Journal of the American Statistical Association

June 2020 • Volume 115 • Number 530

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

On Prediction Properties of Kriging: Uniform Error Bounds and Robustness

Wenjia Wang, Rui Tuo & C. F. Jeff Wu

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 530, 920–930: Theory and Methods
https://doi.org/10.1080/01621459.2019.1598868

Taylor & Francis
Taylor & Francis Group

On Prediction Properties of Kriging: Uniform Error Bounds and Robustness

Wenjia Wang[a], Rui Tuo[b], and C. F. Jeff Wu[c]

[a]The Statistical and Applied Mathematical Sciences Institute, Durham, NC; [b]Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX; [c]The H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA

ABSTRACT

Kriging based on Gaussian random fields is widely used in reconstructing unknown functions. The kriging method has pointwise predictive distributions which are computationally simple. However, in many applications one would like to predict for a range of untried points simultaneously. In this work, we obtain some error bounds for the simple and universal kriging predictor under the uniform metric. It works for a scattered set of input points in an arbitrary dimension, and also covers the case where the covariance function of the Gaussian process is misspecified. These results lead to a better understanding of the rate of convergence of kriging under the Gaussian or the Matérn correlation functions, the relationship between space-filling designs and kriging models, and the robustness of the Matérn correlation functions. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2017
Accepted March 2019

KEYWORDS
Gaussian process modeling; Radial basis functions; Space-filling designs; Spatial statistics; Uniform convergence.

1. Introduction

Kriging is a widely used methodology to reconstruct functions based on their scattered evaluations. Originally, kriging was introduced to geostatistics by Matheron (1963). Later, it has been applied to computer experiments (Sacks et al. 1989), machine learning (Rasmussen 2006), small area estimation from survey data (Rao and Molina 2015), and other areas. With kriging, one can obtain an interpolant of the observed data, that is, the predictive curve or surface goes through all data points. Conventional regression methods, like the linear regression, the local polynomial regression (Fan and Gijbels 1996), and the smoothing splines (Wahba 1990), do not have this property. It is suitable to use interpolation in spatial statistics and machine learning when the random noise of the data is negligible. The interpolation property is particularly helpful in computer experiments, in which the aim is to construct a surrogate model for a deterministic computer code, such as a finite element solver.

A key element of kriging prediction is the use of conditional inference based on Gaussian processes. At each untried point of the design region (i.e., domain for the input variables), the conditional distribution of a Gaussian process is normal with explicit mean and variance. The pointwise confidence interval of the kriging predictor is then constructed using this conditional distribution. In many applications, it is desirable to have a joint confidence region of the kriging predictor over a continuous set of the input variables such as an interval or rectangular region. The pointwise confidence interval for each design point cannot be amalgamated over the points in the design region to give a confidence region/limit with guaranteed coverage probability, even asymptotically. To address this question, it would be desirable to have a theory that gives good bounds on the worst (i.e., maximum) error of the kriging predictor over the design region. This bound can be useful in the construction of confidence regions with guaranteed coverage property, albeit somewhat conservatively.

In this work, we derive error bounds of the simple and universal kriging predictor under a uniform metric. The predictive error is bounded in terms of the maximum pointwise predictive variance of kriging. A key implication of our work is to show that the overall predictive performance of a Gaussian process model is tied to the smoothness of the underlying correlation function as well as the space-filling property of the design (i.e., collection of the design points). This has two major consequences. First, we show that a less smooth correlation function is more *robust* in prediction, in the sense that prediction consistency can be achieved for a broader range of true correlation functions, while a smoother correlation function can achieve a higher rate of convergence provided that it is no smoother than the true correlation. Second, these error bounds are closely related to the *fill distance*, which is a space-filling property of the design. This suggests that it makes a good design by minimizing its fill distance. We also prove a similar error bound for universal kriging with a random kernel function. In addition, our theory shows that the maximum likelihood estimator for the regression coefficient of universal kriging can be *inconsistent*, which is a new result to the best of our knowledge.

This paper is organized as follows. In Section 2, we review the mathematical foundation of simple kriging and state the objectives of this paper. In Section 3, we present our main results on the uniform error bounds for kriging predictors. Comparison with existing results in the literature is given in Section 3.3. Some simulation studies are presented in Section 4, which confirm our theoretical analysis. We extend our theory from simple kriging to universal kriging in Section 5. Concluding remarks and

CONTACT Wenjia Wang ✉ wenjia.wang234@duke.edu The Statistical and Applied Mathematical Sciences Institute, Durham, NC 27709.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

discussion are given in Section 6. Appendix A contains the proof of Theorem 1, the main theorem of this work. Appendices B–D consist of the proofs of Theorems 2–4, respectively. Some necessary mathematical tools are reviewed in the supplementary materials.

2. Preliminaries and Motivation

In Sections 2.1 and 2.2, we review the kriging method and introduce some proper notation. In Section 2.3, we state the primary goal of our work.

2.1. Review on the Simple Kriging Method

Let $Z(\boldsymbol{x})$ be a Gaussian process on $\mathbf{R}^d$. In this work, we suppose that Z has mean zero and is *stationary*, that is, the covariance function of Z depends only on the difference between the two input variables. Specifically, we denote

$$\mathrm{cov}(Z(\boldsymbol{x}), Z(\boldsymbol{x}')) = \sigma^2\Psi(\boldsymbol{x}-\boldsymbol{x}'),$$

for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbf{R}^d$, where σ^2 is the variance and Ψ is the correlation function. The correlation function should be positive definite and satisfy $\Psi(0) = 1$. In particular, we consider two important families of correlation functions. The isotropic Gaussian correlation function is defined as

$$\Psi(\boldsymbol{x};\phi) = \exp\{-\phi\|\boldsymbol{x}\|^2\}, \tag{2.1}$$

with some $\phi > 0$, where $\|\cdot\|$ denotes the Euclidean norm. The isotropic Matérn correlation function (Santner, Williams, and Notz 2003; Stein 1999) is defined as

$$\Psi(\boldsymbol{x};\nu,\phi) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(2\sqrt{\nu}\phi\|\boldsymbol{x}\|)^{\nu}K_{\nu}(2\sqrt{\nu}\phi\|\boldsymbol{x}\|), \tag{2.2}$$

where $\phi, \nu > 0$ and K_ν is the modified Bessel function of the second kind. The parameter ν is often called the smoothness parameter, because it determines the smoothness of the Gaussian process (Cramér and Leadbetter 1967).

Suppose that we have observed $Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)$, in which $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are distinct points. We shall use the terminology in design of experiments (Wu and Hamada 2009) and call $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ the *design points*, although in some situations (e.g., in spatial statistics and machine learning) these points are observed without the use of design. In this article, we do not assume any (algebraic or geometric) structure for the design points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. They are called *scattered points* in the applied mathematics literature.

The aim of *simple kriging* is to predict $Z(\boldsymbol{x})$ at an untried $\boldsymbol{x}$ based on the observed data $Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)$, which is done by calculating the conditional distribution. It follows from standard arguments (Santner, Williams, and Notz 2003; Banerjee, Carlin, and Gelfand 2004) that, conditional on $Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)$, $Z(\boldsymbol{x})$ is normally distributed, with

$$\mathbb{E}[Z(\boldsymbol{x})|Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)] = \boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\boldsymbol{Y}, \text{ a.s.,} \tag{2.3}$$

$$\mathrm{var}[Z(\boldsymbol{x})|Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)] = \sigma^2(1-\boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\boldsymbol{r}(\boldsymbol{x})), \text{ a.s.,} \tag{2.4}$$

where $\boldsymbol{r}(\boldsymbol{x}) = (\Psi(\boldsymbol{x}-\boldsymbol{x}_1), \ldots, \Psi(\boldsymbol{x}-\boldsymbol{x}_n))^T, \mathbf{K} = (\Psi(\boldsymbol{x}_j-\boldsymbol{x}_k))_{jk}$, and $\boldsymbol{Y} = (Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n))^T$.

The conditional expectation $\mathbb{E}[Z(\boldsymbol{x})|Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)]$ is a natural predictor of $Z(\boldsymbol{x})$ using $Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)$, because it is the best linear predictor (Stein 1999; Santner, Williams, and Notz 2003). It is worth noting that a nice property of the Gaussian process models is that the predictor (2.3) has an explicit expression, which explains why kriging is so popular and useful.

The above simple kriging method can be extended. Instead of using a mean zero Gaussian process, one may introduce extra degrees of freedom by assuming that the Gaussian process has an unknown constant mean. More generally, one may assume the mean function is given by a linear combination of known functions. The corresponding methods are referred to as ordinary kriging and universal kriging, respectively. A standard prediction scheme is the best linear unbiased prediction (Santner, Williams, and Notz 2003; Stein 1999). In this work, we shall first consider simple kriging in Sections 2.2–4, and then extend our results to universal kriging in Section 5. This organization is based on the following reasons: (1) the predictive mean of simple kriging (2.3) is identical to the radial basis function interpolant (see Section 2.2), which is an important mathematical tool which our theory relies on, (2) our main theorem for simple kriging (Theorem 1) requires less regularity conditions than those for universal kriging, (3) our theory for simple kriging, together with the techniques we develop to prove Theorem 1, serves as a basis for establishing the results for universal kriging.

2.2. Kriging Interpolant

The conditional expectation in (2.3) defines an interpolation scheme. To see this, let us suppress the randomness in the probability space and then $Z(\boldsymbol{x})$ becomes a deterministic function, often called a sample path. It can be verified that, as a function of $\boldsymbol{x}$, $\boldsymbol{r}^T\mathbf{K}^{-1}\boldsymbol{Y}$ in (2.3) goes through each $Z(\boldsymbol{x}_j), j = 1, \ldots, n$.

The above interpolation scheme can be applied to an arbitrary function f. Specifically, given design points $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and observations $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)$, we define the *kriging interpolant* by

$$\mathcal{I}_{\Psi,\mathbf{X}}f(\boldsymbol{x}) = \boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\boldsymbol{F}, \tag{2.5}$$

where $\boldsymbol{r}(\boldsymbol{x}) = (\Psi(\boldsymbol{x}-\boldsymbol{x}_1), \ldots, \Psi(\boldsymbol{x}-\boldsymbol{x}_n))^T, \mathbf{K} = (\Psi(\boldsymbol{x}_j-\boldsymbol{x}_k))_{jk}$, and $\boldsymbol{F} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))^T$. This interpolation scheme is also referred to as the *radial basis function* interpolation (Wendland 2004). The only difference between (2.5) and (2.3) is that we replace the Gaussian process Z by a function f here. In other words,

$$\mathbb{E}[Z(\boldsymbol{x})|Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)] = \mathcal{I}_{\Psi,\mathbf{X}}Z(\boldsymbol{x}), \text{ a.s.} \tag{2.6}$$

As mentioned in Section 2.1, the conditional expectation $\mathbb{E}[Z(\boldsymbol{x})|Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)]$ is a natural predictor of $Z(\boldsymbol{x})$. A key objective of this work is to derive a uniform bound of the predictive error of the kriging method, given by $Z(\boldsymbol{x}) - \mathbb{E}[Z(\boldsymbol{x})|Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)]$, which is equal to $Z(\boldsymbol{x}) - \mathcal{I}_{\Psi,\mathbf{X}}Z(\boldsymbol{x})$ almost surely.

In practice, Ψ is usually unknown. Thus, it is desirable to develop a theory that also covers the cases with misspecified correlations. In this work, we suppose that we use another correlation function Φ for prediction. We call Ψ the *true correlation function* and Φ the *imposed correlation function*. Under the imposed correlation function, the kriging interplant of the underlying Gaussian process becomes $\mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x})$. In this situation, the interpolant cannot be interpreted as the conditional expectation. With an abuse of terminology, we still call it a kriging predictor.

