On Prediction Properties of Kriging: Uniform Error Bounds and Robustness

Wenjia Wang∗1, Rui Tuo†1,2 and C. F. Jeff Wu‡1

1The H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
2 Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

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) 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.

Keywords: Gaussian Process modeling; Uniform convergence; Space-filling designs; Radial basis functions; Spatial statistics.

∗Wenjia Wang is a Ph.D. in Georgia Institute of Technology, Atlanta, GA 30332, USA (Email: wenjia-wang@gatech.edu);
†Rui Tuo is Assistant Professor in Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China and a visitor at Georgia Institute of Technology Atlanta, GA 30332, USA (Email: tuorui@amss.ac.cn); Tuo’s work is supported by NSF grant DMS 1564438 and also by the National Center for Mathematics and Interdisciplinary Sciences in CAS and NSFC grants 11501551, 11271355 and 11671386.
‡C. F. Jeff Wu is Professor and Coca-Cola Chair in Engineering Statistics at School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA (Email: jeff.wu@isye.gatech.edu). Wu’s work is supported by NSF grant DMS 1564438.
The authors are grateful to an AE and referees for very helpful comments.
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) 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. The theory of radial basis function approximation (Wendland, 2004) and a maximum inequality for Gaussian processes (Adler and Taylor, 2009) are employed as key tools in our technical development.

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. Concluding remarks and discussion are given in Section 5. Appendix A includes some necessary mathematical tools. Appendix B contains the proof of Theorem 1, the main theorem of this work.

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

\[
\text{Cov}(Z(x), Z(x')) = \sigma^2 \Psi(x - x'),
\]

for any \( x, x' \in \mathbb{R}^d \), where \( \sigma^2 \) is the variance and \( \Psi \) 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
\[
\Psi(x; \phi) = \exp\{-\phi \|x\|^2\},
\] (2.1)
with some \( \phi > 0 \), where \( \| \cdot \| \) denotes the Euclidean norm. The isotropic Matérn correlation function (Santner et al., 2003; Stein, 1999) is defined as
\[
\Psi(x; \nu, \phi) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(2\sqrt{\nu \phi \|x\|})^\nu K_\nu(2\sqrt{\nu \phi \|x\|}),
\] (2.2)
where \( \phi, \nu > 0 \) and \( K_\nu \) is the modified Bessel function of the second kind. The parameter \( \nu \) is often called the smoothness parameter, because it determines the smoothness of the Gaussian process (Cramér and Leadbetter, 1967).

Suppose that we have observed \( Z(x_1), \ldots, Z(x_n) \), in which \( x_1, \ldots, x_n \) are distinct points. We shall use the terminology in design of experiments (Wu and Hamada, 2011) and call \( \{x_1, \ldots, x_n\} \) 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 paper, we do not assume any (algebraic or geometric) structure for the design points \( \{x_1, \ldots, x_n\} \). They are called scattered points in the applied mathematics literature.

The aim of (simple) kriging is to predict \( Z(x) \) at an untried \( x \) based on the observed data \( Z(x_1), \ldots, Z(x_n) \), which is done by calculating the conditional distribution. It follows from standard arguments (Santner et al., 2003; Banerjee et al., 2004) that, conditional on \( Z(x_1), \ldots, Z(x_n) \), \( Z(x) \) is normally distributed, with
\[
\mathbb{E}[Z(x)|Z(x_1), \ldots, Z(x_n)] = r^T(x)K^{-1}Y, \quad a.s., \quad (2.3)
\]
\[
\text{Var}[Z(x)|Z(x_1), \ldots, Z(x_n)] = \sigma^2(1 - r^T(x)K^{-1}r(x)), \quad a.s., \quad (2.4)
\]
where \( r(x) = (\Psi(x-x_1), \ldots, \Psi(x-x_n))^T, K = (\Psi(x_j-x_k))_{jk} \) and \( Y = (Z(x_1), \ldots, Z(x_n))^T \).

The conditional expectation \( \mathbb{E}[Z(x)|Z(x_1), \ldots, Z(x_n)] \) is a natural predictor of \( Z(x) \) using \( Z(x_1), \ldots, Z(x_n) \), because it is the best linear predictor (Stein, 1999; Santner et al., 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 et al., 2003; Stein, 1999). For the ease of mathematical treatment, we only consider simple kriging in this work. The convergence theory for ordinary and universal kriging requires separate developments. Further discussions are deferred to Section 5.

### 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 $Z(x)$ becomes a deterministic function, often called a sample path. It can be verified that, as a function of $x$, $r^T k^{-1} y$ in (2.3) goes through each $Z(x_j), j = 1, \ldots, n$.

