Journal of Machine Learning Research 21 (2020) 1-38 Submitted 12/19; Revised 7/20; Published 9/20

Kriging Prediction with Isotropic Matérn Correlations: Robustness and Experimental Designs

Rui Tuo[*†] RUITUO@TAMU.EDU
Wm Michael Barnes '64 Department of Industrial and Systems Engineering
Texas A&M University
College Station, TX 77843, USA

Wenjia Wang[*] 1992WANGWENJIA@GMAIL.COM
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong

Editor: Philipp Hennig

Abstract

This work investigates the prediction performance of the kriging predictors. We derive some error bounds for the prediction error in terms of non-asymptotic probability under the uniform metric and L_p metrics when the spectral densities of both the true and the imposed correlation functions decay algebraically. The Matérn family is a prominent class of correlation functions of this kind. Our analysis shows that, when the smoothness of the imposed correlation function exceeds that of the true correlation function, the prediction error becomes more sensitive to the space-filling property of the design points. In particular, the kriging predictor can still reach the optimal rate of convergence, if the experimental design scheme is quasi-uniform. Lower bounds of the kriging prediction error are also derived under the uniform metric and L_p metrics. An accurate characterization of this error is obtained, when an oversmoothed correlation function and a space-filling design is used.

Keywords: Computer Experiments, Uncertainty Quantification, Scattered Data Approximation, Space-filling Designs, Bayesian Machine Learning

1. Introduction

In contemporary mathematical modeling and data analysis, we often face the challenge of reconstructing smooth functions from scattered observations. Gaussian process regression, also known as kriging, is a widely used approach. The main idea of kriging is to model the underlying function as a realization of a Gaussian process. This probabilistic model assumption endows the reconstructed function with a random distribution. Therefore, unlike the usual interpolation methods, kriging enables uncertainty quantification of the under-

∗. Two authors contributed equally to this work.
†. Tuo's work is supported by NSF grants DMS-1914636 and CCF-1934904.

©2020 Rui Tuo and Wenjia Wang.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v21/19-1045.html.

lying function in terms of its posterior distribution given the data. In spatial statistics and engineering, Gaussian processes are used to reflect the intrinsic randomness of the underlying functions or surfaces (Cressie, 1993; Stein, 1999; Matheron, 1963). In computer experiments, the Gaussian process models are adopted so that the prediction error under limited input data can be accessed (Santner et al., 2003; Sacks et al., 1989; Bayarri et al., 2007). For similar reasons, Gaussian process regression is applied in machine learning (Rasmussen, 2006) and probabilistic numerics (Hennig et al., 2015); specifically, in the area of Bayesian optimization, Gaussian process models are imposed and the probabilistic error of the reconstructed function are used to determine the next input point in a sequential optimization scheme for a complex black-box function (Shahriari et al., 2016; Frazier, 2018; Bull, 2011; Klein et al., 2017).

Under a Gaussian process model, the conditional distribution of the function value at an untried point given the data is normal, and can be expressed explicitly. In practice, we usually use the curve of conditional expectation as a surrogate model of the underlying function. Despite the known pointwise distributions, many basic properties of the kriging predictive curves remain as open problems. In this work, we focus on three fundamental aspects of kriging: 1) convergence of kriging predictive curves in function spaces; 2) robustness of kriging prediction against misspecification of the correlation functions; 3) effects of the design of experiments. Understanding the above properties of kriging can provide guidelines for choosing suitable correlation functions and experimental designs, which would potentially help the practical use of the method.

In this article, we focus on the isotropic Matérn correlation family. We suppose the underlying function is a random realization of a Gaussian process with an isotropic Matérn correlation function, and we reconstruct this function using kriging with a *misspecified* isotropic Matérn correlation function. We summarize our main results in Section 1.1. In Section 1.2, we make some remarks on related areas and research problems, and discuss the differences between the existing and the present results. In Section 2, we state our problem formulation and discuss the required technical conditions. Our main results are presented in Section 3. A simulation study is reported in Section 5, which assesses our theoretical findings regarding the effects of the experimental designs. Technical proofs are given in Section 7.

1.1 Summary of our results

We consider the reconstruction of a sample path of a Gaussian process over a compact set $\Omega \subset \mathbb{R}^d$. The shape of Ω can be rather general, subject to a few regularity conditions presented in Section 2.2. Table 1 shows a list of results on the rate of convergence of Gaussian process regression in the $L_p(\Omega)$ norm, with $1 \leq p \leq \infty$ under different designs and misspeficied correlation functions. Table 1 covers results on both the upper bounds and the lower bounds. The lower bounds are given in terms of the sample size n and the true smoothness ν_0; and the upper bounds depend also on the imposed smoothness ν, and two space-filling metrics of the design: the fill distance $h_{X,\Omega}$ and the mesh ratio $\rho_{X,\Omega}$. Details of the above notation are described in Section 2.2. The variance of the (stationary) Gaussian

process at each point is denoted as σ^2. Recall that we consider interpolation of Gaussian processes only, so there is no extra random error at each observed point given the Gaussian process sample path.