2.3. Goal of This Work

Our aim is to study the approximation power of the kriging predictor. For simple kriging, we are interested in bounding the *maximum prediction error* over a region Ω,

$$\sup_{\boldsymbol{x}\in\Omega} |Z(\boldsymbol{x}) - \mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x})| \tag{2.7}$$

in a probabilistic manner, where Ω is the region of interest, also called the experimental region, and $\Omega \supset \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$. For universal kriging, our aim is to bound a quantity similar to (2.7), but in which $\mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x})$ should be replaced by the best linear unbiased predictor, or a more general predictor given by universal kriging with an estimated kernel function.

Our obtained results on the error bound in (2.7) can be used to address or answer the following three questions.

First, the quantity (2.7) captures the worst case prediction error of kriging. In many practical problems, we are interested in recovering a whole function rather than predicting for just one point. Therefore, obtaining uniform error bounds are of interest because they provide some insight on how we can modify the pointwise error bound to achieve a uniform coverage.

Second, we study the case of misspecified correlation functions. This address a common question in kriging when the true correlation function is unknown: how to gain model robustness under a misspecified correlation function and how much efficiency loss is incurred.

Third, our framework allows the study of an arbitrary set of design points (also called scattered points). Thus, our results can facilitate the study of kriging with fixed or random designs. In addition, our theory can be used to justify the use of space-filling designs (Santner, Williams, and Notz 2003), in which the design points spread (approximately) evenly in the design region.

3. Uniform Error Bounds for Simple Kriging

This section contains our main theoretical results on the prediction error of simple kriging.

3.1. Error Bound in Terms of the Power Function

It will be shown that, the predictive variance (2.4) plays a curial role on the prediction error, when the true correlation function is known, that is, $\Phi = \Psi$. To incorporate the case of misspecified correlation functions, we define the *power function* as

$$P^2_{\Phi,\mathbf{X}}(\boldsymbol{x}) = 1 - \boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\boldsymbol{r}(\boldsymbol{x}), \tag{3.1}$$

where $\boldsymbol{r}(\boldsymbol{x}) = (\Phi(\boldsymbol{x}-\boldsymbol{x}_1), \dots, \Phi(\boldsymbol{x}-\boldsymbol{x}_n))^T$ and $\mathbf{K} = (\Phi(\boldsymbol{x}_j - \boldsymbol{x}_k))_{jk}$.

The statistical interpretation of the power function is evident. From (2.4) it can be seen that, if $\Psi = \Phi$, the power function is the kriging predictive variance for a Gaussian process with $\sigma^2 = 1$. Clearly, we have $P^2_{\Phi,\mathbf{X}}(\boldsymbol{x}) \le 1$.

To pursue a convergence result under the uniform metric, we define

$$P_{\Phi,\mathbf{X}} := \sup_{\boldsymbol{x}\in\Omega} P_{\Phi,\mathbf{X}}(\boldsymbol{x}). \tag{3.2}$$

We now state the main results on the error bounds for kriging predictors. Recall that the prediction error under the uniform metric is given by (2.7).

The results depend on some smoothness conditions on the imposed kernel. Given any function f, let $\tilde{f}$ be its Fourier transform. According to the inversion formula in Fourier analysis, $\tilde{\Psi}/(2\pi)^d$ is the spectral density of the stationary process Z if Ψ is continuous and integrable on $\mathbf{R}^d$.

Condition 1. The kernels Ψ and Φ are continuous and integrable on $\mathbf{R}^d$, satisfying

$$\|\tilde{\Psi}/\tilde{\Phi}\|_{L_\infty(\mathbf{R}^d)} =: A_1^2 < +\infty. \tag{3.3}$$

In addition, there exists $\alpha \in (0, 1]$, such that

$$\int_{\mathbf{R}^d} \|\boldsymbol{\omega}\|^\alpha \tilde{\Phi}(\boldsymbol{\omega}) d\boldsymbol{\omega} =: A_0 < +\infty. \tag{3.4}$$

Now we are able to state the first main theorem of this paper. Recall that σ^2 is the variance of $Z(\boldsymbol{x})$. The proofs of Theorem 1 and the theorems in Section 5 make extensive use of the scattered data approximation theory and a maximum inequality for Gaussian processes. Detailed discussions of relevant areas are given in, for example, Wendland (2004) and van der Vaart and Wellner (1996), respectively. We also collect the required mathematical tools and results in the supplementary materials.

Theorem 1. Suppose Condition 1 holds, and the design set $\mathbf{X}$ is dense enough in the sense that $P_{\Phi,\mathbf{X}}$ defined in (3.2) is no more than some given constant C. Then for any $u > 0$, with probability at least $1 - 2\exp\{-u^2/(2A_1^2\sigma^2P^2_{\Phi,\mathbf{X}})\}$, the kriging prediction error has the upper bound

$$\sup_{\boldsymbol{x}\in\Omega} |Z(\boldsymbol{x}) - \mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x})| \le K\sigma P_{\Phi,\mathbf{X}} \log^{1/2}(e/P_{\Phi,\mathbf{X}}) + u. \tag{3.5}$$

Here the constants $C, K > 0$ depend only on Ω, α, A_0, and A_1.

Theorem 1 presents an upper bound on the maximum prediction error of kriging. This answers the first question posed in Section 2.3. We will give more explicit error bounds in terms of the design $\mathbf{X}$ and the kernel Φ in Section 3.2. Theorem 1 can also be used to study the case of misspecified correlation functions, provided that condition (3.3) is fulfilled. Condition (3.3) essentially requires that the imposed correlation Φ is *no smoother* than the true correlation function Ψ. Theorem 1 can also be used to address the third question posed in Section 2.3. Note that the right side of (3.5) is a deterministic function depending on the design, and is decreasing in $P_{\Phi,\mathbf{X}}$ if $P_{\Phi,\mathbf{X}}$ is large enough. Therefore, it is reasonable to consider designs which minimize $P_{\Phi,\mathbf{X}}$. Such a construction depends on the specific form of Φ. In Section 3.2, we will further show that, by maximizing certain space-filling measure, one can arrive at the optimal rate of convergence for a broad class of correlation functions.

From Theorem 1, we also observe that the constant A_1 in (3.3) determines the decay rate of the maximum prediction error. In other words, the maximum prediction error appears more concentrated around its mean when the imposed kernel is closer to the true correlation function. Note that condition (3.4) requires a moment condition on the spectral density, which is fulfilled for any Matérn or Gaussian kernel.

The nonasymptotic upper bound in Theorem 1 implies some asymptotic results which are of traditional interests in spatial statistics and related areas. For instance, suppose we adopt a classic setting of fixed-domain asymptotics (Stein 1999) in which the probabilistic structure of $Z(\boldsymbol{x})$ and the kernel function Φ are fixed, and the number of design points increases so that $P_{\Phi,\mathbf{X}}$ tends to zero. Corollary 1 is an immediate consequence of Theorem 1, which shows the weak convergence and L_p convergence of the maximum prediction error.

Corollary 1. For fixed Ψ, Φ, Ω, and σ, we have the following asymptotic results

$$\sup_{\boldsymbol{x}\in\Omega}|Z(\boldsymbol{x})-\mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x})| = O_{\mathbb{P}}(P_{\Phi,\mathbf{X}}\log^{1/2}(1/P_{\Phi,\mathbf{X}})), \quad (3.6)$$

$$\left(\mathbb{E}\left[\sup_{\boldsymbol{x}\in\Omega}|Z(\boldsymbol{x})-\mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x})|^p\right]\right)^{1/p} = O(P_{\Phi,\mathbf{X}}\log^{1/2}(1/P_{\Phi,\mathbf{X}})), \quad (3.7)$$

for any $1 \le p < +\infty$, as $P_{\Phi,\mathbf{X}} \to 0$.

Proof. Theorem 1 implies (3.6) directly. For (3.7), it follows from

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\boldsymbol{x}\in\Omega}|Z(\boldsymbol{x})-\mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x})|^p\right]\\
&= \int_0^\infty \mathbb{P}(\sup_{\boldsymbol{x}\in\Omega}|Z(\boldsymbol{x})-\mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x})| > t^{1/p})dt\\
&= \left(\int_0^{[K\sigma P_{\Phi,\mathbf{X}}\log^{1/2}(e/P_{\Phi,X})]^p} + \int_{[K\sigma P_{\Phi,\mathbf{X}}\log^{1/2}(e/P_{\Phi,X})]^p}^\infty\right)\\
&\quad\times\mathbb{P}(\sup_{\boldsymbol{x}\in\Omega}|Z(\boldsymbol{x})-\mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x})| > t^{1/p})dt\\
&\le \left[K\sigma P_{\Phi,\mathbf{X}}\log^{1/2}(e/P_{\Phi,X})\right]^p\\
&\quad+\int_0^\infty 2\exp\{-t^{2/p}/(2A_1^2\sigma^2P_{\Phi,\mathbf{X}}^2)\}dt\\
&= O(P_{\Phi,\mathbf{X}}^p\log^{p/2}(1/P_{\Phi,\mathbf{X}})),
\end{aligned}$$

where the inequality follows from Theorem 1. □

We believe that (3.6) and (3.7) are the full convergence rate because from (1.3) in the supplementary materials we can see that the convergence rate of the radial basis approximation for deterministic functions in the reproducing kernel Hilbert space is $O(P_{\Phi,\mathbf{X}})$ and these two rates are nearly at the same order of magnitude, expect for a logarithmic factor. This is reasonable because the support of a Gaussian process is typically larger than the corresponding reproducing kernel Hilbert space (van der Vaart and van Zanten 2008). As said earlier in this section, if $\Psi = \Phi$, $P_{\Phi,\mathbf{X}}$ is the supremum of the pointwise predictive SD. Thus, Corollary 1 implies that, if Ψ is known, the predictive error of kriging under the uniform metric is not much larger than its pointwise error.

3.2. Error Bounds in Terms of the Fill Distance

Our next step is to find error bounds which are easier to interpret and compute than that in Theorem 1. To this end, we wish to find an upper bound of $P_{\Phi,\mathbf{X}}$, in which the effects of the design $\mathbf{X}$ and the kernel Φ can be made explicit and separately. This step is generally more complicated, but fortunately some upper bounds are available in the literature, especially for the Gaussian and the Matérn kernels. These bounds are given in terms of the *fill distance*, which is a quantity depending only on the design $\mathbf{X}$. Given the experimental region Ω, the fill distance of a design $\mathbf{X}$ is defined as

$$h_{\mathbf{X}} := \sup_{\boldsymbol{x}\in\Omega}\min_{\boldsymbol{x}_j\in\mathbf{X}}\|\boldsymbol{x}-\boldsymbol{x}_j\|. \quad (3.8)$$

Clearly, the fill distance quantifies the space-filling property (Santner, Williams, and Notz 2003) of a design. A design having the minimum fill distance among all possible designs with the same number of points is known as a minimax distance design (Johnson, Moore, and Ylvisaker 1990).

The upper bounds of $P_{\Phi,\mathbf{X}}$ in terms of the fill distance for Gaussian and Matérn kernels are given in Lemmas 1 and 2, respectively.

Lemma 1 (Wendland 2004, Theorem 11.22). Let $\Omega = [0,1]^d$; $\Phi(x)$ be a Gaussian kernel given by (2.1). Then there exist constants c, h_0 depending only on Ω and the scale parameter ϕ in (2.1), such that $P_{\Phi,\mathbf{X}} \le h_{\mathbf{X}}^{c/h_{\mathbf{X}}}$ provided that $h_{\mathbf{X}} \le h_0$.

Lemma 2 (Wu and Schaback 1993, Theorem 5.14). Let Ω be compact and convex with a positive Lebesgue measure; $\Phi(x)$ be a Matérn kernel given by (2.2) with the smoothness parameter ν. Then there exist constants c, h_0 depending only on Ω, ν and the scale parameter ϕ in (2.2), such that $P_{\Phi,\mathbf{X}} \le ch_{\mathbf{X}}^\nu$ provided that $h_{\mathbf{X}} \le h_0$.