The above interpolation scheme can be applied to an arbitrary function $f$. Specifically, given design points $X = (x_1, \ldots, x_n)$ and observations $f(x_1), \ldots, f(x_n)$, we define the kriging interpolant by

$$I_{\psi,X} f(x) = r^T (x) k^{-1} F,$$

(2.5)

where $r(x) = (\psi(x-x_1), \ldots, \psi(x-x_n))^T$, $K = (\psi(x_j-x_k))_{jk}$ and $F = (f(x_1), \ldots, f(x_n))^T$.

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,

$$E[Z(x)|Z(x_1), \ldots, Z(x_n)] = I_{\psi,X} Z(x), \ a.s.$$

(2.6)

As mentioned in Section 2.1, the conditional expectation $E[Z(x)|Z(x_1), \ldots, Z(x_n)]$ is a natural predictor of $Z(x)$. A key objective of this work is to derive a uniform bound of the predictive error of the kriging method, given by $Z(x) - E[Z(x)|Z(x_1), \ldots, Z(x_n)]$, which is equal to $Z(x) - I_{\psi,X} Z(x)$ almost surely.
In practice, $\Psi$ 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 $\Phi$ for prediction. We call $\Psi$ the \textit{true correlation function} and $\Phi$ the \textit{imposed correlation function}. Under the imposed correlation function, the kriging interpolant of the underlying Gaussian process becomes $\mathcal{I}_{\Phi, X} Z(x)$. 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. Specifically, we are interested in bounding the \textit{maximum prediction error} over a region $\Omega$,

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

(2.7)

in a probabilistic manner, where $\Omega$ is the region of interest, also called the experimental region, and $\Omega \supset \{x_1, \ldots, x_n\}$.

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 et al., 2003), in which the design points spread (approximately) evenly in the design region.
3 Uniform error bounds for kriging

This section contains our main theoretical results on the prediction error of 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, i.e., \( \Phi = \Psi \). To incorporate the case of misspecified correlation functions, we define the power function as

\[
P_{\Phi, X}(x) = 1 - r^T(x)K^{-1}r(x),
\]

where \( r(x) = (\Phi(x - x_1), \ldots, \Phi(x - x_n))^T \), and \( K = (\Phi(x_j - x_k))_{jk} \).

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 \( P_{\Phi, X}(x) \leq 1 \).

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

\[
P_{\Phi, X} := \sup_{x \in \Omega} P_{\Phi, X}(x).
\]

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 \( \Psi \) is continuous and integrable on \( \mathbb{R}^d \).

**Condition 1.** The kernels \( \Psi \) and \( \Phi \) are continuous and integrable on \( \mathbb{R}^d \), satisfying

\[
\| \tilde{\Psi} / \tilde{\Phi} \|_{L_\infty(\mathbb{R}^d)} =: A_1^2 < +\infty.
\]

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

\[
\int_{\mathbb{R}^d} \| \omega \|^\alpha \tilde{\Phi}(\omega) d\omega =: A_0 < +\infty.
\]

Now we are able to state the main theorem of this paper. Recall that \( \sigma^2 \) is the variance of \( Z(x) \).
Theorem 1. Suppose Condition 1 holds, and the design set $X$ is dense enough in the sense that $P_{\Phi,X}$ defined in (3.2) is no more than some given constant $C$. Then for any $u > 0$, with probability at least $1 - 2\exp\{-u^2/(2A_1^2\sigma^2P_{\Phi,X})\}$, the kriging predictive error has the upper bound

$$\sup_{x \in \Omega} |Z(x) - I_{\Phi,X}Z(x)| \leq K\sigma P_{\Phi,X} \log^{1/2}(e/P_{\Phi,X}) + u. \quad (3.5)$$

Here the constants $C, K > 0$ depend only on $\Omega, \alpha, A_0$ 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 $X$ and the kernel $\Phi$ 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 $\Phi$ is no smoother than the true correlation function $\Psi$. 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 $P_{\Phi,X}$ if $P_{\Phi,X}$ is large enough. Therefore, it is reasonable to consider designs which minimize $P_{\Phi,X}$. Such a construction depends on the specific form of $\Phi$. 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 non-asymptotic 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 $Z(x)$ and the kernel function $\Phi$ are fixed, and the number of design points increases so that $P_{\Phi,X}$ 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 $\Psi, \Phi$ and $\Omega$, we have the following asymptotic results

\[
\sup_{x \in \Omega} |Z(x) - \mathcal{I}_{\Phi, X} Z(x)| = O_P(P_{\Phi, X} \log^{1/2}(1/P_{\Phi, X})),
\]

