric Gaussian Process Methods". *Journal of Machine Learning Research* 12:2095–2119.

Vazquez, E., and J. Bect. 2010. "Pointwise Consistency of the Kriging Predictor with Known Mean and Covariance Functions". In *mODa9–Advances in Model-Oriented Design and Analysis*, edited by A. Giovagnoli, A. Atkinson, B. Torsney, and C. May, 221–228. Physica-Verlag HD.

Wang, B., and J. Q. Shi. 2014. "Generalized Gaussian Process Regression Model for Non-Gaussian Functional Data". *Journal of the American Statistical Association* 109:1123–1133.

Wang, W., and X. Chen. 2018. "An Adaptive Two-stage Dual Metamodeling Approach for Stochastic Simulation Experiments". *IISE Transactions* 50:820–836.

AUTHOR BIOGRAPHIES

Yutong Zhang is Ph.D. student in the Grado department of Industrial and Systems Engineering at Virginia Tech. Her research interest lies in machine learning, stochastic modeling, and simulation methodology. Her email address is yutongz@vt.edu.

Xi Chen is an associate professor in the Grado department of Industrial and Systems Engineering at Virginia Tech. Her research interests include stochastic modeling and simulation, applied probability and statistics, computer experiment design and analysis, and simulation optimization. Her email address is xchen6@vt.edu and her web page is https://sites.google.com/vt.edu/xi-chen-ise/home.