Using the upper bounds of $P_{\Phi,\mathbf{X}}$ given in Lemmas 1 and 2, we can further deduce error bounds of the kriging predictor in terms of the fill distance defined in (3.8). We demonstrate these results in Examples 1–3.

Example 1. Here we assume Φ is a Matérn kernel in (2.2) with smoothness parameter ν. It is known that

$$\begin{aligned}\tilde{\Phi}(\boldsymbol{\omega}) = 2^d\pi^{d/2}\frac{\Gamma(\nu+d/2)}{\Gamma(\nu)}(4\nu\phi^2)^\nu\\ \times(4\nu\phi^2+\|\boldsymbol{\omega}\|^2)^{-(\nu+d/2)},\end{aligned} \quad (3.9)$$

where ϕ is the scale parameter in (2.2). See, for instance, Wendland (2004) and Tuo and Wu (2015). Suppose Ψ is a Matérn correlation function with smoothness ν_0. It can be verified that Condition 1 holds if and only if $0 < \nu \le \nu_0$. Therefore, if $0 < \nu \le \nu_0$, we can invoke Lemma 2 and Theorem 1 to obtain that the kriging predictor converges to the true Gaussian process with a rate at least $O_{\mathbb{P}}(h_{\mathbf{X}}^\nu\log^{1/2}(1/h_{\mathbf{X}}))$ as $h_{\mathbf{X}}$ tends to zero. It can be seen that the rate of convergence is maximized at $\nu = \nu_0$. In other words, if the true smoothness is known *a priori*, one can obtain the greatest rate of convergence.

Example 2. Suppose Φ is the same as in Example 1, and Ψ is a Gaussian correlation function in (2.1), with spectral density (Santner, Williams, and Notz 2003) $\tilde{\Psi}(\boldsymbol{\omega}) =$

Table 1. Comparison between our work and some existing results.

Article/book	Model assumption	Predictor	Design	Type of convergence	Rate of convergence (Matérn kernels)
Present work	Gaussian process with misspecification	Kriging	Scattered points	L_p conv., uniform in x	$h_{\mathbf{X}}^{\nu}(\log(1/h_{\mathbf{X}}))^{1/2}$
Yakowitz and Szidarovszky (1985)	Stochastic process with misspecification	Kriging	Scattered points	Mean square conv., pointwise in x	NA
Stein (1990b)	Stochastic process with misspecification	Kriging	Regular grid points	Mean square conv., pointwise in x	$n^{-\nu/d}$
Buslaev and Seleznjev (1999)	Gaussian process	Best linear approximation	Optimally chosen points in an interval	L_p conv., uniform in x	$n^{-\nu}(\log n)^{1/2}$
Ritter (2000)	Gaussian process	Kriging	Optimally chosen points	Mean square conv., L_2 in x	$n^{-\nu/d}$
Wu and Schaback (1993)	Deterministic function	Kriging	Scattered points	Uniform in x	$h_{\mathbf{X}}^{\nu}$

$(\pi/\phi)^{2/d}\exp\{-\|\boldsymbol{\omega}\|^2/(4\phi)\}$, where ϕ is the scale parameter in (2.1). Then Condition 1 holds for any choice of ν. Then we can invoke Lemma 2 and Theorem 1 to obtain the same rate of convergence as in Example 1.

Example 3. Suppose $\Phi = \Psi$, and Φ is a Gaussian kernel in (2.1). Then we can invoke Lemmas 1 and Theorem 1 to obtain the rate of convergence $O_{\mathbb{P}}(h_{\mathbf{X}}^{c/h_{\mathbf{X}}-1/2}\log^{1/2}(1/h_{\mathbf{X}}))$ for some constant $c > 0$. Note that this rate is faster than the rates obtained in Examples 1–3, because it decays faster than any polynomial of $h_{\mathbf{X}}$. Such a rate is known as a spectral convergence order (Xiu 2010; Wendland 2004).

The upper bounds in Lemmas 1 and 2 explain more explicitly how the choice of designs can affect the prediction performance. Note that in Examples 1–3, the upper bounds are increasing in $h_{\mathbf{X}}$. This suggests that we should consider the designs with a minimum $h_{\mathbf{X}}$ value, which are known as the maximin distance designs (Johnson, Moore, and Ylvisaker 1990). Therefore, our theory shows that the maximin distance designs enjoy nice theoretical guarantees for all Gaussian and Matérn kernels. In contrast with the designs minimizing $P_{\Phi,\mathbf{X}}$ as discussed after Theorem 1, it would be practically beneficial to use the maximin distance designs because they can be constructed without knowing which specific kriging model is to be used.

3.3. Comparison With Some Existing Results

We make some remarks on the relationship between our results and some existing results. In Table 1, we list some related results in the literature concerning the prediction of some underlying function, which is either a realization of a Gaussian process or a determinisitc function in a reproducing kernel Hilbert spaces. It can be seen from Table 1 that only Buslaev and Seleznjev (1999) and Wu and Schaback (1993) address the uniform convergence problem.

Buslaev and Seleznjev (1999) study the rate of convergence of the best linear approximation under an optimally chosen points in an interval. In other words, the predictor is constructed using the best linear combination of the observed data. Thus, this predictor is in general different from the kriging predictor. Also, note that their work is limited to the one-dimensional case where the points are chosen in a specific way. Therefore, this theory does not directly address the question raised in this paper. However, their result, together with our findings in Example 1, does imply an interesting property of kriging. Recall that in Example 1, the rate of convergence for a (known) Matérn correlation is at least $O_{\mathbb{P}}(h_{\mathbf{X}}^{\nu}\log^{1/2}(1/h_{\mathbf{X}}))$. If a space-filling design is used in an interval, then $h_{\mathbf{X}} \sim 1/n$ and the convergence rate is $O_{\mathbb{P}}(n^{-\nu}(\log n)^{1/2})$, which coincides with the best possible rate of convergence given by a linear predictor. Because the kriging predictor is a linear predictor, we can conclude that our uniform upper bound for kriging is sharp in the sense that it captures the actual rate of convergence.

Among the papers listed in Table 1, Wu and Schaback (1993) is the only one comparable to ours, in the sense that they consider a uniform prediction error under a scatter set of design points. They obtain error estimates of the kriging-type interpolants for a deterministic function, known as the radial basis function approximation. Although the mathematical formulations of the interpolants given by kriging and radial basis functions are similar, the two methods are different in their mathematical settings and assumptions. In radial basis function approximation, the underlying function is assumed *fixed*, while kriging utilizes a probabilistic model, driven by a Gaussian random field.

Kriging with misspecified correlation functions is discussed in Yakowitz and Szidarovszky (1985) and Stein (1988, 1990a, 1990b). It has been proven in these papers that some correlation functions, especially the Matérn correlation family, are robust against model misspecification. However, they do not consider convergence under a uniform metric. More discussions on this point are given in Section 6.

4. Simulation Studies

In Example 1, we have shown that if Ψ and Φ are Matérn kernels with smoothness parameters ν_0 and ν, respectively, and $\nu \leq \nu_0$, then the kriging predictor converges with a rate at least $O_{\mathbb{P}}(h_{\mathbf{X}}^{\nu}\log^{1/2}(1/h_{\mathbf{X}}))$. In this section, we report simulation studies that verify that this rate is sharp, that is, the true convergence rate coincides with that given by the theoretical upper bound.

Table 2. Numerical studies on the convergence rates of kriging prediction.

ν_0	ν	Regression coefficient	Theoretical assertion	Relative difference
3	2.5	2.697	2.5	0.0788
5	3.5	3.544	3.5	0.0126
3.5	3.5	3.582	3.5	0.0234
5	5	4.846	5	0.0308

NOTE: The first two columns show the true and imposed smoothness parameters of the Matérn kernels. The third column shows the convergence rate obtained from the simulation. The fourth column shows the convergence rate given by Theorem 1. The last column shows the relative difference between the third and the fourth columns, given by |regression coefficient − theoretical assertion|/(theoretical assertion).

We denote the expectation of the left-hand side of (3.5) by $\mathcal{E}$. If the error bound (3.5) is sharp, we have the approximation

$$\mathcal{E} \approx c h_{\mathbf{X}}^{\nu} \log^{1/2}(1/h_{\mathbf{X}})$$

for some constant c independent of $h_{\mathbf{X}}$. Taking logarithm on both sides of the above formula yields

$$\log \mathcal{E} \approx \nu \log h_{\mathbf{X}} + \frac{1}{2} \log(-\nu \log h_{\mathbf{X}}) + \log c. \tag{4.1}$$

Since $\log(-\nu \log h_{\mathbf{X}})$ is much smaller than $\log h_{\mathbf{X}}$, the effect of $\log(-\nu \log h_{\mathbf{X}})$ is negligible in (4.1). Consequently, we get our second approximation

$$\log \mathcal{E} \approx \nu \log h_{\mathbf{X}} + \log c. \tag{4.2}$$

As shown in (4.2), $\log \mathcal{E}$ is approximately a linear function in $\log h_{\mathbf{X}}$ with slope ν. Therefore, to assess whether (3.5) is sharp, we should verify if the regression coefficient (slope) of $\log \mathcal{E}$ with respect to $\log h_{\mathbf{X}}$ is close to ν.

In our simulation studies, the experimental region is chosen to be $\Omega = [0,1]^2$. To estimate the regression coefficient ν in (4.2), we choose 50 different maximin Latin hypercube designs (Santner, Williams, and Notz 2003) with sample sizes $10k$, for $k = 1, 2, \ldots, 50$. Note that each design corresponds to a specific value of the fill distance $h_{\mathbf{X}}$. For each k, we simulate the Gaussian processes 100 times to reduce the simulation error. For each simulated Gaussian process, we compute $\sup_{\boldsymbol{x} \in \Omega_1} |Z(\boldsymbol{x}) - \mathcal{I}_{\Phi,\mathbf{X}} Z(\boldsymbol{x})|$ to approximate the sup-error $\sup_{\boldsymbol{x} \in \Omega} |Z(\boldsymbol{x}) - \mathcal{I}_{\Phi,\mathbf{X}} Z(\boldsymbol{x})|$, where Ω_1 is the set of grid points with grid length 0.01. This should give a good approximation since the grid is dense enough. Next, we calculate the average of $\sup_{\boldsymbol{x} \in \Omega_1} |Z(\boldsymbol{x}) - \mathcal{I}_{\Phi,\mathbf{X}} Z(\boldsymbol{x})|$ over the 100 simulations to approximate $\mathcal{E}$. Then the regression coefficient is estimated using the least squares method.

We conduct four simulation studies with different choices of the true and imposed smoothness of the Matérn kernels, denoted by ν_0 and ν, respectively. We summarize the simulation results in Table 2.

It is seen in Table 2 that the regression coefficients are close to the values given by our theoretical analysis, with relative error no more than 0.08. By comparing the third and the fourth rows of Table 2, we find that the regression coefficient does not have a significant change when ν remains the same, even if ν_0 changes. On the other hand, the third and the fifth rows show that, the regression coefficient changes significantly as ν changes, even if ν_0 keeps unchanged. This shows convincingly that the convergence rate is independent of the true smoothness of the Gaussian process, and the rate given by Theorem 1 is sharp. Note that our simulation studies justify the use of the leading term $\log h_{\mathbf{X}}$ in (4.1) to assess the convergence rate but they do not cover the second term $\log(-\nu \log h_{\mathbf{X}})$, which is of lower order. Figure 1 in the supplementary materials shows that the regression line for the logarithm of the fill distance and the logarithm of the average prediction error fits the data very well.