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

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

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

\[
\mathbb{E} \sup_{x \in \Omega} |Z(x) - \mathcal{I}_{\Phi, X} Z(x)|^p
\]

\[
= \int_0^\infty \mathbb{P}(\sup_{x \in \Omega} |Z(x) - \mathcal{I}_{\Phi, X} Z(x)| > t^{1/p}) dt
\]

\[
= \left( \int_0^{[K\sigma P_{\Phi, X} \log^{1/2}(e/P_{\Phi, X})]^p} + \int_{[K\sigma P_{\Phi, X} \log^{1/2}(e/P_{\Phi, X})]^p}^\infty \mathbb{P}(\sup_{x \in \Omega} |Z(x) - \mathcal{I}_{\Phi, X} Z(x)| > t^{1/p}) dt \right)
\]

\[
\leq \left[ K\sigma P_{\Phi, X} \log^{1/2}(e/P_{\Phi, X}) \right]^p + \int_0^\infty 2 \exp\{-t^{2/p}/(2A^2\sigma^2 P_{\Phi, X}^2)\} dt
\]

\[
= O(P_{\Phi, X}^p \log^{p/2}(1/P_{\Phi, X})),
\]

where the inequality follows from Theorem 1.

We believe that (3.6) and (3.7) are the full convergence rate because from (A.1) in Appendix A.1 we can see that the convergence rate of the radial basis approximation for deterministic functions in the reproducing kernel Hilbert space is $O(P_{\Phi, X})$ and these two rates are nearly at the same order of magnitude, except 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 $\Psi = \Phi$, $P_{\Phi, X}$ is the supremum of the pointwise predictive standard deviation. Thus Corollary 1 implies that, if $\Psi$ 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 $P_{\Phi, X}$, in which the effects of the
design $X$ and the kernel $\Phi$ 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 $X$. Given the experimental region $\Omega$, the fill distance of a design $X$ is defined as

$$h_X := \sup_{x \in \Omega} \min_{x_j \in X} \| x - x_j \|.$$  \hspace{1cm} (3.8)

Clearly, the fill distance quantifies the space-filling property (Santner et al., 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 et al., 1990).

The upper bounds of $P_{\Phi,X}$ 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 $\Omega$ and the scale parameter $\phi$ in (2.1), such that $P_{\Phi,X} \leq c h_X^j h_X$ provided that $h_X \leq h_0$.

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

Using the upper bounds of $P_{\Phi,X}$ 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 $\Phi$ is a Matérn kernel in (2.2) with smoothness parameter $\nu$. It is known that

$$\tilde{\Phi}(\omega) = 2^d \pi^{d/2} \frac{\Gamma(\nu + d/2)}{\Gamma(\nu)} (4\nu\phi^2)^\nu (4\nu\phi^2 + \| \omega \|^2)^{-(\nu + d/2)},$$

where $\phi$ is the scale parameter in (2.2). See, for instance, Wendland (2004); Tuo and Wu (2015). Suppose $\Psi$ is a Matérn correlation function with smoothness $\nu_0$. It can be verified that Condition 1 holds if and only if $0 < \nu \leq \nu_0$. Therefore, if $0 < \nu \leq \nu_0$, 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 $O_p(h_X^\nu \log^{1/2}(1/h_X))$ as $h_X$ 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 $\Phi$ is the same as in Example 1, and $\Psi$ is a Gaussian correlation function in (2.1), with spectral density (Santner et al., 2003) $\tilde{\Psi}(\omega) = (\pi/\phi)^{d/2} \exp\{-\|\omega\|^2/(4\phi)\}$, where $\phi$ is the scale parameter in (2.1). Then Condition 1 holds for any choice of $\nu$. 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 $\Phi$ is a Gaussian kernel in (2.1). Then we can invoke Lemmas 1 and Theorem 1 to obtain the rate of convergence $O_p(h_X^{c/h_X} - 1/2 \log^{1/2}(1/h_X))$ 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 $h_X$. 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 $h_X$. This suggests that we should consider the designs with a minimum $h_X$ value, which are known as the maximin distance designs (Johnson et al., 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 $P_{\Phi, X}$ 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 deterministic 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.

| Article/Book                | Model assumption            | Predictor | Design            | Type of convergence | Rate of convergence |
|-----------------------------|----------------------------|-----------|-------------------|---------------------|---------------------|
| present work                | Gaussian process            | kriging   | scattered points   | $L_p$ conv., uniform in $x$ | $h_X^{(log(1/h_X))1/2}$ |
| Yakowitz and Szidarovszky (1985) | stochastic process with misspecification | kriging   | scattered points   | Mean square conv., pointwise in $x$ | $N/A$ |
| Stein (1990b)               | Stochastic process with misspecification | kriging   | regular grid points | Mean square conv., pointwise in $x$ | $n^{-v/d}$ |
| Buslaev and Seleznjev (1999) | Gaussian process            | best linear approximation | optimally chosen points in an interval | $L_p$ conv., uniform in $x$ | $n^{-v/(log n)^{1/2}}$ |
| Ritter (2000)               | Gaussian process            | kriging   | optimally chosen points | Mean square conv., $L_2$ in $x$ | $n^{-v/d}$ |
| Wu and Schaback (1993)      | Deterministic function      | kriging   | scattered points   | uniform in $x$ | $h_X^{v}$ |

Table 1: Comparison between our work and some existing results

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 $O_p(h_X^{v}log^{1/2}(1/h_X))$. If a space-filling design is used in an interval, then $h_X \sim 1/n$ and the convergence rate is $O_p(n^{-v}(log n)^{1/2})$, 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); Stein (1988, 1990a,b). 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 5.