All results in Table 1 are obtained by the present work, except the shaded row which was obtained by our previous work (Wang et al., 2020). Compared to Wang et al. (2020), this work makes significant advances. First, this work establishes the convergence results when an oversmoothed correlation function is used, i.e., $\nu > \nu_0$. Specifically, the results in Wang et al. (2020) depends only on $h_{X,\Omega}$, and cannot be extended to oversmoothed correlations. In this work, we prove some new approximation results for radial basis functions (see Section 4), and establish the theoretical framework for oversmoothed correlations. In the present theory, the upper bounds in oversmoothed cases depend on both $h_{X,\Omega}$ and $\rho_{X,\Omega}$. We also present the bounds under the $L_p(\Omega)$ norms with $1 \leq p < \infty$ as well as the lower-bound-type results in this article.

Our findings in Table 1 lead to a remarkable result for the so-called *quasi-uniform sampling* (see Section 2.2). We show that under quasi-uniform sampling and oversmoothed correlation functions, the lower and upper rates coincide, which means that the optimal rates are achievable. This result also implies that the prediction performance does not deteriorate largely as an oversmoothed correlation function is imposed, provided that the experimental design scheme is quasi-uniform.

Case		Design	
		General design	Quasi-uniform design
$\nu \leq \nu_0$, $1 \leq p < \infty$	Upper rate	$\sigma h_{X,\Omega}^{\nu}$	$\sigma n^{-\nu/d}$
	Lower rate	$\sigma n^{-\nu_0/d}$	
$\nu \leq \nu_0$, $p = \infty$	Upper rate	$\sigma h_{X,\Omega}^{\nu} \log^{1/2}(1/h_{X,\Omega})$	$\sigma n^{-\nu/d}\sqrt{\log n}$
	Lower rate	$\sigma n^{-\nu_0/d}\sqrt{\log n}$	
$\nu > \nu_0$, $1 \leq p < \infty$	Upper rate	$\sigma h_{X,\Omega}^{\nu_0}\rho_{X,\Omega}^{\nu-\nu_0}$	$\sigma n^{-\nu_0/d}$
	Lower rate	$\sigma n^{-\nu_0/d}$	
$\nu > \nu_0$, $p = \infty$	Upper rate	$\sigma h_{X,\Omega}^{\nu_0}\rho_{X,\Omega}^{\nu-\nu_0} \log^{1/2}(1/h_{X,\Omega})$	$\sigma n^{-\nu_0/d}\sqrt{\log n}$
	Lower rate	$\sigma n^{-\nu_0/d}\sqrt{\log n}$	

Table 1: Summary of the L_p convergence rates for kriging prediction error with isotropic Matérn correlation functions. In addition to the rates of convergence, all kriging prediction errors in Table 1 decay at sub-Gaussian rates. The rates on the shaded row were presented in our previous work (Wang et al., 2020). The results for all other cases are obtained in the current work.

1.2 Comparison with related areas

Although the general context of function reconstruction is of interest in a broad range of areas, the particular settings of this work include: 1) **Random underlying function:** the underlying function is random and follows the law of a Gaussian process; 2) **Interpolation:** besides the Gaussian process, no random error is present, and therefore an interpolation scheme should be adopted; 3) **Misspecification:** Gaussian process regression is used to

reconstruct the underlying true function, and the imposed Gaussian process may have a misspecified correlation function; 4) **Scattered inputs:** the input points are fixed, with no particular structure. These features differentiate our objective from the existing areas of function reconstruction. In this section, we summarize the distinctions between the current work and four existing areas: average-case analysis of numerical problems, nonparametric regression, posterior contraction of Gaussian process priors, and scattered data approximation. Despite the differences in the scope, some of the mathematical tools in these areas are used in the present work, including a lower-bound result from the average-case analysis (Lemma 20), and some results from the scattered data approximation (see Section 4).

1.2.1 Average-case analysis of numerical problems

Among the existing areas, the average-case analysis of numerical problems has the closest model settings compared with ours, where the reconstruction of Gaussian process sample paths is considered. The primary distinction between this area and our work is the objective of the study: the average-case analysis aims at finding the optimal algorithms (which are generally *not* the Gaussian process regression, where a misspecified correlation can be used). In this work, we are interested in the *robustness* of the Gaussian process regression. Besides, the average-case analysis focuses on the optimal designs, while our study also covers general scattered designs.

One specific topic in the average-case analysis focuses on the following quantity,

$$e_p^{\text{avg}}(\phi, N) = \left(\int_{F_1} \|f - \phi(Nf)\|^p_{L_p(\Omega)} \mu(df) \right)^{1/p}, \tag{1}$$

where $\phi : N(F_1) \to L_p(\Omega)$ is an algorithm, $Nf = [f(x_1), ..., f(x_n)]$ with $x_i \in \Omega$, and F_1 is a function space equipped with Gaussian measure μ. It is worth noting that the results in the present work also imply some results in the form (1), where ϕ has to be a kriging algorithm. Specifically, Theorem 6 implies lower bounds of (1), and Corollary 9 shows that these lower bounds can be achieved, which also implies upper bounds of (1).