From the simulation studies, we can see that if the smoothness of the imposed kernel is lower, the kriging predictor converges slower. Therefore, to maximize the prediction efficiency, it is beneficial to set the smoothness parameter of the imposed kernel the same as the true correlation function.

5. Extensions to Universal Kriging

In this section, we extend the main result in Theorem 1 from simple kriging to universal kriging. As an extension of simple kriging, universal kriging is widely used in practice. Instead of using a zero mean Gaussian process, universal kriging assumes that the Gaussian process has a nonzero mean function, modeled as a linear combination of a set of basis functions with unknown regression coefficients. Specifically, we consider the following model

$$Y(\boldsymbol{x}) = \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} + Z(\boldsymbol{x}), \tag{5.1}$$

where $\boldsymbol{f}(\boldsymbol{x}) := (f_1(\boldsymbol{x}), \ldots, f_p(\boldsymbol{x}))^T$ is a vector of p linearly independent known functions over Ω, $\boldsymbol{\beta}$ is an unknown vector of regression coefficients, and $Z(\boldsymbol{x})$ is a zero mean stationary Gaussian process with correlation function Ψ. The goal of universal kriging is to reconstruct $Y(\boldsymbol{x})$ based on scattered observations $Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_n)$.

A common practice is to use the maximum likelihood estimation (MLE) method to estimate $\boldsymbol{\beta}$ and the best linear unbiased predictor (BLUP) to predict $Y(\boldsymbol{x})$ at an untried $\boldsymbol{x}$. The following facts can be found in the book by Santner, Williams, and Notz (2003). As before, let $\mathbf{K} := (\Psi(\boldsymbol{x}_j - \boldsymbol{x}_k))_{jk}$. Define $\mathbf{F} := (\boldsymbol{f}(\boldsymbol{x}_1), \ldots, \boldsymbol{f}(\boldsymbol{x}_n))^T$. We temporarily suppose that Ψ is known. Then the MLE of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{K}^{-1}\mathbf{Y}, \tag{5.2}$$

with $\mathbf{Y} := (Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_n))^T$. To use (5.2), we should require $n \geq p$ so that $\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F}$ is invertible. The BLUP of $Y(\boldsymbol{x})$ is

$$\begin{aligned}\hat{Y}_{\mathbf{BLUP}}(\boldsymbol{x}) =& [\boldsymbol{f}^T(\boldsymbol{x})(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{K}^{-1} \\ &+ \boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}(\mathbf{I}_n - \mathbf{F}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{K}^{-1})]\mathbf{Y},\end{aligned} \tag{5.3}$$

where $\boldsymbol{r}(\boldsymbol{x}) := (\Psi(\boldsymbol{x} - \boldsymbol{x}_1), \ldots, \Psi(\boldsymbol{x} - \boldsymbol{x}_n))^T$.

As before, we are interested in the situation with a misspecified correlation function, also denoted by Φ. Using (5.3), we can calculate the "BLUP" of $Y(\boldsymbol{x})$ under Φ, denoted by $\hat{Y}_{\mathbf{BLUP},\Phi}(\boldsymbol{x})$, with redefined $\boldsymbol{r}$ and $\mathbf{K}$ given by $\boldsymbol{r} := (\Phi(\boldsymbol{x} - \boldsymbol{x}_1), \ldots, \Phi(\boldsymbol{x} - \boldsymbol{x}_n))^T$ and $\mathbf{K} := (\Phi(\boldsymbol{x}_j - \boldsymbol{x}_k))_{jk}$.

The goal of our theoretical study is to bound $\sup_{\boldsymbol{x} \in \Omega} |Y(\boldsymbol{x}) - \hat{Y}_{\mathbf{BLUP},\Phi}(\boldsymbol{x})|$ in a way similar to Theorem 1. The results are given in Theorem 2. We denote the minimum eigenvalue of a matrix $\mathbf{H}$ by $\lambda_{\min}(\mathbf{H})$. Let $\mathcal{A}$ be the set of $p \times p$ submatrices of $\mathbf{F}$.

Theorem 2. Suppose the conditions of Theorem 1 are fulfilled. In addition, $f_j \in \mathcal{N}_\Phi(\Omega)$ for $j = 1, \dots, p$. Then

$$\mathbb{E}\left[\sup_{\boldsymbol{x}\in\Omega} |Y(\boldsymbol{x}) - \hat{Y}_{\mathbf{BLUP},\Phi}(\boldsymbol{x})|\right] = O(P_{\Phi,\mathbf{X}}[pA + \log^{1/2}(1/P_{\Phi,\mathbf{X}})]), \qquad (5.4)$$

where the asymptotic constant is independent of $\mathbf{X}$, p and f_j's; and

$$A = \left(\sum_{j=1}^{p} \|f_j\|^2_{\mathcal{N}_\Phi(\Omega)} / \max_{\mathbf{F}_p \in \mathcal{A}} \lambda_{\min}(\mathbf{F}_p^T \mathbf{F}_p)\right)^{1/2}.$$

The condition $f_i \in \mathcal{N}_\Phi(\Omega)$ in Theorem 2 is mild if Φ is a Matérn kernel. It is known that in such case, $\mathcal{N}_\Phi(\Omega)$ coincides with a Sobolev space, which contains all smooth functions, such as polynomials. See Corollary 10.13 of Wendland (2004).

Compared to the uniform error bound for simple kriging, (5.4) has an additional term $O(P_{\Phi,\mathbf{X}} pA)$, which is caused by the unknown regression coefficient β. In many situations, pA is bounded above by a constant, for example, when p and f_j's are fixed and $\boldsymbol{x}_j$'s are independent random samples. In this case $O(P_{\Phi,\mathbf{X}} pA) = O(P_{\Phi,\mathbf{X}})$ and can be absorbed by the simple kriging uniform error bound.

As a by-product of our analysis, we show that the MLE $\hat{\boldsymbol{\beta}}$ is *inconsistent* when Ψ is known. The result in Theorem 3 shows that the covariance matrix of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ is no less than the inverse of the inner product matrix of f_j's in the reproducing kernel Hilbert space, in the sense that the former subtracting the latter is positive semidefinite.

Theorem 3. If $\Phi = \Psi$ and $f_j \in \mathcal{N}_\Phi(\Omega)$ for $j = 1, \dots, p$, we have $\mathrm{var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \geq \mathbf{V}^{-1}$, where $\mathbf{V} := (\langle f_j, f_k \rangle_{\mathcal{N}_\Phi(\Omega)})_{jk}$ is positive definite.

The proof of Theorem 3 shows that the estimation of $\boldsymbol{\beta}$ is perturbed by the Gaussian process, and thus becomes inconsistent. In addition, the well-known theory by Ying (1991) and Zhang (2004) suggests that the model parameters in the covariance functions may not have consistent estimators. Therefore, it would be more meaningful to use Gaussian process models for prediction, rather than for parameter identification.

Next, we study the uniform error bound when a *random* kernel function is used. Such a result can be useful when an estimated correlation function is used. The main idea here is to study a more sophisticated type of uniform error, which also takes supremum over a family of correlation functions.

Let $\Phi_{\boldsymbol{\theta}}$ be a family of correlation functions indexed by $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)^T \in \boldsymbol{\Theta}$. Suppose Θ is a compact subregion of $\mathbf{R}^q$. For notational simplicity, denote $P_{\mathbf{X}} = \max_{\boldsymbol{\theta}\in\Theta} P_{\Phi_{\boldsymbol{\theta}},\mathbf{X}}$. Theorem 4 provides a uniform error bound in terms of both untried $\boldsymbol{x}$ and parameter $\boldsymbol{\theta}$.

Condition 2. The kernels $\Phi_{\boldsymbol{\theta}}$ are continuous and integrable on $\mathbf{R}^d$, satisfying

$$\|\tilde{\Psi}/\tilde{\Phi}_{\boldsymbol{\theta}}\|_{L_\infty(\mathbf{R}^d)} =: A_1^2 < +\infty \qquad (5.5)$$

and

$$\int_{\mathbf{R}^d} \|\boldsymbol{\omega}\|^\alpha \tilde{\Phi}_{\boldsymbol{\theta}}(\boldsymbol{\omega}) d\boldsymbol{\omega} =: A_0 < +\infty, \qquad (5.6)$$

for constants $\alpha \in (0, 1], A_0, A_1$ independent of $\boldsymbol{\theta}$.

Theorem 4. Suppose the conditions of Theorem 2 and Condition 2 are fulfilled. Suppose $\Phi_{\boldsymbol{\theta}}(\boldsymbol{x})$ is differentiable in $\boldsymbol{\theta}$ for each $\boldsymbol{x}$. In addition, suppose the differentiation in $\boldsymbol{\theta}$ and the Fourier transform in $\boldsymbol{x}$ are interchangeable, that is,

$$\frac{\widetilde{\partial \Phi_{\boldsymbol{\theta}}}}{\partial \theta_j} = \frac{\partial \widetilde{\Phi_{\boldsymbol{\theta}}}}{\partial \theta_j}, \quad j = 1, \dots, q. \qquad (5.7)$$

Moreover, suppose

$$\sup_{\boldsymbol{x}\in\mathbf{R}^d, \boldsymbol{\theta}\in\Theta} \left|\frac{\partial}{\partial \theta_j} \log \widetilde{\Phi_{\boldsymbol{\theta}}}\right| \leq A_2 < +\infty, \quad j = 1, \dots, q. \qquad (5.8)$$

Then,

$$\mathbb{E}\left[\sup_{\boldsymbol{x}\in\Omega, \boldsymbol{\theta}\in\Theta} |Y(\boldsymbol{x}) - \hat{Y}_{\mathbf{BLUP},\Phi_{\boldsymbol{\theta}}}(\boldsymbol{x})|\right] = O(P_{\mathbf{X}}[pA + \log^{1/2}(1/P_{\mathbf{X}})]), \qquad (5.9)$$

where A is the same as in Theorem 2 and the asymptotic constant is independent of $\mathbf{X}$, p, and f_j's.

The uniform error bound in (5.9) can govern the error bound when a random kernel is used. Specifically, suppose a random kernel, denoted by $\Phi_{\hat{\boldsymbol{\theta}}}$, is used, and the support of the random variable (estimator) $\hat{\boldsymbol{\theta}}$ is Θ. Then we have $\mathbb{E}\sup_{\boldsymbol{x}\in\Omega} |Y(\boldsymbol{x}) - \hat{Y}_{\mathbf{BLUP},\Phi_{\hat{\boldsymbol{\theta}}}}(\boldsymbol{x})| = O(P_{\mathbf{X}}[pA + \log^{1/2}(1/P_{\mathbf{X}})])$.

We now verify the conditions of Theorem 4 for Matérn and Gaussian kernels. Suppose Ψ is a Matérn kernel with smoothness ν_0. Let $\boldsymbol{\theta} = (\phi, \nu)$ and Θ is a compact subregion of $(0, +\infty) \times (0, \nu_0)$. Clearly, Condition 2 is fulfilled. Recall that $\tilde{\Phi}_{\boldsymbol{\theta}}$ is given in (3.9). Dominated convergence theorem ensures (5.7) and we can verify (5.8) via direct calculations. Suppose Ψ is a Gaussian kernel in (2.1) with $\phi = \phi_0$. Suppose $\boldsymbol{\theta} = \phi$ and Θ is a compact subregion of $[\phi_0, +\infty)$. Similar direct calculations can show that the conditions in Theorem 4 are also fulfilled in this case.

The proofs of Theorems 2 and 4 use the techniques we developed for proving Theorem 1. These theorems show that the general rate of convergence $O_P(P_{\Phi,X} \log^{1/2} P_{\Phi,X})$ is still valid even if estimated mean and covariance functions are used.