4 Simulation studies

In Example 1, we have shown that if $\Psi$ and $\Phi$ are Matérn kernels with smoothness parameters $\nu_0$ and $\nu$, respectively, and $\nu \leq \nu_0$, then the kriging predictor converges with a rate at least $O_p(h_X^{\nu} \log^{1/2}(1/h_X))$. In this section we report simulation studies that verify that this rate is sharp, i.e., the true convergence rate coincides with that given by the theoretical upper bound.

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

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

for some constant $c$ independent of $h_X$. Taking logarithm on both sides of the above formula yields

$$\log \mathcal{E} \approx \nu \log h_X + \frac{1}{2} \log(-\nu \log h_X) + \log c.$$  \hfill (4.1)

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

$$\log \mathcal{E} \approx \nu \log h_X + \log c.$$  \hfill (4.2)

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

13
In our simulation studies, the experimental region is chosen to be $\Omega = [0, 1]^2$. To estimate the regression coefficient $\nu$ in (4.2), we choose 50 different maximin Latin hypercube designs (Santner et al., 2003) with sample sizes $10k$, for $k = 1, 2, \ldots, 50$. Note that each design corresponds to a specific value of the fill distance $h_X$. For each $k$, we simulate the Gaussian processes 100 times to reduce the simulation error. For each simulated Gaussian process, we compute $\sup_{x \in \Omega_1} |Z(x) - I_{\Phi, X} Z(x)|$ to approximate the sup-error $\sup_{x \in \Omega} |Z(x) - I_{\Phi, X} Z(x)|$, where $\Omega_1$ 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 $\sup_{x \in \Omega_1} |Z(x) - I_{\Phi, X} Z(x)|$ 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 $\nu_0$ and $\nu$, respectively. We summarize the simulation results in Table 2.

| $\nu_0$ | $\nu$ | 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 |

Table 2: Numerical studies on the convergence rates of kriging prediction. (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 $|\text{regression coefficient-theoretical assertion}|/(\text{theoretical assertion}).)

Figure 1 shows the relationship between the logarithm of the fill distance (i.e., $\log h_X$) and the logarithm of the average prediction error (i.e., $\log \mathcal{E}$) in scatter plots for the four cases given in Table 2. The solid line in each panel shows the linear regression fit calculated from the data. Each of the regression lines in Figure 1 fits the data very well, which gives an empirical confirmation of the approximation in (4.2). It is also observed from Figure 1...
that, as the fill distance decreases, the maximum prediction error also decreases.

![Graph 1](image1)

![Graph 2](image2)

![Graph 3](image3)

![Graph 4](image4)

Figure 1: The regression line of $\log \sup_{x \in \Omega} \epsilon(x)$ on $\log h_x$. Each point denotes one average prediction error for each $n$. Panel 1: $\nu_0 = 3, \nu = 2.5$. Panel 2: $\nu_0 = 5, \nu = 3.5$. Panel 3: $\nu_0 = \nu = 3.5$. Panel 4: $\nu_0 = \nu = 5$.

Returning to Table 2, it is seen 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 $\nu$ remains the same, even if $\nu_0$ changes. On the other hand,
the third and the fifth rows show that, the regression coefficient changes significantly as \( \nu \) changes, even if \( \nu_0 \) 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 \( \log h_X \) in (4.1) to assess the convergence rate but they do not cover the second term \( \log(-\nu \log h_X) \), which is of lower order.

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 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. 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.

In this paper, we only consider Gaussian process models with mean zero, which is referred to as the simple kriging. A natural extension of this work is to include the Gaussian process models with a mean function modeled as a linear combination of a finite set of functions, known as the universal kriging. In this situation, one would consider the best linear unbiased predictor (BLUP) instead of the conditional expectation. The mathematical treatments should follow the same lines but the details are more cumbersome. Another extension is to consider the impact of a misspecified mean function. See, for example, Jiang et al. (2011) who addresses this problem in the context of small area estimation. We can
consider the study of the effect of joint misspecification of mean and correlation functions in a future work.

We have proved in Theorem 1 that the kriging predictor is consistent if the imposed correlation function is *undersmoothed*, i.e., 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,b, 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 important topic is the kriging predictive performance when the correlation function is estimated. In this paper, we consider kriging with a misspecified but fixed kernel function. It is shown that the prediction error can be minimized if the imposed kernel has the same smoothness as the true correlation function. A natural question is whether the optimal rate of convergence can be achieved with an estimated correlation function.