Results on the lower bounds of (1)**.** For $p = 2$, the lower bound was provided by Papageorgiou and Wasilkowski (1990); also see Lemma 20. If one further assumes that $\Omega = [0, 1]^d$, Proposition VI.8 of Ritter (2007) shows that the error (1) has a lower bound with the rate $n^{-\nu_0/d}$. One dimensional problems with correlation functions satisfying the Sacks-Ylvisaker conditions are extensively studied; see Müller-Gronbach and Ritter (1997); Ritter (2007); Ritter et al. (1995); Sacks and Ylvisaker (1966, 1968, 1970).

Results on the upper bound of (1)**.** Upper-bound-type results are pursued in average-case analysis under the optimal algorithm ϕ and optimal designs of $\{x_1, ..., x_n\}$. If $\Omega = [0, 1]^d$, Ritter (2007) shows that the rate $n^{-\nu_0/d}$ can be achieved by piecewise polynomial interpolation and specifically chosen designs; see Remark VI.3 of Ritter (2007), also see page 34 of Novak (2006) and Ivanov (1971).

For $1 \le p < \infty$ and the Matérn correlation function in one dimension, the error in average case $e_p^{\text{avg}}(\phi, N)$ can achieve the rate $n^{-\nu_0}$ by using piecewise polynomial interpolation; See

Proposition IV.36 of Ritter (2007). For the Matérn correlation function in one dimension, the quantity

$$e^{\text{avg}}_{L_\infty,p}(\phi, N) = \left(\int_{F_1} \|f - \phi(Nf)\|^p_{L_\infty(\Omega)} \mu(df) \right)^{1/p}, \tag{2}$$

can achieve the rate $n^{-\nu_0}\sqrt{\log n}$ by using Hermite interpolating splines (Buslaev and Seleznjev, 1999) for $1 \le p < \infty$.

Other definitions of the error are also studied in average-case analysis. See Chen and Wang (2019); Fasshauer et al. (2012); Khartov and Zani (2019); Lifshits and Zani (2015); Luschgy and Pagès (2004, 2007) for examples.

1.2.2 Nonparametric regression and statistical learning

The problem of interest in nonparameteric regression is to recover a *deterministic* function f under n *noisy* observations $(x_i, y_i), i = 1, \ldots, n$, under the model

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{3}$$

where ϵ_i's are the measurement error. Assuming that the function f has smoothness ν_0,[1] the optimal (minimax) rate of convergence is $n^{-\nu_0/(2\nu_0+d)}$ (Stone, 1982). A vast literature proposes and discusses methodologies regarding the nonparametric regression model (3), such as smoothing splines (Gu, 2013), kernel ridge regression (van de Geer, 2000), local polynomials (Tsybakov, 2008), etc. Because of the random noise, the rates for nonparametric regression are slower than those of the present work, as well as those in other interpolation problems. Some cross-cutting theory and approaches between regression and scattered data approximation are also discussed in the statistical learning literature; see, for example, Cucker and Zhou (2007).

1.2.3 Posterior contraction of Gaussian process priors

In this area, the model setting is similar to nonparametric regression, i.e., the underlying function is assumed *deterministic* and the observations are subject to random *noise*. The problem of interest is the posterior contraction of the Gaussian process prior. An incomplete list of papers in this area includes Castillo (2008, 2014); Giordano and Nickl (2019); Nickl and Söhl (2017); Pati et al. (2015); van der Vaart and van Zanten (2011, 2008a); van Waaij and van Zanten (2016). Despite the use of Gaussian process priors, to the best of our knowledge, the theory under this framework does not consider noiseless observations, and no error bounds in terms of the our settings, i.e., the fill and separation distances, are reported in this area.

1. See Section 4 for a discussion on the smoothness of a deterministic function.

1.2.4 Scattered data approximation

In the field of scattered data approximation, the goal is to approximate, or often, interpolate a *deterministic* function f with its exact observations $f(x_i), i = 1, ..., n$, where x_i's are data sites. For function f with smoothness m, the L_p convergence rate is $n^{-m/d+(1/2-1/q)_+}$ for $1 \leq p \leq \infty$ (Wendland, 2004), where a_+ stands for $\max\{a, 0\}$. A sharper characterization of the upper bounds are related to the fill distance and separation distance of the design points. Although this area focuses on a purely deterministic problem, some of the results in this field will serve as the key mathematical tool in this work.

It is worth noting that the existing research in scattered data approximation also covered the circumstances where the underlying function is rougher than the kernel function, so that the function is outside of the reproducing kernel Hilbert space generated by the kernel. See Narcowich et al. (2006) for example. Such results can be interpreted as using "misspecified" kernels in interpolating deterministic functions. More discussions are deferred to Section 4.

2. Problem formulation

In this section we discuss the interpolation method considered in this work, and the required technical conditions.