6. Conclusions and Further Discussion

We first summarize the statistical implications of this work. We prove that the kriging predictive error converges to zero under a uniform metric, which justifies the use of kriging as a function reconstruction tool. Our analysis covers both simple and universal kriging. Kriging with a misspecified correlation function is also studied. Theorem 1 shows that there is a tradeoff between the predictive efficiency and the robustness. Roughly speaking, a less smooth correlation function is more robust against model misspecification. However, the price for robustness is to incur a small loss in prediction efficiency. With the help of the classic results in radial basis function approximation (in Lemmas 1 and 2), we find that the predictive error of kriging is associated with the fill distance, which is a space-filling measure of the design. This justifies the use of space-filling designs for (stationary) kriging models.

We have proved in Theorem 1 that the kriging predictor is consistent if the imposed correlation function is *undersmoothed*, that is, the imposed correlation function is no smoother than the true correlation function. One would ask whether a similar result can be proven for the case of oversmoothed correlation functions. Yakowitz and Szidarovszky (1985) proved that kriging with an oversmoothed Matérn correlation function also achieves (pointwisely) predictive consistency. In light of this result, we may consider extensions of Theorem 1 to the oversmoothed case in a future work.

In a series of papers, Stein (1988, 1990a, 1990b, 1993) investigated the asymptotic efficiency of the kriging predictor. The theory in our work does not give assertions about prediction efficiency, although we provide explicit error bounds for kriging predictors with scattered design points in general dimensions. Another possible extension of this work is to consider the impact of a misspecified mean function. Jiang, Nguyen, and Rao (2011) address this problem in the context of small area estimation.

Appendix A: Proof of Theorem 1

Because $\mathcal{I}_{\Phi,\mathbf{X}}$ is a linear map between two functions, $\mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x})$ is also a Gaussian process. Therefore, the problem in (2.7) is to bound the maximum value of a Gaussian process. The main idea of the proof is to invoke a maximum inequality for Gaussian processes, which states that the supremum of a Gaussian process is no more than a multiple of the integral of the covering number with respect to the natural distance $\mathfrak{d}$. The details are given in the supplementary materials. Also see Adler and Taylor (2009) and van der Vaart and Wellner (1996) for related discussions.

Without loss of generality, assume $\sigma = 1$, because otherwise we can consider the upper bound of $\sup_{\boldsymbol{x}\in\Omega} |Z(\boldsymbol{x}) - \mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x})|/\sigma$ instead. Let $g(\boldsymbol{x}) = Z(\boldsymbol{x}) - \mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x})$. For any $\boldsymbol{x}, \boldsymbol{x}' \in \Omega$,

$$
\begin{aligned}
\mathfrak{d}(\boldsymbol{x},\boldsymbol{x}')^2 &= \mathbb{E}(g(\boldsymbol{x}) - g(\boldsymbol{x}'))^2 \\
&= \mathbb{E}(Z(\boldsymbol{x}) - \mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x}) - (Z(\boldsymbol{x}') - \mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x}')))^2 \\
&= \Psi(\boldsymbol{x}-\boldsymbol{x}) - 2\boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\boldsymbol{r}_1(\boldsymbol{x}) + \boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\mathbf{K}_1\mathbf{K}^{-1}\boldsymbol{r}(\boldsymbol{x}) \\
&\quad +\Psi(\boldsymbol{x}'-\boldsymbol{x}') - 2\boldsymbol{r}^T(\boldsymbol{x}')\mathbf{K}^{-1}\boldsymbol{r}_1(\boldsymbol{x}') \\
&\quad +\boldsymbol{r}^T(\boldsymbol{x}')\mathbf{K}^{-1}\mathbf{K}_1\mathbf{K}^{-1}\boldsymbol{r}(\boldsymbol{x}') \\
&\quad -2[\Psi(\boldsymbol{x}-\boldsymbol{x}') - \boldsymbol{r}^T(\boldsymbol{x}')\mathbf{K}^{-1}\boldsymbol{r}_1(\boldsymbol{x}) - \boldsymbol{r}_1^T(\boldsymbol{x}')\mathbf{K}^{-1}\boldsymbol{r}(\boldsymbol{x}) \\
&\quad +\boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\mathbf{K}_1\mathbf{K}^{-1}\boldsymbol{r}(\boldsymbol{x}')],
\end{aligned}
$$

where $\boldsymbol{r}_1(\cdot) = (\Psi(\cdot - \boldsymbol{x}_1), \dots, \Psi(\cdot - \boldsymbol{x}_n))^T$, $\boldsymbol{r}(\cdot) = (\Phi(\cdot - \boldsymbol{x}_1), \dots, \Phi(\cdot - \boldsymbol{x}_n))^T$, $\mathbf{K}_1 = (\Psi(\boldsymbol{x}_j - \boldsymbol{x}_k))_{jk}$, and $\mathbf{K} = (\Phi(\boldsymbol{x}_j - \boldsymbol{x}_k))_{jk}$.

The rest of our proof consists of the following steps. In step 1, we bound the covering number $N(\epsilon, \Omega, \mathfrak{d})$. Next we bound the diameter D. In step 3, we invoke Lemma 1 in the supplementary materials to obtain a bound for the entropy integral. In the last step, we use (1.8) in the supplementary materials to obtain the desired results.

Step 1: Bounding the covering number

Let $h(\cdot) = \Psi(\boldsymbol{x} - \cdot) - \Psi(\boldsymbol{x}' - \cdot)$ and $h_1(\cdot) = \boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\boldsymbol{r}_1(\cdot) - \boldsymbol{r}^T(\boldsymbol{x}')\mathbf{K}^{-1}\boldsymbol{r}_1(\cdot)$. It can verified that

$$
\begin{aligned}
\mathfrak{d}(\boldsymbol{x},\boldsymbol{x}')^2 = &- [h(\boldsymbol{x}') - \mathcal{I}_{\Phi,\mathbf{X}}h(\boldsymbol{x}')] + [h(\boldsymbol{x}) - \mathcal{I}_{\Phi,\mathbf{X}}h(\boldsymbol{x})] \\
&+ [h_1(\boldsymbol{x}') - \mathcal{I}_{\Phi,\mathbf{X}}h_1(\boldsymbol{x}')] - [h_1(\boldsymbol{x}) - \mathcal{I}_{\Phi,\mathbf{X}}h_1(\boldsymbol{x})].
\end{aligned}
$$

By Condition 1, $h \in \mathcal{N}_\Phi(\mathbf{R}^d)$, since $\Psi(\boldsymbol{x} - \cdot) \in \mathcal{N}_\Phi(\mathbf{R}^d)$ for any $\boldsymbol{x} \in \Omega$. Thus, by (1.3) in the supplementary materials,

$$
\mathfrak{d}(\boldsymbol{x},\boldsymbol{x}')^2 \le 2P_{\Phi,\mathbf{X}}(\|h\|_{\mathcal{N}_\Phi(\mathbf{R}^d)} + \|h_1\|_{\mathcal{N}_\Phi(\mathbf{R}^d)}). \tag{A.1}
$$

By Theorem 1 in the supplementary materials,

$$
\|h\|^2_{\mathcal{N}_\Phi(\mathbf{R}^d)} = (2\pi)^{-d} \int_{\mathbf{R}^d} \frac{|\tilde{h}(\boldsymbol{\omega})|^2}{\tilde{\Phi}(\boldsymbol{\omega})} d\boldsymbol{\omega}. \tag{A.2}
$$

Using Condition 1 and (A.2), we obtain

$$
\begin{aligned}
\|h\|^2_{\mathcal{N}_\Phi(\mathbf{R}^d)} &= (2\pi)^{-d} \int_{\mathbf{R}^d} \frac{|\tilde{h}(\boldsymbol{\omega})|^2}{\tilde{\Phi}(\boldsymbol{\omega})} d\boldsymbol{\omega} \\
&\le A_1^2 (2\pi)^{-d} \int_{\mathbf{R}^d} \frac{|\tilde{h}(\boldsymbol{\omega})|^2}{\tilde{\Psi}(\boldsymbol{\omega})} d\boldsymbol{\omega} \\
&= A_1^2 \|h\|^2_{\mathcal{N}_\Psi(\mathbf{R}^d)}.
\end{aligned} \tag{A.3}
$$

We need the following inequality to bound $\|h\|^2_{\mathcal{N}_\Psi(\mathbf{R}^d)}$. For any $0 < \beta \le 1$ and $x \in \mathbf{R}$, we have

$$
|1 - \cos x| \le 2|x|^\beta. \tag{A.4}
$$

This inequality is trivial when $|x| \ge 1$ because $|1 - \cos x| \le 2$; and for the case that $|x| < 1$, (A.4) can be proven using the mean value theorem and the fact that $|x| \le |x|^\beta$. Note that the definition of h implies that $\|h\|^2_{\mathcal{N}_\Psi(\mathbf{R}^d)} = \Psi(\boldsymbol{x}-\boldsymbol{x}) - 2\Psi(\boldsymbol{x}'-\boldsymbol{x}) + \Psi(\boldsymbol{x}'-\boldsymbol{x}')$. Thus, by the Fourier inversion theorem and (A.4), we have

$$
\begin{aligned}
\|h\|^2_{\mathcal{N}_\Psi(\mathbf{R}^d)} &= \Psi(\boldsymbol{x}-\boldsymbol{x}) - 2\Psi(\boldsymbol{x}'-\boldsymbol{x}) + \Psi(\boldsymbol{x}'-\boldsymbol{x}') \\
&= 2(2\pi)^{-d} \int_{\mathbf{R}^d} (1 - e^{i(\boldsymbol{x}-\boldsymbol{x}')^T\boldsymbol{\omega}}) \tilde{\Psi}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\
&\le \left(4(2\pi)^{-d} \int_{\mathbf{R}^d} \|\omega\|^\beta \tilde{\Psi}(\boldsymbol{\omega}) d\boldsymbol{\omega}\right) \|\boldsymbol{x}-\boldsymbol{x}'\|^\beta && \text{(A.5)} \\
&=: C_1 \|\boldsymbol{x}-\boldsymbol{x}'\|^\beta, && \text{(A.6)}
\end{aligned}
$$

for any $0 < \beta \le \alpha$. In particular, we now choose $\beta = \alpha/2$. Now we consider $h_1(\cdot)$. It follows from a similar argument that $\|h_1\|^2_{\mathcal{N}_\Phi(\mathbf{R}^d)} \le A_1^2\|h_1\|^2_{\mathcal{N}_\Psi(\mathbf{R}^d)}$. The definition of h_1 implies $\|h_1\|^2_{\mathcal{N}_\Psi(\mathbf{R}^d)} = (\boldsymbol{r}(\boldsymbol{x}') - \boldsymbol{r}(\boldsymbol{x}))^T\mathbf{K}^{-1}\mathbf{K}_1\mathbf{K}^{-1}(\boldsymbol{r}(\boldsymbol{x}') - \boldsymbol{r}(\boldsymbol{x}))$.