### Appendix

#### A Auxiliary tools

In this section, we review some mathematical tools which are used in the proof of Theorem 1 in Appendix B.

##### A.1 Reproducing kernel Hilbert spaces

There are several equivalent ways to define the reproducing kernel Hilbert spaces. See, for example, Wahlba (1990); Gu (2013); Wendland (2004); van der Vaart and van Zanten
Here we adopt the one using the Fourier transform. See Girosi et al. (1995) and Theorem 10.12 of Wendland (2004). Let $L_2(\mathbb{R}^d)$ be the space of complex-valued square integrable functions on $\mathbb{R}^d$, and $C(\mathbb{R}^d)$ be the space of continuous real-valued functions on $\mathbb{R}^d$.

**Definition 1.** Let $\Phi$ be a positive definite kernel function which is continuous and integrable in $\mathbb{R}^d$. Define the reproducing kernel Hilbert space $\mathcal{N}_\Phi(\mathbb{R}^d)$ as

$$\mathcal{N}_\Phi(\mathbb{R}^d) := \{ f \in L_2(\mathbb{R}^d) \cap C(\mathbb{R}^d) : \tilde{f} / \sqrt{\tilde{\Phi}} \in L_2(\mathbb{R}^d) \},$$

with the inner product

$$\langle f, g \rangle_{\mathcal{N}_\Phi(\mathbb{R}^d)} = (2\pi)^{-d} \int_{\mathbb{R}^d} \frac{\tilde{f}(\omega)\tilde{g}(\omega)}{\tilde{\Phi}(\omega)} d\omega.$$ 

Denote the norm of $\mathcal{N}_\Phi(\mathbb{R}^d)$ by $\| \cdot \|_{\mathcal{N}_\Phi(\mathbb{R}^d)}$. For $f \in \mathcal{N}_\Phi(\mathbb{R}^d)$, a pointwise error bound for the radial basis function interpolation is given by (Wu and Schaback (1993), Theorem 4.5):

$$|f(x) - \mathcal{I}_{\Phi,X} f(x)| \leq P_{\Phi,X}(x) \|f\|_{\mathcal{N}_\Phi(\mathbb{R}^d)},$$

for each $x \in \mathbb{R}^d$, where $P_{\Phi,X}(x)$ is defined in (3.1).

### A.2 A Maximum inequality for Gaussian processes

It is worth noting that $\mathcal{I}_{\Phi,X}$ is a linear map between two functions, and therefore $\mathcal{I}_{\Phi,X} Z(x)$ is also a Gaussian process. Therefore, the problem in (2.7) is to bound the maximum value of a Gaussian process.

The theory of bounding the maximum value of a Gaussian process is well-established in the literature. The main step of finding an upper bound is to calculate the *covering number* of the index space. Here we review the main results. Detailed discussions can be found in Adler and Taylor (2009).

Let $Z_t$ be a Gaussian process indexed by $t \in T$. Here $T$ can be an arbitrary set. The Gaussian process $Z_t$ induces a metric on $T$, defined by

$$\vartheta(t_1, t_2) = \sqrt{\mathbb{E}(Z_{t_1} - Z_{t_2})^2}. \quad (A.2)$$
The \( \epsilon \)-covering number of the metric space \((T, \varnothing)\), denoted as \( N(\epsilon, T, \varnothing) \), is the minimum integer \( N \) so that there exist \( N \) distinct balls in \((T, \varnothing)\) with radius \( \epsilon \), and the union of these balls covers \( T \). Let \( D \) be the diameter of \( T \). The supremum of a Gaussian process is closely tied to a quantity called the \textit{entropy integral}, defined as

\[
\int_0^{D/2} \sqrt{\log N(\epsilon, T, \varnothing)} \, d\epsilon.
\]

(A.3)

Lemma 3 gives a maximum inequality for Gaussian processes, which is a direct consequence of Theorems 1.3.3 and 2.1.1 of Adler and Taylor (2009).

**Lemma 3.** Let \( Z_t \) be a centered separable Gaussian process on a \( \varnothing \)-compact \( T \), \( \varnothing \) the metric, and \( N \) the \( \epsilon \)-covering number. Then there exists a universal constant \( K \) such that for all \( u > 0 \),

\[
\mathbb{P}(\sup_{t \in T} |Z_t| > K \int_0^{D/2} \sqrt{\log N(\epsilon, T, \varnothing)} \, d\epsilon + u) \leq 2e^{-u^2/2\sigma_T^2},
\]

(A.4)

where \( \sigma_T^2 = \sup_{t \in T} \mathbb{E}Z_t^2 \).