2.1 Background

Let $Z(x)$ be an underlying Gaussian process, with $x \in \mathbb{R}^d$. We suppose $Z(\cdot)$ is a stationary Gaussian process with mean zero. The covariance function of Z is denoted as

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

where σ^2 is the variance, and Ψ is the correlation function, or kernel, satisfying $\Psi(0) = 1$. The correlation function Ψ is a symmetric positive semi-definite function on $\mathbb{R}^d$. Since we are interested in interpolation, we require that $Z(\cdot)$ is mean square continuous, or equivalently, Ψ is continuous on $\mathbb{R}^d$. Then it follows from the Bochner's theorem (Gihman and Skorokhod, 1974, page 208; Wendland, 2004, Theorem 6.6) that, there exists a finite nonnegative Borel measure F_Ψ on $\mathbb{R}^d$, such that

$$\Psi(x) = \int_{\mathbb{R}^d} e^{i\omega^T x} F_\Psi(d\omega). \tag{4}$$

In particular, we are interested in the case where Ψ is also positive definite and integrable on $\mathbb{R}^d$. In this case, it can be proven that F_Ψ has a density with respect to the Lebesgue measure. See Theorem 6.11 of Wendland (2004). The density of F_Ψ, denoted as f_Ψ, is known as the *spectral density* of Z or Ψ.

In this work, we suppose that f_Ψ decays algebraically. A prominent class of correlation functions of this type is the isotropic Matérn correlation family (Santner et al., 2003; Stein,

1999), given by

$$\Psi(x;\nu,\phi) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(2\sqrt{\nu}\phi\|x\|)^{\nu}K_{\nu}(2\sqrt{\nu}\phi\|x\|), \tag{5}$$

with the spectral density (Tuo and Wu, 2016)

$$f_{\Psi}(\omega;\nu,\phi) = \pi^{-d/2}\frac{\Gamma(\nu+d/2)}{\Gamma(\nu)}(4\nu\phi^2)^{\nu}(4\nu\phi^2+\|\omega\|^2)^{-(\nu+d/2)}, \tag{6}$$

where $\phi,\nu>0$, K_ν is the modified Bessel function of the second kind and $\|\cdot\|$ denotes the Euclidean metric. It is worth noting that (6) is bounded above and below by $(1+\|\omega\|^2)^{-(\nu+d/2)}$ multiplied by two constants, respectively. The parameter ν for the Matérn kernels is called the *smoothness* parameter, as it governs the smoothness (or differentiability) of the Gaussian processes. Further discussions are deferred to Section 4.

Another example of correlation functions with algebraically decayed spectral densities is the generalized Wendland correlation function (Wendland, 2004; Gneiting, 2002; Chernih and Hubbert, 2014; Bevilacqua et al., 2019; Fasshauer and McCourt, 2015), defined as

$$\Psi_{GW}(x) = \begin{cases} \frac{1}{B(2\kappa,\mu+1)}\int_{\|\phi x\|}^{1} u(u^2-\|\phi x\|^2)^{\kappa-1}(1-u)^{\mu}du, & 0\le\|x\|<\frac{1}{\phi},\\ 0, & \|x\|\ge\frac{1}{\phi}, \end{cases}$$

where $\phi,\kappa>0$ and $\mu\ge(d+1)/2+\kappa$, and B denotes the beta function. See Theorem 1 of Bevilacqua et al. (2019).

Now we consider the interpolation problem. Suppose we have a scattered set of points $X=\{x_1,\dots,x_n\}\subset\Omega$. Here the set Ω is the region of interest, which is a subset of $\mathbb{R}^d$. The goal of kriging is to recover $Z(x)$ given the observed data $Z(x_1),\dots,Z(x_n)$. A standard predictor is the best linear predictor (Santner et al., 2003; Stein, 1999), given by the conditional expectation of $Z(x)$ on $Z(x_1),\dots,Z(x_n)$, as

$$\mathbb{E}[Z(x)|Z(x_1),\dots,Z(x_n)] = r_{\Psi}^T(x)K_{\Psi}^{-1}Y, \tag{7}$$

where $r_\Psi(x)=(\Psi(x-x_1),\dots,\Psi(x-x_n))^T, K_\Psi=(\Psi(x_j-x_k))_{jk}$ and $Y=(Z(x_1),\dots,Z(x_n))^T$.

The best linear predictor in (21) depends on the correlation function Ψ. However, in practice Ψ is commonly unknown. Thus, we may inevitably use a misspecified correlation function, denoted by Φ. Suppose that Φ has a spectral density f_Φ. We also suppose that f_Φ decays algebraically, but the decay rate of f_Φ can differ from that of f_Ψ.

We consider the predictor given by the right-hand side of (21), in which the true correlation function Ψ is replaced by the misspecified correlation function Φ. Clearly, such a predictor is no longer the best linear predictor. Nevertheless, it still defines an interpolant, denoted by

$$\mathcal{I}_{\Phi,X}Z(x) = r_{\Phi}^T(x)K_{\Phi}^{-1}Y, \tag{8}$$

where $r_\Phi(x)=(\Phi(x-x_1),\dots,\Phi(x-x_n))^T, K_\Phi=(\Phi(x_j-x_k))_{jk}$ and $Y=(Z(x_1),\dots,Z(x_n))^T$. In (8), $\mathcal{I}_{\Phi,X}$ denotes the interpolation operator given by the kriging predictor, which can be applied not only to a Gaussian process, but also to a deterministic function in the same vein.

2.2 Notation and conditions

We do not assume any particular structure of the design points $X = \{x_1, \ldots, x_n\}$. Our error estimate for the kriging predictor will depend on two dispersion indices of the design points.