For any $\boldsymbol{u} = (u_1, \dots, u_n)^T$, the Fourier inversion theorem and Condition 1 yield

$$
\begin{aligned}
&\sum_{j,k=1}^n u_j\bar{u}_k\Psi(\boldsymbol{x}_j - \boldsymbol{x}_k) \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \sum_{j,k=1}^n u_j\bar{u}_k e^{i(\boldsymbol{x}_j-\boldsymbol{x}_k)^T\boldsymbol{\omega}} \tilde{\Psi}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left|\sum_{j=1}^n u_j e^{i\boldsymbol{x}_j^T\boldsymbol{\omega}}\right|^2 \tilde{\Psi}(\boldsymbol{\omega}) d\boldsymbol{\omega} && \text{(A.7)} \\
&\le \frac{A_1^2}{(2\pi)^d} \int_{\mathbb{R}^d} \left|\sum_{j=1}^n u_j e^{i\boldsymbol{x}_j^T\boldsymbol{\omega}}\right|^2 \tilde{\Phi}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\
&= A_1^2 \sum_{j,k=1}^n u_j\bar{u}_k\Phi(\boldsymbol{x}_j - \boldsymbol{x}_k).
\end{aligned}
$$

Then we choose $\boldsymbol{u} = \mathbf{K}^{-1}(\boldsymbol{r}(\boldsymbol{x}') - \boldsymbol{r}(\boldsymbol{x}))$ to get

$$
\|h_1\|^2_{\mathcal{N}_\Psi(\mathbf{R}^d)} \le A_1^2(\boldsymbol{r}(\boldsymbol{x}') - \boldsymbol{r}(\boldsymbol{x}))^T\mathbf{K}^{-1}(\boldsymbol{r}(\boldsymbol{x}') - \boldsymbol{r}(\boldsymbol{x})). \tag{A.8}
$$

Let $h_2(\cdot) = \Phi(\cdot - \boldsymbol{x}') - \Phi(\cdot - \boldsymbol{x})$. Then $\mathcal{I}_{\Phi,\mathbf{X}}h_2(\cdot) = \boldsymbol{r}^T(\cdot)\mathbf{K}^{-1}(\boldsymbol{r}(\boldsymbol{x}') - \boldsymbol{r}(\boldsymbol{x}))$. By (1.3) in the supplementary materials and the fact that $\|h_2\|^2_{\mathcal{N}_\Phi(\mathbf{R}^d)} = \Phi(\boldsymbol{x}-\boldsymbol{x}) - 2\Phi(\boldsymbol{x}'-\boldsymbol{x}) + \Phi(\boldsymbol{x}'-\boldsymbol{x}')$, we have

$$\begin{aligned}
&(\boldsymbol{r}(\boldsymbol{x}') - \boldsymbol{r}(\boldsymbol{x}))^T\mathbf{K}^{-1}(\boldsymbol{r}(\boldsymbol{x}') - \boldsymbol{r}(\boldsymbol{x}))\\
&\quad\le |h_2(\boldsymbol{x}') - \mathcal{I}_{\Phi,\mathbf{X}}h_2(\boldsymbol{x}')| + |h_2(\boldsymbol{x}) - \mathcal{I}_{\Phi,\mathbf{X}}h_2(\boldsymbol{x})|\\
&\qquad + |h_2(\boldsymbol{x}')| + |h_2(\boldsymbol{x})|\\
&\quad\le 2P_{\Phi,\mathbf{X}}\sqrt{\Phi(\boldsymbol{x}-\boldsymbol{x}) - 2\Phi(\boldsymbol{x}'-\boldsymbol{x}) + \Phi(\boldsymbol{x}'-\boldsymbol{x}')}\\
&\qquad + 2(\Phi(\boldsymbol{x}-\boldsymbol{x}) - 2\Phi(\boldsymbol{x}'-\boldsymbol{x}) + \Phi(\boldsymbol{x}'-\boldsymbol{x}'))\\
&\quad\le 2(P_{\Phi,\mathbf{X}} + \sqrt{\Phi(\boldsymbol{x}-\boldsymbol{x}) - 2\Phi(\boldsymbol{x}'-\boldsymbol{x}) + \Phi(\boldsymbol{x}'-\boldsymbol{x}')})\\
&\qquad \times \sqrt{\Phi(\boldsymbol{x}-\boldsymbol{x}) - 2\Phi(\boldsymbol{x}'-\boldsymbol{x}) + \Phi(\boldsymbol{x}'-\boldsymbol{x}')}\\
&\quad\le 2(P_{\Phi,\mathbf{X}} + 2)\sqrt{\Phi(\boldsymbol{x}-\boldsymbol{x}) - 2\Phi(\boldsymbol{x}'-\boldsymbol{x}) + \Phi(\boldsymbol{x}'-\boldsymbol{x}')}. \qquad \text{(A.9)}
\end{aligned}$$

Thus, if $P_{\Phi,\mathbf{X}} < 1$, by a similar argument in (A.5) (with the choice $\beta = \alpha$), and together with (A.8) and (A.9), we have

$$\|h_1\|^2_{\mathcal{N}_\Psi(\mathbf{R}^d)} \le C_2\|\boldsymbol{x}-\boldsymbol{x}'\|^{\alpha/2}, \qquad \text{(A.10)}$$

for a constant C_2.

In view of (A.1), (A.6), and (A.10), there exists a constant C_3 such that

$$\mathfrak{d}(\boldsymbol{x},\boldsymbol{x}')^2 \le C_3P_{\Phi,\mathbf{X}}\|\boldsymbol{x}-\boldsymbol{x}'\|^{\alpha/4}. \qquad \text{(A.11)}$$

Therefore, the covering number is bounded above by

$$\log N(\epsilon,\Omega,\mathfrak{d}) \le \log N\left(\frac{\epsilon^{8/\alpha}}{C_3^{4/\alpha}P_{\Phi,\mathbf{X}}^{4/\alpha}},\Omega,\|\cdot\|\right). \qquad \text{(A.12)}$$

The right side of (A.12) involves the covering number of a Euclidean ball, which is well understood in the literature. See Lemma 4.1 of Pollard (1990). This result leads to the bound

$$\begin{aligned}
\log N(\epsilon,\Omega,\mathfrak{d}) \;&\le\; C_{4,0}\log\left(\frac{C_{5,0}C_3^{4/\alpha}P_{\Phi,\mathbf{X}}^{4/\alpha}}{\epsilon^{8/\alpha}}\right)\\
&=:\; C_4\log\left(\frac{C_5P_{\Phi,\mathbf{X}}^{1/2}}{\epsilon}\right), \qquad \text{(A.13)}
\end{aligned}$$

provided that

$$\epsilon < C_5P_{\Phi,\mathbf{X}}^{1/2}, \qquad \text{(A.14)}$$

where $C_{4,0}$ and $C_{5,0}$ are constants depending on the dimension and the Eucliden diameter of Ω.

Step 2: Bounding the diameter D

Recall that the diameter is defined by $D = \sup_{\boldsymbol{x},\boldsymbol{x}'\in\Omega}\mathfrak{d}(\boldsymbol{x},\boldsymbol{x}')$. For any $\boldsymbol{x},\boldsymbol{x}'\in\Omega$,

$$\begin{aligned}
\mathfrak{d}(\boldsymbol{x},\boldsymbol{x}')^2 &= \mathbb{E}(g(\boldsymbol{x}) - g(\boldsymbol{x}'))^2 \le 4\sup_{\boldsymbol{x}\in\Omega}\mathbb{E}(g(\boldsymbol{x}))^2\\
&= 4\sup_{\boldsymbol{x}\in\Omega}\mathbb{E}(Z(\boldsymbol{x}) - \mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x}))^2\\
&= 4\sup_{\boldsymbol{x}\in\Omega}(\Psi(\boldsymbol{x}-\boldsymbol{x}) - 2\boldsymbol{r}_1^T(\boldsymbol{x})\mathbf{K}^{-1}r(\boldsymbol{x})\\
&\quad +\boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\mathbf{K}_1\mathbf{K}^{-1}\boldsymbol{r}(\boldsymbol{x})), \qquad \text{(A.15)}
\end{aligned}$$

where $\boldsymbol{r}$, $\boldsymbol{r}_1$, $\mathbf{K}$, and $\mathbf{K}_1$ are defined in the beginning of Appendix A.

Combining identity (A.7) with

$$\Psi(\boldsymbol{x}_j - \boldsymbol{x}) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}e^{i(\boldsymbol{x}-\boldsymbol{x}_j)^T\boldsymbol{\omega}}\tilde{\Psi}(\boldsymbol{\omega})d\boldsymbol{\omega},$$

for any $\boldsymbol{u} = (u_1,\ldots,u_n)$, under Condition 1, we have

$$\begin{aligned}
&\boldsymbol{u}^T\mathbf{K}_1\boldsymbol{u} - 2\boldsymbol{u}^T\boldsymbol{r}_1(\boldsymbol{x}) + \Psi(\boldsymbol{x}-\boldsymbol{x})\\
&=\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left|\sum_{j=1}^n u_je^{i\boldsymbol{x}_j^T\boldsymbol{\omega}} - e^{i\boldsymbol{x}^T\boldsymbol{\omega}}\right|^2\tilde{\Psi}(\boldsymbol{\omega})d\boldsymbol{\omega}\\
&\le\frac{A_1^2}{(2\pi)^d}\int_{\mathbb{R}^d}\left|\sum_{j=1}^n u_je^{i\boldsymbol{x}_j^T\boldsymbol{\omega}} - e^{i\boldsymbol{x}^T\boldsymbol{\omega}}\right|^2\tilde{\Phi}(\boldsymbol{\omega})d\boldsymbol{\omega}\\
&=A_1^2(\boldsymbol{u}^T\mathbf{K}\boldsymbol{u} - 2\boldsymbol{u}^T\boldsymbol{r}(\boldsymbol{x}) + \Phi(\boldsymbol{x}-\boldsymbol{x})). \qquad \text{(A.16)}
\end{aligned}$$

We can combine (A.16) with (A.15) by substituting $\boldsymbol{u}$ in (A.16) by $\mathbf{K}^{-1}\boldsymbol{r}(\boldsymbol{x})$ and arrive at

$$\mathfrak{d}(\boldsymbol{x},\boldsymbol{x}')^2 \le 4A_1^2\sup_{\boldsymbol{x}\in\Omega}(\Phi(\boldsymbol{x}-\boldsymbol{x}) - \boldsymbol{r}(\boldsymbol{x})\mathbf{K}^{-1}\boldsymbol{r}(\boldsymbol{x})).$$

Note that the upper bound of $\Phi(\boldsymbol{x}-\boldsymbol{x}) - \boldsymbol{r}(\boldsymbol{x})\mathbf{K}^{-1}\boldsymbol{r}(\boldsymbol{x})$ is $P^2_{\Phi,\mathbf{X}}$, which implies $\mathfrak{d}(\boldsymbol{x},\boldsymbol{x}')^2 \le 4A_1^2P^2_{\Phi,\mathbf{X}}$. Thus, we conclude that

$$D \le 2A_1P_{\Phi,\mathbf{X}}. \qquad \text{(A.17)}$$

Step 3: Bounding the entropy integral

Under Condition 1, if

$$P_{\Phi,\mathbf{X}} < C_5^2/A_1^2 := C,$$

(A.14) is satisfied for all $\epsilon\in[0,D/2]$. Thus, by (A.13) and (A.17),

$$\begin{aligned}
&\int_0^{D/2}\sqrt{\log N(\epsilon,\Omega,\mathfrak{d})}d\epsilon\\
&\le\int_0^{A_1P_{\Phi,\mathbf{X}}}\sqrt{C_4\log\left(\frac{C_5P_{\Phi,\mathbf{X}}^{1/2}}{\epsilon}\right)}d\epsilon\\
&\le\left(\int_0^{A_1P_{\Phi,\mathbf{X}}}d\epsilon\right)^{1/2}\left(\int_0^{A_1P_{\Phi,\mathbf{X}}}C_4\log\left(\frac{C_5P_{\Phi,\mathbf{X}}}{\epsilon}d\epsilon\right)\right)^{1/2}\\
&\qquad \text{(A.18)}\\
&=C_4^{1/2}A_1P_{\Phi,\mathbf{X}}\sqrt{\log\left(\frac{C_5e}{A_1P_{\Phi,\mathbf{X}}^{1/2}}\right)}. \qquad \text{(A.19)}
\end{aligned}$$

Because $P_{\Phi,\mathbf{X}} \le 1$, the quantity inside the logarithm can be replaced by $e/P_{\Phi,\mathbf{X}}$ at the cost of (possibly) increasing the constant C_4.