**B Proof of Theorem 1**

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

\[
\varnothing(x, x')^2 = \mathbb{E}(g(x) - g(x'))^2
\]

\[
= \mathbb{E}(Z(x) - \mathcal{I}_{\Phi, \mathbf{X}}Z(x) - (Z(x') - \mathcal{I}_{\Phi, \mathbf{X}}Z(x')))^2
\]

\[
= \Psi(x - x) - 2r^T(x)K^{-1}r(x) + r^T(x)K^{-1}K_1K^{-1}r(x)
\]

\[
+ \Psi(x' - x') - 2r^T(x')K^{-1}r(x') + r^T(x')K^{-1}K_1K^{-1}r(x')
\]

\[
- 2[\Psi(x - x') - r^T(x')K^{-1}r(x) - r^T(x')K^{-1}K_1K^{-1}r(x') + r^T(x)K^{-1}K_1K^{-1}r(x')],
\]

where \( r_1(\cdot) = (\Psi(\cdot - x_1), \ldots, \Psi(\cdot - x_n))^T \), \( r(\cdot) = (\Phi(\cdot - x_1), \ldots, \Phi(\cdot - x_n))^T \), \( K_1 = (\Psi(x_j - x_k))_{jk} \), and \( K = (\Phi(x_j - x_k))_{jk} \).

The rest of our proof consists of the following steps. In step 1, we bound the covering number \( N(\epsilon, \Omega, \varnothing) \). Next we bound the diameter \( D \). In step 3, we invoke Theorem 3 to
obtain a bound for the entropy integral. In the last step, we use (A.4) to obtain the desired results.

**Step 1: Bounding the covering number**

Let \( h(\cdot) = \Psi(x - \cdot) - \Psi(x' - \cdot) \) and \( h_1(\cdot) = r^T(x)K^{-1}r_1(\cdot) - r^T(x')K^{-1}r_1(\cdot) \). It can be verified that

\[
\vartheta(x, x')^2 = -[h(x') - I_{\Psi, X}h(x)] + [h(x) - I_{\Psi, X}h(x)] \\
+ [h_1(x') - I_{\Psi, X}h_1(x')] - [h_1(x) - I_{\Psi, X}h_1(x)].
\]

By Condition 1, \( h \in \mathcal{N}_\Psi(R^d) \), since \( \Psi(x - \cdot) \in \mathcal{N}_\Psi(R^d) \) for any \( x \in \Omega \). Thus, by (A.1),

\[
\vartheta(x, x')^2 \leq 2P_{\Psi, X}(\|h\|_{\mathcal{N}_\Psi(R^d)} + \|h_1\|_{\mathcal{N}_\Psi(R^d)}).
\]

By Definition 1,

\[
\|h\|_{\mathcal{N}_\Psi(R^d)}^2 = (2\pi)^{-d} \int_{R^d} \frac{|\hat{h}(\omega)|^2}{\Phi(\omega)} d\omega. \tag{B.2}
\]

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

\[
\|h\|_{\mathcal{N}_\Psi(R^d)}^2 = (2\pi)^{-d} \int_{R^d} \frac{|\hat{h}(\omega)|^2}{\Phi(\omega)} d\omega \leq A_1^2(2\pi)^{-d} \int_{R^d} \frac{|\hat{h}(\omega)|^2}{\Psi(\omega)} d\omega = A_1^2\|h\|_{\mathcal{N}_\Psi(R^d)}^2. \tag{B.3}
\]

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

\[
|1 - \cos x| \leq 2|x|^\beta. \tag{B.4}
\]

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

\[
\|h\|_{\mathcal{N}_\Psi(R^d)}^2 = \Psi(x - x) - 2\Psi(x' - x) + \Psi(x' - x') \\
= 2(2\pi)^{-d} \int_{R^d} (1 - e^{i(x-x')\omega}) \bar{\Psi}(\omega) d\omega \\
\leq \left(4(2\pi)^{-d} \int_{R^d} \|\omega\|^\beta \bar{\Psi}(\omega) d\omega\right) \|x - x'\|^\beta \tag{B.5}
\]

\[
=: C_1 \|x - x'\|^\beta, \tag{B.6}
\]
for any $0 < \beta \leq \alpha$. In particular, we now choose $\beta = \alpha/2$. Now we consider $h_1(\cdot)$. It follows from a similar argument that $\|h_1\|^2_{\mathcal{N}_q(\mathbb{R}^d)} \leq A_1^2\|h_1\|^2_{\mathcal{N}_q(\mathbb{R}^d)}$. The definition of $h_1$ implies $\|h_1\|^2_{\mathcal{N}_q(\mathbb{R}^d)} = (r(x') - r(x))^T K^{-1} K_1 K^{-1} (r(x') - r(x))$.