The first one is the fill distance, defined as

$$h_{X,\Omega} := \sup_{x \in \Omega} \inf_{x_j \in X} \|x - x_j\|.$$

The second is the separation radius, given by

$$q_X := \min_{1 \le j \ne k \le n} \|x_j - x_k\|/2.$$

It is easy to check that $h_{X,\Omega} \ge q_X$ (Wendland, 2004). Define the mesh ratio $\rho_{X,\Omega} := h_{X,\Omega}/q_X \ge 1$. Because we are only interested in the prediction error when the design points are sufficiently dense, for notational simplicity, we assume that $h_{X,\Omega} < 1$. In the rest of this paper, we use the following conventions. For two positive sequences a_n and b_n, we write $a_n \asymp b_n$ if, for some constants $C, C' > 0$, $C \le a_n/b_n \le C'$ for all n, and write $a_n \gtrsim b_n$ if $a_n \ge C b_n$ for some constant $C > 0$. Let $\mathrm{card}(X)$ denote the cardinality of set X.

In this work, we consider both the non-asymptotic case, i.e., the design X is fixed, and the asymptotic case, i.e., the number of design points increases to infinity. To state the asymptotic results, suppose we have a sequence of designs with increasing number of points, denoted by $\mathcal{X} = \{X_1, X_2, \ldots\}$. We regard $\mathcal{X}$ as a *sampling scheme* which generates a sequence of designs, for instance, a design sequence generated by random sampling or maximin Latin hypercube designs.

Without loss of generality, assume that $\mathrm{card}(X_n) = n$, where n takes its value in an infinite subset of $\mathbb{N}$. This assumption enables direct comparison between our upper and lower bounds. Given the sampling scheme $\mathcal{X}$, we denote $h_n := h_{X_n,\Omega}, q_n := q_{X_n}$ and $\rho_n = h_n/q_n$. For any sampling scheme, it can be shown that $h_n \gtrsim n^{-1/d}$ and $q_n \lesssim n^{-1/d}$ (Borodachov et al., 2007; Joseph et al., 2015). In fact, it is possible to have $h_n \asymp q_n \asymp n^{-1/d}$, if and only if ρ_n is uniformly bounded above by a constant (Müller, 2009).

Definition 1 *A sampling scheme $\mathcal{X}$ is said quasi-uniform if there exists a constant $C > 0$ such that $\rho_n \le C$ for all n.*

It is not hard to find a quasi-uniform sampling scheme. For example, a hypercube grid sampling in $\Omega = [0,1]^d$ is quasi-uniform because $\rho_n = \sqrt{d}$ is a constant (Wendland, 2004). However, random samplings do not belong to the quasi-uniform class; see Example 1 in Section 3.3.

We assume the Conditions 1-3 throughout this article.

Definition 2 *A set $\Omega \subset \mathbb{R}^d$ is said to satisfy an interior cone condition if there exists an angle $\alpha \in (0, \pi/2)$ and a radius $R > 0$ such that for every $x \in \Omega$, a unit vector $\xi(x)$ exists such that the cone*

$$C(x, \xi(x), \alpha, R) := \left\{ x + \lambda y : y \in \mathbb{R}^d, \|y\| = 1, y^T \xi(x) \geq \cos\alpha, \lambda \in [0, R] \right\}$$

is contained in Ω.

Condition 1 *The experimental region $\Omega \subset \mathbb{R}^d$ is a compact set with Lipschitz boundary and satisfies an interior cone condition.*

Condition 2 *There exist constants $c_2 \geq c_1 > 0$ and $\nu_0 > 0$ such that, for all $\omega \in \mathbb{R}^d$,*

$$c_1(1 + \|\omega\|^2)^{-(\nu_0 + d/2)} \leq f_\Psi(\omega) \leq c_2(1 + \|\omega\|^2)^{-(\nu_0 + d/2)}.$$

Condition 3 *There exist constants $c_4 \geq c_3 > 0$ and $\nu > 0$ such that, for all $\omega \in \mathbb{R}^d$,*

$$c_3(1 + \|\omega\|^2)^{-(\nu + d/2)} \leq f_\Phi(\omega) \leq c_4(1 + \|\omega\|^2)^{-(\nu + d/2)}.$$

Condition 1 is a geometric condition on the experimental region Ω, which holds in most practical situations, because the commonly encountered experimental regions, like the rectangles or balls, satisfy interior cone conditions. Figure 1 (page 258 of Roy and Couchman (2001)) is an illustration of the α-interior cone condition.

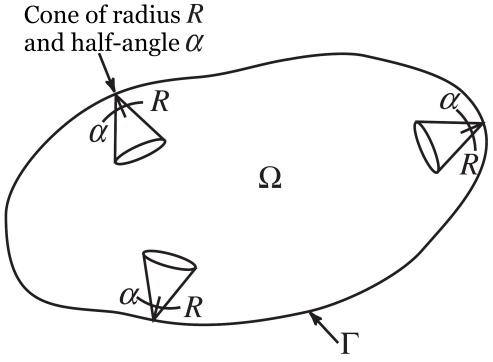


Figure 1: An illustration of an interior cone condition (page 258 of Roy and Couchman (2001)).