Step 4: Bounding $\mathbb{P}(\sup_{\boldsymbol{x}\in\Omega}|Z(\boldsymbol{x}) - \mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x})| > K\int_0^{D/2}\sqrt{\log N(\epsilon,T,\mathfrak{d})}d\epsilon + u)$

Noting that $\sup_{\boldsymbol{x}\in\Omega}\mathbb{E}(Z(\boldsymbol{x}) - \mathcal{I}_{\Phi,\mathbf{X}}Z(\boldsymbol{x}))^2 = D^2$, by plugging (A.17) into (1.8) in the supplementary materials, we obtain the desired inequality, which completes the proof.

Appendix B: Proof of Theorem 2

Denote $\mathbf{Z} = (Z(\boldsymbol{x}_1),\ldots,Z(\boldsymbol{x}_n))^T$. Direct calculations show that

$$\begin{aligned}
&Y(\boldsymbol{x}) - \hat{Y}_{\mathbf{BLUP},\Phi}(\boldsymbol{x})\\
&=\underbrace{Z(\boldsymbol{x}) - \boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\mathbf{Z}}_{I_1} - \underbrace{(\boldsymbol{f}^T(\boldsymbol{x}) - \boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\mathbf{F})(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{K}^{-1}\mathbf{Z}}_{I_2}.
\end{aligned} \qquad \text{(B.1)}$$

Thus, $\sup_{\boldsymbol{x}\in\Omega}|Y(\boldsymbol{x}) - \hat{Y}_{\mathbf{BLUP},\Phi}(\boldsymbol{x})| \le \sup_{\boldsymbol{x}\in\Omega}|I_1| + \sup_{\boldsymbol{x}\in\Omega}|I_2|$. Clearly, $\sup_{\boldsymbol{x}\in\Omega}|I_1|$ is the uniform error of simple kriging, which is studied in Section 3. Corollary 1 suggests that $\mathbb{E}\sup_{\boldsymbol{x}\in\Omega}|I_1| = O(P_{\Phi,\mathbf{X}}\log^{1/2}(1/P_{\Phi,\mathbf{X}}))$.

Now we turn to I_2. By Cauchy–Schwarz inequality,

$$\begin{aligned}|I_2| =&|(\boldsymbol{f}^T(\boldsymbol{x}) - \boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\mathbf{F})(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{K}^{-1}\mathbf{Z}| \\ &\leq \left\{(\boldsymbol{f}^T(\boldsymbol{x}) - \boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\mathbf{F})(\boldsymbol{f}^T(\boldsymbol{x}) - \boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\mathbf{F})^T\right\}^{1/2} \\ &\cdot \left\{\mathbf{Z}^T\mathbf{K}^{-1}\mathbf{F}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{K}^{-1}\mathbf{Z}\right\}^{1/2}. \qquad \text{(B.2)}\end{aligned}$$

Note that the right-hand side of (B.2) is the product of a deterministic function and a random variable independent of $\boldsymbol{x}$.

Clearly, the jth entry of $\boldsymbol{f}^T(\boldsymbol{x}) - \boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\mathbf{F}$ is $f_j(\boldsymbol{x}) - \mathcal{I}_{\Phi,X}f_j(\boldsymbol{x})$, whose absolute value is bounded above by $P_{\Phi,\mathbf{X}}\|f_j\|_{\mathcal{N}_\Phi(\Omega)}$. See Theorem 11.4 of Wendland (2004), also see (1.3) in the supplementary materials. Therefore,

$$(\boldsymbol{f}^T(\boldsymbol{x}) - \boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\mathbf{F})(\boldsymbol{f}^T(\boldsymbol{x}) - \boldsymbol{r}^T(\boldsymbol{x})\mathbf{K}^{-1}\mathbf{F})^T \leq P^2_{\Phi,\mathbf{X}}\sum_{j=j}^{p}\|f_j\|^2_{\mathcal{N}_\Phi(\Omega)}.$$

Our final goal is to bound

$$\begin{aligned}&\left(\mathbb{E}\left\{\mathbf{Z}^T\mathbf{K}^{-1}\mathbf{F}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{K}^{-1}\mathbf{Z}\right\}^{1/2}\right)^2 \\ &\leq \mathbb{E}\mathbf{Z}^T\mathbf{K}^{-1}\mathbf{F}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{K}^{-1}\mathbf{Z} \\ &= \mathbb{E}\mathbf{Tr}[(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{K}^{-1}\mathbf{Z}\mathbf{Z}^T\mathbf{K}^{-1}\mathbf{F}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}] \\ &= \mathbf{Tr}[(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{K}^{-1}\mathbf{K}_1\mathbf{K}^{-1}\mathbf{F}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}],\end{aligned}$$

where $\mathbf{K}_1 = (\Psi(\boldsymbol{x}_j - \boldsymbol{x}_k))_{jk}$ is the true correlation. Via the treatment used in (A.7)–(A.8), it can be shown that

$$\boldsymbol{\alpha}^T\mathbf{K}^{-1}\mathbf{K}_1\mathbf{K}^{-1}\boldsymbol{\alpha} \leq C\boldsymbol{\alpha}^T\mathbf{K}^{-1}\boldsymbol{\alpha}, \qquad \text{(B.3)}$$

for any $\boldsymbol{\alpha}$ and a constant C depending only on $\|\tilde{\Psi}/\tilde{\Phi}\|_{L_\infty(\mathbf{R}^d)}$, which implies

$$\begin{aligned}&\mathbf{Tr}[(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{K}^{-1}\mathbf{K}_1\mathbf{K}^{-1}\mathbf{F}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}] \\ \leq& C\mathbf{Tr}[(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}] \leq Cp/\lambda_{\min}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F}).\end{aligned}$$

For $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$, we have

$$\begin{aligned}\lambda_{\min}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F}) &= \min_{\|\boldsymbol{\alpha}\|=1}\boldsymbol{\alpha}^T\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F}\boldsymbol{\alpha} \\ &= \min_{\|\boldsymbol{\alpha}\|=1}\left\|\mathcal{I}_{\Phi,\mathbf{X}}\sum_{j=1}^{p}\alpha_j f_j(\boldsymbol{x})\right\|^2_{\mathcal{N}_\Phi(\Omega)}.\end{aligned}$$

Now take a p-point subset of $\mathbf{X}$, denoted by $\mathbf{X}_p = \{\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_p\}$. Define $\mathbf{K}_p = (\Phi(\boldsymbol{x}'_j - \boldsymbol{x}'_k))_{jk}$ and $\mathbf{F}_p = (\boldsymbol{f}(\boldsymbol{x}'_1), \ldots, \boldsymbol{f}(\boldsymbol{x}'_p))$. Then by (1.5) in the supplementary materials, we have

$$\begin{aligned}&\lambda_{\min}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F}) \geq \min_{\|\boldsymbol{\alpha}\|=1}\left\|\mathcal{I}_{\Phi,\mathbf{X}_p}\sum_{j=1}^{p}\alpha_j f_j(\boldsymbol{x})\right\|^2_{\mathcal{N}_\Phi(\Omega)} \\ &= \min_{\|\boldsymbol{\alpha}\|=1}\boldsymbol{\alpha}^T\mathbf{F}_p^T\mathbf{K}_p^{-1}\mathbf{F}_p\boldsymbol{\alpha} \geq \min_{\boldsymbol{\alpha}\neq 0}\frac{\boldsymbol{\alpha}^T\mathbf{F}_p^T\mathbf{K}_p^{-1}\mathbf{F}_p\boldsymbol{\alpha}}{\alpha^T\mathbf{F}_p^T\mathbf{F}_p\boldsymbol{\alpha}}\min_{\boldsymbol{\alpha}\neq 0}\frac{\alpha^T\mathbf{F}_p^T\mathbf{F}_p\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\boldsymbol{\alpha}} \\ &\geq \lambda_{\min}(\mathbf{K}_p^{-1})\lambda_{\min}(\mathbf{F}_p^T\mathbf{F}_p) \geq \lambda_{\min}(\mathbf{F}_p^T\mathbf{F}_p)/\mathbf{Tr}(\mathbf{K}_p) = \lambda_{\min}(\mathbf{F}_p^T\mathbf{F}_p)/p. \qquad \text{(B.4)}\end{aligned}$$

Because $\mathbf{X}'_p$ can be chosen as an arbitrary p-point subset, the right-hand side of (B.4) can be replaced by the maximum value over all possible choices of $\mathbf{F}_p$, which completes the proof.

Appendix C: Proof of Theorem 3

We use the notation in the proof of Theorem 2. It is easily verified that $\hat{\beta} - \beta = (\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{K}^{-1}\mathbf{Z}$. Because $\Phi = \Psi$, $\text{var}(Z) = \mathbf{K}$. Therefore,

$$\text{var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}.$$

Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$ be an arbitrary vector. Then

$$\begin{aligned}\boldsymbol{\alpha}^T\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F}\boldsymbol{\alpha} &= \left\|\mathcal{I}_{\Phi,\mathbf{X}}\sum_{j=1}^{p}\alpha_j f_j\right\|^2_{\mathcal{N}_\Phi(\Omega)} \leq \left\|\sum_{j=1}^{p}\alpha_j f_j\right\|^2_{\mathcal{N}_\Phi(\Omega)} \\ &= \boldsymbol{\alpha}^T\mathbf{V}\boldsymbol{\alpha},\end{aligned}$$

where the inequality follows from Corollary 10.25 of Wendland (2004); also see (1.4) in the supplementary materials. Clearly, $\mathbf{V}$ is positive definite, because f_j's are linearly independent. Then then desired result follows from the fact that if $\mathbf{A} \leq \mathbf{B}$ then $\mathbf{A}^{-1} \geq \mathbf{B}^{-1}$.

Appendix D: Proof of Theorem 4

First we prove the simple kriging version. Similar to the proof of Theorem 1, we examine the distance defined by

$$\begin{aligned}\mathfrak{d}^2((\boldsymbol{x}_1, \boldsymbol{\theta}_1), (\boldsymbol{x}_2, \boldsymbol{\theta}_2)) = \mathbb{E}\Big[&Z(\boldsymbol{x}_1) - \mathcal{I}_{\Phi_{\boldsymbol{\theta}_1},\mathbf{X}}Z(\boldsymbol{x}_1) - Z(\boldsymbol{x}_2) \\ &+ \mathcal{I}_{\Phi_{\boldsymbol{\theta}_2},\mathbf{X}}Z(\boldsymbol{x}_2)\Big]^2.\end{aligned}$$

It follows from a similar argument as in Theorem 1 that the diameter of $\mathfrak{d}$ is no more than a multiple of $P_{\mathbf{X}}$. It remains to study the cover number given by $\mathfrak{d}$. First we can separate the effect of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ using the following inequality

$$\begin{aligned}&\mathfrak{d}^2((\boldsymbol{x}_1, \boldsymbol{\theta}_1), (\boldsymbol{x}_2, \boldsymbol{\theta}_2)) \\ &\leq 2\mathbb{E}\left[Z(\boldsymbol{x}_1) - \mathcal{I}_{\Phi_{\boldsymbol{\theta}_1},\mathbf{X}}Z(\boldsymbol{x}_1) - Z(\boldsymbol{x}_2) + \mathcal{I}_{\Phi_{\boldsymbol{\theta}_1},\mathbf{X}}Z(\boldsymbol{x}_2)\right]^2 \\ &+ 2\mathbb{E}\left[\mathcal{I}_{\Phi_{\boldsymbol{\theta}_1},\mathbf{X}}Z(\boldsymbol{x}_2) - \mathcal{I}_{\Phi_{\boldsymbol{\theta}_2},\mathbf{X}}Z(\boldsymbol{x}_2)\right]^2. \qquad \text{(D.1)}\end{aligned}$$