For any $u = (u_1, ..., u_n)^T$, the Fourier inversion theorem and Condition 1 yield

$$
\sum_{j,k=1}^n u_j \bar{u}_k \Psi(x_j - x_k) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \sum_{j,k=1}^n u_j \bar{u}_k e^{i(x_j - x_k)^T \omega} \tilde{\Psi}(\omega) d\omega
$$

$$
= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{i x_j \cdot \omega} \right|^2 \tilde{\Psi}(\omega) d\omega
$$

$$
\leq A_1^2 \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{i x_j \cdot \omega} \right|^2 \tilde{\Phi}(\omega) d\omega
$$

$$
= A_1^2 \sum_{j,k=1}^n u_j \bar{u}_k \Phi(x_j - x_k).
$$

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

$$
\|h_1\|^2_{\mathcal{N}_q(\mathbb{R}^d)} \leq A_1^2(r(x') - r(x))^T K^{-1}(r(x') - r(x)).
$$

(B.8)

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

$$
\sum_{j,k=1}^n u_j \bar{u}_k \Psi(x_j - x_k) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{i x_j \cdot \omega} \right|^2 \tilde{\Phi}(\omega) d\omega
$$

$$
\leq A_1^2 \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{i x_j \cdot \omega} \right|^2 \tilde{\Phi}(\omega) d\omega
$$

$$
= A_1^2 \sum_{j,k=1}^n u_j \bar{u}_k \Phi(x_j - x_k).
$$

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

$$
\|h_1\|^2_{\mathcal{N}_q(\mathbb{R}^d)} \leq C_2 \|x - x'\|^\alpha/2,
$$

(B.10)

for a constant $C_2$. 

21
In view of (B.1), (B.6) and (B.10), there exists a constant $C_3$ such that
\[
\vartheta(x, x')^2 \leq C_3 P_{\Phi, X} \| x - x' \|^{\alpha/4}.
\] (B.11)

Therefore, the covering number is bounded above by
\[
\log N(\epsilon, \Omega, \vartheta) \leq \log N\left( \frac{\epsilon^{8/\alpha} P_{\Phi, X}^{4/\alpha}}{C_3^4 P_{\Phi, X}^{4/\alpha}} , \Omega, \| \cdot \| \right).
\] (B.12)

The right side of (B.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
\[
\log N(\epsilon, \Omega, \vartheta) \leq C_{4, 0} \log \left( \frac{C_{5, 0} C_3^{4/\alpha} P_{\Phi, X}^{4/\alpha}}{\epsilon^{8/\alpha}} \right) =: C_4 \log \left( \frac{C_{5, 0} P_{\Phi, X}^{1/2}}{\epsilon} \right),
\] (B.13)
provided that
\[
\epsilon < C_{5, 0} P_{\Phi, X}^{1/2},
\] (B.14)
where $C_{4, 0}$ and $C_{5, 0}$ are constants depending on the dimension and the Euclidean diameter of $\Omega$.

**Step 2: Bounding the diameter $D$**

Recall that the diameter is defined by $D = \sup_{x, x' \in \Omega} \vartheta(x, x')$. For any $x, x' \in \Omega$,
\[
\vartheta(x, x')^2 = \mathbb{E}(g(x) - g(x'))^2 \leq 4 \sup_{x \in \Omega} \mathbb{E}(g(x))^2
\]
\[
= 4 \sup_{x \in \Omega} \mathbb{E}(Z(x) - T_{\Phi, X} Z(x))^2
\]
\[
= 4 \sup_{x \in \Omega} (\Psi(x - x) - 2r_1^T(x)K^{-1}r(x) + r^T(x)K^{-1}K_1 K^{-1}r(x)),
\] (B.15)
where $r$, $r_1$, $K$ and $K_1$ are defined in the beginning of Appendix B.

Combining identity (B.7) with
\[
\Psi(x_j - x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i(x - x_j)^T \omega} \tilde{\Psi}(\omega) d\omega,
\]
for any $\mathbf{u} = (u_1, ..., u_n)$, under Condition 1, we have

$$u^T K_1 \mathbf{u} - 2 \mathbf{u}^T \mathbf{r}(\mathbf{x}) + \Psi(\mathbf{x} - \mathbf{x})$$

$$= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{ix_j^T \omega} - e^{ix^T \omega} \right|^2 \Psi(\omega) d\omega$$

$$\leq A_1^2 \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{ix_j^T \omega} - e^{ix^T \omega} \right|^2 \Phi(\omega) d\omega$$

$$= A_1^2 (u^T K \mathbf{u} - 2 \mathbf{u}^T \mathbf{r}(\mathbf{x}) + \Phi(\mathbf{x} - \mathbf{x})). \quad (B.16)$$