Conditions 2 and 3 require that the spectral densities decay in an algebraic order. For example, if Ψ and Φ are Matérn correlation functions with smoothness parameter ν_0 and ν, respectively, they satisfy Conditions 2 and 3. The decay rates in Conditions 2 and 3 determine the smoothness of the correlation function Ψ and Φ; see Section 4 for the discussion of the relation between the smoothness of the correlation functions and the smoothness of the sample path of a Gaussian process.

3. Main results

In this section, we present our main theoretical results on the prediction error of kriging.

3.1 Upper and lower bounds of the uniform kriging prediction error

This work aims at studying the prediction error of the kriging algorithm (8), i.e., $|Z(x) - \mathcal{I}_{\Phi,X}Z(x)|$. In this subsection, we consider the prediction error of the kriging algorithm (8) under a uniform metric, given by

$$\sup_{x\in\Omega} |Z(x) - \mathcal{I}_{\Phi,X}Z(x)|, \tag{9}$$

which was considered previously in Wang et al. (2020). Under Conditions 1-3, they derived an upper bound of (9) under the case $\nu \le \nu_0$. This result is shown in Theorem 1 for the completeness of work. Here we are interested in the case $\nu > \nu_0$, that is, the imposed correlation function is smoother than the true correlation function. In Theorem 2, we provide an upper bound of the prediction error for $\nu > \nu_0$. In addition to the upper bounds, we obtain a lower bound of the uniform kriging prediction error in Theorem 3.

Theorem 1 *Suppose Conditions 1-3 hold and $\nu \le \nu_0$. Then there exist constants $C_1, C_2 > 0, C_3 > e$ and $h_0 \in (0,1]$, such that for any design X with $h_{X,\Omega} \le h_0$ and any $t > 0$, with probability at least $1 - \exp\{-t^2/(C_1\sigma^2 h_{X,\Omega}^{2\nu})\}$,[2] the kriging prediction error has the upper bound*

$$\sup_{x\in\Omega} |Z(x) - \mathcal{I}_{\Phi,X}Z(x)| \le C_2\sigma h_{X,\Omega}^{\nu} \log^{1/2}(C_3/h_{X,\Omega}) + t.$$

Here the constants C_1, C_2, C_3 depend only on Ω, Φ, and Ψ, including ν and ν_0.

Theorem 2 *Suppose Conditions 1-3 hold and $\nu > \nu_0$. Then there exist constants $C_1, C_2 > 0, C_3 > e$ and $h_0 \in (0,1]$, such that for any design X with $h_{X,\Omega} \le h_0$ and any $t > 0$, with probability at least $1 - \exp\{-t^2/(C_1\sigma^2 h_{X,\Omega}^{2\nu_0}\rho_{X,\Omega}^{2(\nu-\nu_0)})\}$, the kriging prediction error has the upper bound*

$$\sup_{x\in\Omega} |Z(x) - \mathcal{I}_{\Phi,X}Z(x)| \le C_2\sigma h_{X,\Omega}^{\nu_0}\rho_{X,\Omega}^{\nu-\nu_0} \log^{1/2}(C_3/h_{X,\Omega}) + t.$$

Here the constants C_1, C_2, C_3 depend only on Ω, Φ, and Ψ, including ν and ν_0.

Theorem 3 *Suppose Conditions 1-3 hold. Then there exist constants $C_1, C_2 > 0$, such that for any design X satisfying $\mathrm{card}(X) = n$ and any $t > 0$, with probability at least $1 - \exp\{-t^2/(2C_1\sigma^2 A)\}$, the kriging prediction error has the lower bound*

$$\sup_{x\in\Omega} |Z(x) - \mathcal{I}_{\Phi,X}Z(x)| \ge C_2\sigma n^{-\frac{\nu_0}{d}}\sqrt{\log n} - t,$$

where $A = h_{X,\Omega}^{2\nu}$ if $\nu \le \nu_0$, and $A = h_{X,\Omega}^{2\nu_0}\rho_{X,\Omega}^{2(\nu-\nu_0)}$ if $\nu > \nu_0$. Here the constants $C_1, C_2 > 0$ depend only on Ω, Φ, and Ψ, including ν and ν_0.

2. In Wang et al. (2020), this probability is $1 - 2\exp\{-t^2/(C_1\sigma^2 h_{X,\Omega}^{2\nu})\}$. The constant two can be removed by applying a different version of the Borell-TIS inequality given by Lemma 21 in Section 7.2.

3.2 Bounds for the L_p norms of the kriging prediction error

Now we consider the L_p norm of the kriging prediction error, given by

$$\|Z - \mathcal{I}_{\Phi,X} Z\|_{L_p(\Omega)} := \left(\int_\Omega |Z(x) - \mathcal{I}_{\Phi,X} Z(x)|^p dx \right)^{1/p}, \tag{10}$$

with $1 \le p < \infty$. The upper bounds of the L_p norms of the kriging prediction error with undersmoothed and oversmoothed correlation functions are provided in Theorems 4 and 5, respectively.