The first term in (D.1) is studied in the proof of Theorem 1. It suffices to show that

$$\mathbb{E}\left[\mathcal{I}_{\Phi_{\boldsymbol{\theta}_1},\mathbf{X}}Z(\boldsymbol{x}) - \mathcal{I}_{\Phi_{\boldsymbol{\theta}_2},\mathbf{X}}Z(\boldsymbol{x})\right]^2 \leq CP^2_{\mathbf{X}}\|\theta_1 - \theta_2\|^2,$$

for all $x \in \Omega$ and some constant C. Let $\mathbf{K}_{\boldsymbol{\theta}_l} = (\Phi_{\boldsymbol{\theta}_l}(\boldsymbol{x}_j - \boldsymbol{x}_k))_{jk}$, and $\boldsymbol{r}_{\boldsymbol{\theta}_l} = (\Phi_{\boldsymbol{\theta}_l}(\boldsymbol{x} - \boldsymbol{x}_1), \ldots, \Phi_{\boldsymbol{\theta}_l}(\boldsymbol{x} - \boldsymbol{x}_n))^T$, for $l = 1, 2$. By the mean value theorem, we have

$$\begin{aligned}&\left|\mathcal{I}_{\Phi_{\boldsymbol{\theta}_1},\mathbf{X}}Z(\boldsymbol{x}) - \mathcal{I}_{\Phi_{\boldsymbol{\theta}_2},\mathbf{X}}Z(\boldsymbol{x})\right| = \left|\mathbf{Z}^T(\mathbf{K}_{\boldsymbol{\theta}_1}^{-1}\boldsymbol{r}_{\boldsymbol{\theta}_1} - \mathbf{K}_{\boldsymbol{\theta}_2}^{-1}\boldsymbol{r}_{\boldsymbol{\theta}_2})\right| \\ &\leq \max_{\boldsymbol{\theta}\in\Theta}\left\|\mathbf{Z}^T\frac{\partial}{\partial\boldsymbol{\theta}}(\mathbf{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{r}_{\boldsymbol{\theta}})\right\|\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.\end{aligned}$$

It remains to prove that $\mathbb{E}[\mathbf{Z}^T\partial(\mathbf{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{r}_{\boldsymbol{\theta}})/\partial\theta_l]^2 \leq CP^2_{\mathbf{X}}$, for all $\boldsymbol{\theta} \in \Theta$ and $l = 1, \ldots, q$. For notational simplicity, we denote $\mathbf{K} := \mathbf{K}_{\boldsymbol{\theta}}, \boldsymbol{r} := \boldsymbol{r}_{\boldsymbol{\theta}}, \dot{\mathbf{K}} = \partial\mathbf{K}/\partial\theta_l, \dot{\boldsymbol{r}} = \partial\boldsymbol{r}/\partial\theta_l, \dot{\Phi} = \partial\Phi/\partial\theta_l$, and $\dot{\tilde{\Phi}} = \partial\tilde{\Phi}/\partial\theta_l$. As before, denote the covariance matrix of $\mathbf{Z}$ by $\mathbf{K}_1$. Then

$$\begin{aligned}\mathbb{E}[\mathbf{Z}^T\partial(\mathbf{K}^{-1}\boldsymbol{r})/\partial\theta_l]^2 &= (\boldsymbol{r}^T\mathbf{K}^{-1}\dot{\mathbf{K}}\mathbf{K}^{-1} - \dot{\boldsymbol{r}}^T\mathbf{K}^{-1})\mathbf{K}_1 \\ &\quad\times(\mathbf{K}^{-1}\dot{\mathbf{K}}\mathbf{K}^{-1}\boldsymbol{r} - \mathbf{K}^{-1}\dot{\boldsymbol{r}}) \\ &\leq C_1(\boldsymbol{r}^T\mathbf{K}^{-1}\dot{\mathbf{K}} - \dot{\boldsymbol{r}}^T)\mathbf{K}^{-1}(\dot{\mathbf{K}}\mathbf{K}^{-1}\boldsymbol{r} - \dot{\boldsymbol{r}}),\end{aligned}$$

where the inequality follows from (B.3).

Define $\boldsymbol{u} = (u_1, \ldots, u_n)^T := \mathbf{K}^{-1}\boldsymbol{r}$, and $h(\boldsymbol{y}) := \sum_{j=1}^n u_j \dot{\Phi}(\boldsymbol{y} - \boldsymbol{x}_j) - \dot{\Phi}(\boldsymbol{y} - \boldsymbol{x})$. Clearly, the jth entry of $\dot{\mathbf{K}}\mathbf{K}^{-1}\boldsymbol{r} - \dot{\boldsymbol{r}}$ is $h(\boldsymbol{x}_j)$. Thus,

$$(\boldsymbol{r}^T\mathbf{K}^{-1}\dot{\mathbf{K}} - \dot{\boldsymbol{r}}^T)\mathbf{K}^{-1}(\dot{\mathbf{K}}\mathbf{K}^{-1}\boldsymbol{r} - \dot{\boldsymbol{r}}) = \|\mathcal{I}_{\Phi,\mathbf{X}}h\|^2_{\mathcal{N}_\Phi(\mathbf{R}^d)} \leq \|h\|^2_{\mathcal{N}_\Phi(\mathbf{R}^d)}.$$

Finally, we use Fourier transform to calculate $\|h\|^2_{\mathcal{N}_\Phi(\mathbf{R}^d)}$. It is worth noting that the Fourier transform is performed with respect to $\boldsymbol{y}$, not $\boldsymbol{x}$. It is easy to find that $\tilde{h}(\omega) = (\sum_{j=1}^n u_j e^{-i\omega \boldsymbol{x}_j} - e^{-i\omega\boldsymbol{x}})\dot{\tilde{\Phi}}(\omega)$, which implies

$$\begin{aligned}
\|h\|^2_{\mathcal{N}_\Phi(\mathbf{R}^d)} &= \int_{\mathbf{R}^d} \frac{\tilde{h}(\omega)\bar{\tilde{h}}(\omega)}{\tilde{\Phi}(\omega)} d\omega \\
&= \int_{\mathbf{R}^d} \left| \sum_{j=1}^n u_j e^{i\omega\boldsymbol{x}_j} - e^{i\omega\boldsymbol{x}} \right|^2 \frac{\dot{\tilde{\Phi}}^2(\omega)}{\tilde{\Phi}(\omega)} d\omega \qquad \text{(D.2)} \\
&\leq A_2 \int_{\mathbf{R}^d} \left| \sum_{j=1}^n u_j e^{i\omega\boldsymbol{x}_j} - e^{i\omega\boldsymbol{x}} \right|^2 \tilde{\Phi}(\omega) d\omega \\
&= C_2 P^2_{\Phi,\mathbf{X}} \leq C_2 P^2_{\mathbf{X}}, \qquad \text{(D.3)}
\end{aligned}$$

where the first inequality follows from the condition that $|\partial \log \tilde{\Phi}/\partial \boldsymbol{\theta}_l| = |\dot{\tilde{\Phi}}/\tilde{\Phi}| \leq A_2$.

The proof for the universal kriging case follows from similar lines as that of Theorem 2. Hence, we complete the proof.

Supplementary Materials

In the supplementary materials, we review some mathematical tools which are used in the proofs presented in Appendix, and provide an additional figure related to Table 2.

Acknowledgments

The authors are grateful to an AE and referees for very helpful comments.

Funding

Tuo's work is supported by NSF grant DMS 1564438 and NSFC grants 11501551, 11271355, and 11671386. Wu's work is supported by NSF grant DMS 1564438.

References

Adler, R. J., and Taylor, J. E. (2009), *Random Fields and Geometry*, New York: Springer Science & Business Media. [927]

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: CRC Press, Taylor & Francis Group. [921]

Buslaev, A., and Seleznjev, O. (1999), "On Certain Extremal Problems in the Theory of Approximation of Random Processes," *East Journal on Approximations*, 5, 467–481. [924]

Cramér, H., and Leadbetter, M. R. (1967), *Stationary and Related Stochastic Processes: Sample Function Properties and Their Applications*, Mineola, NY: Courier Corporation. [921]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications: Monographs on Statistics and Applied Probability* (Vol. 66), Boca Raton, FL: CRC Press. [920]

Jiang, J., Nguyen, T., and Rao, J. S. (2011), "Best Predictive Small Area Estimation," *Journal of the American Statistical Association*, 106, 732–745. [927]

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990), "Minimax and Maximin Distance Designs," *Journal of Statistical Planning and Inference*, 26, 131–148. [923,924]

Matheron, G. (1963), "Principles of Geostatistics," *Economic Geology*, 58, 1246–1266. [920]

Pollard, D. (1990), "Empirical Processes: Theory and Applications," in *NSF-CBMS Regional Conference Series in Probability and Statistics, JSTOR*, pp. i–86. [928]

Rao, J. N. K., and Molina, I. (2015), *Small-Area Estimation*, Hoboken, NJ: Wiley. [920]

Rasmussen, C. E. (2006), *Gaussian Processes for Machine Learning*, Cambridge, MA: MIT Press. [920]

Ritter, K. (2000), *Average-Case Analysis of Numerical Problems*, Berlin, Heidelberg: Springer. [924]

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), "Design and Analysis of Computer Experiments," *Statistical Science*, 4, 409–423. [920]

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer Science & Business Media. [921,922,923,925]

Stein, M. L. (1988), "Asymptotically Efficient Prediction of a Random Field With a Misspecified Covariance Function," *The Annals of Statistics*, 16, 55–63. [924,927]

——— (1990a), "Bounds on the Efficiency of Linear Predictions Using an Incorrect Covariance Function," *The Annals of Statistics*, 18, 1116–1138. [924,927]

——— (1990b), "Uniform Asymptotic Optimality of Linear Predictions of a Random Field Using an Incorrect Second-Order Structure," *The Annals of Statistics*, 18, 850–872. [924,927]

——— (1993), "A Simple Condition for Asymptotic Optimality of Linear Predictions of Random Fields," *Statistics & Probability Letters*, 17, 399–404. [927]

——— (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer Science & Business Media. [921,923]

Tuo, R., and Wu, C. F. J. (2015), "Efficient Calibration for Imperfect Computer Models," *The Annals of Statistics*, 43, 2331–2352. [923]

van der Vaart, A. W., and van Zanten, J. H. (2008), "Reproducing Kernel Hilbert Spaces of Gaussian Priors," in *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, Beachwood, OH: Institute of Mathematical Statistics, pp. 200–222. [923]

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer. [922,927]

Wahba, G. (1990), *Spline Models for Observational Data* (Vol. 59), Philadelphia, PA: SIAM. [920]

Wendland, H. (2004), *Scattered Data Approximation* (Vol. 17), Cambridge, UK: Cambridge University Press. [921,922,923,924,926,929]

Wu, C. F. J., and Hamada, M. S. (2009), *Experiments: Planning, Analysis, and Optimization* (2nd ed.), New York: Wiley. [921]

Wu, Z., and Schaback, R. (1993), "Local Error Estimates for Radial Basis Function Interpolation of Scattered Data," *IMA Journal of Numerical Analysis*, 13, 13–27. [923,924]

Xiu, D. (2010), *Numerical Methods for Stochastic Computations: A Spectral Method Approach*, Princeton, NJ: Princeton University Press. [924]

Yakowitz, S., and Szidarovszky, F. (1985), "A Comparison of Kriging With Nonparametric Regression Methods," *Journal of Multivariate Analysis*, 16, 21–53. [924,927]

Ying, Z. (1991), "Asymptotic Properties of a Maximum Likelihood Estimator With Data From a Gaussian Process," *Journal of Multivariate Analysis*, 36, 280–296. [926]

Zhang, H. (2004), "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics," *Journal of the American Statistical Association*, 99, 250–261. [926]