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

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

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

$$D \leq 2A_1 P_{\Phi.\mathbf{x}}. \quad (B.17)$$

**Step 3: Bounding the entropy integral**

Under Condition 1, if

$$P_{\Phi.\mathbf{x}} < C^2_5 / A_1^2 := C,$$

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

$$\int_0^{D/2} \sqrt{\log N(\epsilon, \Omega, \delta)} d\epsilon \leq \int_0^{A_1 P_{\Phi.\mathbf{x}}} \sqrt{C_4 \log \left( \frac{C_5 P_{\Phi.\mathbf{x}}^{1/2}}{\epsilon} \right)} d\epsilon$$

$$\leq \left( \int_0^{A_1 P_{\Phi.\mathbf{x}}} d\epsilon \right)^{1/2} \left( \int_0^{A_1 P_{\Phi.\mathbf{x}}} C_4 \log \left( \frac{C_5 P_{\Phi.\mathbf{x}} \epsilon}{d\epsilon} \right) \right)^{1/2} \quad (B.18)$$

$$= C_4^{1/2} A_1 P_{\Phi.\mathbf{x}} \sqrt{\log \left( \frac{C_5 \epsilon}{A_1 P_{\Phi.\mathbf{x}}^{1/2}} \right)}. \quad (B.19)$$

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

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

Noting that $\sup_{x \in \Omega} \mathbb{E}(Z(\mathbf{x}) - \mathcal{I}_{\Phi.\mathbf{x}} Z(\mathbf{x}))^2 = D^2$, by plugging (B.17) into (A.4), we obtain the desired inequality, which completes the proof.
References

Adler, R. J. and J. E. Taylor (2009). Random Fields and Geometry. Springer Science & Business Media.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2004). Hierarchical Modeling and Analysis for Spatial Data. CRC Press, Taylor & Francis Group.

Buslaev, A. and O. Seleznjev (1999). On certain extremal problems in the theory of approximation of random processes. East Journal on Approximations 5(4), 467–481.

Cramér, H. and M. R. Leadbetter (1967). Stationary and Related Stochastic Processes: Sample Function Properties and Their Applications. Courier Corporation.

Fan, J. and I. Gijbels (1996). Local Polynomial Modeling and Its Applications: Monographs on Statistics and Applied Probability, Volume 66. CRC Press.

Girosi, F., M. Jones, and T. Poggio (1995). Regularization theory and neural networks architectures. Neural Computation 7(2), 219–269.

Gu, C. (2013). Smoothing Spline ANOVA Models. Springer.

Jiang, J., T. Nguyen, and J. S. Rao (2011). Best predictive small area estimation. Journal of the American Statistical Association 106(494), 732–745.

Johnson, M. E., L. M. Moore, and D. Ylvisaker (1990). Minimax and maximin distance designs. Journal of Statistical Planning and Inference 26(2), 131–148.

Matheron, G. (1963). Principles of geostatistics. Economic Geology 58(8), 1246–1266.

Pollard, D. (1990). Empirical processes: theory and applications. In NSF-CBMS regional conference series in probability and statistics, pp. i–86. JSTOR.

Rao, J. N. K. and I. Molina (2015). Small-Area Estimation. Wiley Online Library.

Rasmussen, C. E. (2006). Gaussian Processes for Machine Learning. MIT Press.

Ritter, K. (2000). Average-case Analysis of Numerical Problems. Springer.
Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn (1989). Design and analysis of computer experiments. *Statistical Science*, 409–423.

Santner, T. J., B. J. Williams, and W. I. Notz (2003). *The Design and Analysis of Computer Experiments*. Springer Science & Business Media.

Stein, M. L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *The Annals of Statistics* 16(1), 55–63.

Stein, M. L. (1990a). Bounds on the efficiency of linear predictions using an incorrect covariance function. *The Annals of Statistics*, 1116–1138.

Stein, M. L. (1990b). Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. *The Annals of Statistics*, 850–872.

Stein, M. L. (1993). A simple condition for asymptotic optimality of linear predictions of random fields. *Statistics & Probability Letters* 17(5), 399–404.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media.

Tuo, R. and C. F. J. Wu (2015). Efficient calibration for imperfect computer models. *The Annals of Statistics* 43(6), 2331–2352.

van der Vaart, A. W. and J. H. van Zanten (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, pp. 200–222. Institute of Mathematical Statistics.

Wahba, G. (1990). *Spline Models for Observational Data*, Volume 59. SIAM.

Wendland, H. (2004). *Scattered Data Approximation*, Volume 17. Cambridge university press.

Wu, C. F. J. and M. S. Hamada (2011). *Experiments: Planning, Analysis, and Optimization*, Volume 552. John Wiley & Sons.

Wu, Z. and R. Schaback (1993). Local error estimates for radial basis function interpolation of scattered data. *IMA Journal of Numerical Analysis* 13(1), 13–27.
Xiu, D. (2010). *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press.

Yakowitz, S. and F. Szidarovszky (1985). A comparison of kriging with nonparametric regression methods. *Journal of Multivariate Analysis 16*(1), 21–53.