Theorem 4 *Suppose Conditions 1-3 hold and $\nu \le \nu_0$. Then there exist constants $C_1, C_2 > 0$ and $h_0 \in (0, 1]$, such that for any design X with $h_{X,\Omega} \le h_0$ and any $t > 0$, with probability at least $1 - \exp\{-t^2/(C_1 \sigma^2 h_{X,\Omega}^{2\nu})\}$, the kriging prediction error has the upper bound*

$$\|Z - \mathcal{I}_{\Phi,X} Z\|_{L_p(\Omega)} \le C_2 \sigma h_{X,\Omega}^{\nu} + t.$$

The constants C_1, C_2 depend only on Ω, p, Φ, and Ψ, including ν and ν_0.

Theorem 5 *Suppose Conditions 1-3 hold and $\nu > \nu_0$. Then there exist constants $C_1, C_2 > 0$ and $h_0 \in (0, 1]$, such that for any design X with $h_{X,\Omega} \le h_0$ and any $t > 0$, with probability at least $1 - \exp\{-t^2/(C_1 \sigma^2 h_{X,\Omega}^{2\nu_0} \rho_{X,\Omega}^{2(\nu-\nu_0)})\}$, the kriging prediction error has the upper bound*

$$\|Z - \mathcal{I}_{\Phi,X} Z\|_{L_p(\Omega)} \le C_2 \sigma h_{X,\Omega}^{\nu_0} \rho_{X,\Omega}^{\nu-\nu_0} + t.$$

Here the constants C_1, C_2 depend only on Ω, p, Φ, and Ψ, including ν and ν_0.

Regarding the lower prediction error bounds under the L_p norm, we obtain a result analogous to Theorem 3. Theorem 6 suggests a lower bound under the L_p norm, which differs from that in Theorem 3 only by a $\sqrt{\log n}$ factor.

Theorem 6 *Suppose Conditions 1-3 hold. There exist constants $C_1, C_2 > 0$, such that for any design X satisfying $\mathrm{card}(X) = n$ and any $t > 0$, with probability at least $1 - 2\exp\{-t^2/(2C_1\sigma^2 A)\}$, the kriging prediction error has the lower bound*

$$\|Z - \mathcal{I}_{\Phi,X} Z\|_{L_p(\Omega)} \ge C_2 \sigma n^{-\frac{\nu_0}{d}} - t$$

for $1 \le p < \infty$, where $A = h_{X,\Omega}^{2\nu}$ if $\nu \le \nu_0$, and $A = h_{X,\Omega}^{2\nu_0} \rho_{X,\Omega}^{2(\nu-\nu_0)}$ if $\nu > \nu_0$. Here the constants $C_1, C_2 > 0$ depend only on Ω, p, Φ, and Ψ, including ν and ν_0.

The results in Theorems 1, 2, 4 and 5 are presented in a non-asymptotic manner, i.e., the design X is fixed. The asymptotic results, which are traditionally of interest in spatial statistics, can be inferred from these non-asymptotic results. Here we consider the so-called fixed-domain asymptotics (Stein, 1999; Loh, 2005), in which the domain Ω is kept unchanged and the design points become dense over Ω.

We collect the asymptotic rates analogous to the upper bounds in Corollaries 7 and 8. Their proofs are straightforward.

Corollary 7 *Suppose Conditions 1-3 hold. In addition, we suppose the sampling scheme $\mathcal{X}$ is asymptotically dense over Ω, that is, $h_n \to 0$ as $n \to \infty$. We further assume $h_n^{\nu_0}\rho_n^{(\nu-\nu_0)} \to 0$ if $\nu > \nu_0$. Then the uniform kriging prediction error has the order of magnitude*

$$\sup_{x\in\Omega} |Z(x) - \mathcal{I}_{\Phi,X_n} Z(x)| = \begin{cases} O_{\mathbb{P}}\left(h_n^{\nu} \log^{1/2}(1/h_n)\right) & \text{if } \nu \le \nu_0, \\ O_{\mathbb{P}}\left(h_n^{\nu_0}\rho_n^{\nu-\nu_0} \log^{1/2}(1/h_n)\right) & \text{if } \nu > \nu_0. \end{cases}$$

Corollary 8 *Under the conditions of Corollary 7, for $1 \le p < \infty$, the kriging prediction error has the order of magnitude in $L_p(\Omega)$*

$$\|Z(x) - \mathcal{I}_{\Phi,X_n} Z(x)\|_{L_p(\Omega)} = \begin{cases} O_{\mathbb{P}}\left(h_n^{\nu}\right) & \text{if } \nu \le \nu_0, \\ O_{\mathbb{P}}\left(h_n^{\nu_0}\rho_n^{\nu-\nu_0}\right) & \text{if } \nu > \nu_0. \end{cases}$$

From Corollaries 7 and 8, we find that the upper bounds of kriging prediction error strongly depend on the sampling scheme $\mathcal{X}$.

If a sampling scheme is quasi-uniform and $\nu \ge \nu_0$, then the orders of magnitude in Corollaries 7 and 8 agree with the lower bounds in Theorems 3 and 6, respectively, implying that these bounds are sharp. We summarize the results in Corollary 9.

Corollary 9 *Suppose Conditions 1-3 hold and $\nu \ge \nu_0$. In addition, we suppose the sampling scheme $\mathcal{X}$ is quasi-uniform. Then the kriging prediction error has the exact order of magnitude*

$$\sup_{x\in\Omega} |Z(x) - \mathcal{I}_{\Phi,X_n} Z(x)| \asymp n^{-\nu_0/d} \log^{1/2} n,$$
$$\|Z(x) - \mathcal{I}_{\Phi,X_n} Z(x)\|_{L_p(\Omega)} \asymp n^{-\nu_0/d}, \quad 1 \le p < \infty.$$

3.3 An example

We illustrate the impact of the experimental designs in Example 1.

Example 1 *The random sampling in $[0,1]$ is not quasi-uniform. To see this, let $x_1, \ldots, x_n$ be mutually independent random variables following the uniform distribution on $[0,1]$. Denote their order statistics as*

$$0 = x_{(0)} \le x_{(1)} \le \cdots \le x_{(n)} \le x_{(n+1)} = 1.$$

Clearly, we have

$$\rho_n = \frac{\max_{0\le j\le n} |x_{(j+1)} - x_{(j)}|}{\min_{0\le j\le n} |x_{(j+1)} - x_{(j)}|}.$$

Let $y_1, \ldots, y_n, y_{n+1}$ be mutually independent random variables following the exponential distribution with mean one. It is well known that $(x_{(1)}, \ldots, x_{(n)})$ has the same distribution as

$$\left(\frac{y_1}{\sum_{j=1}^{n+1} y_j}, \ldots, \frac{\sum_{j=1}^{n} y_j}{\sum_{j=1}^{n+1} y_j}\right).$$

Thus ρ_n has the same distribution as $\max y_j / \min y_j$. Clearly, $\max y_j \asymp \log n$ and $\min y_j \asymp 1/n$. This implies $\rho_n \asymp n \log n$. Similarly, we can see that h_n has the same distribution as $\max y_j / \sum_{k=1}^{n+1} y_k$, which is of the order $O_{\mathbb{P}}(n^{-1} \log n)$. See Appendix A for proofs of the above statements.

Now consider the kriging predictive curve under $\Omega = [0,1]$ and random sampled design points and an oversmoothed correlation, i.e., $\nu > \nu_0$. According to Corollary 7, its uniform error has the order of magnitude $O_{\mathbb{P}}(n^{\nu-2\nu_0} \log^{\nu+1/2} n)$, which decays to zero if $\nu < 2\nu_0$.

In Section 5, we will conduct simulation studies to verify our theoretical assertions on the rates of convergence in this example. It can be seen from Table 2 in Section 5 that the numerical results agree with our theory.

4. Discussion on a major mathematical tool and the notion of smoothness

The theory of radial basis function approximation is an essential mathematical tool for developing the bounds in this work, as well as those in our previous work Wang et al. (2020). We refer to Wendland (2004) for an introduction of the radial basis function approximation theory.

A primary objective of the radial basis function approximation theory is to study the approximation error

$$g - \mathcal{I}_{\Phi,X} g,$$

for a deterministic function g. Here we consider the circumstance that g lies in a (fractional) Sobolev space.

Our convention of the Fourier transform is $\hat{g}(\omega) = \int_{\mathbb{R}^d} g(x) e^{-i\omega^T x} dx$. Regarding the Fourier transform as a mapping $\hat{g} : L_1(\mathbb{R}^d) \cap L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d)$, we can uniquely extend it to a mapping $\hat{g} : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d)$ (Wendland, 2004). The norm of the (fractional) Sobolev space $W_2^\beta(\mathbb{R}^d)$ for a real number $\beta > 0$ (also known as the Bessel potential space) is

$$\|g\|^2_{W_2^\beta(\mathbb{R}^d)} = \int_{\mathbb{R}^d} |\hat{g}(\omega)|^2 (1 + \|\omega\|^2)^\beta d\omega,$$

for $g \in L_2(\mathbb{R}^d)$.

Remark 10 *An equivalent norm of the Sobolev space $W_2^\beta(\mathbb{R}^d)$ for $\beta \in \mathbb{N}$ can be defined via derivatives. For $\alpha = (\alpha_1, \ldots, \alpha_d)^T \in \mathbb{N}_0^d$, we shall use the notation $|\alpha| = \sum_{j=1}^d \alpha_j$. For $x = (x_1, \ldots, x_d)^T$, denote*

$$D^\alpha g = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} g \qquad \text{and} \qquad x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d}.$$

Define $\|g\|'_{W_2^\beta(\mathbb{R}^d)} = \left(\sum_{|\alpha| \le \beta} \|D^\alpha g\|^2_{L_2(\mathbb{R}^d)} \right)^{\frac{1}{2}}$. It can be shown that $\|\cdot\|'_{W_2^\beta(\mathbb{R}^d)}$ and $\|\cdot\|_{W_2^\beta(\mathbb{R}^d)}$ are equivalent for $\beta \in \mathbb{N}$ (Adams and Fournier, 2003).

The classic framework on the error analysis for radial basis function approximation employs the reproducing kernel Hilbert spaces (RKHS, see Section 7.1 for more details) as a necessary mathematical tool. The development of Wang et al. (2020) relies on these classic results. These results, however, are not applicable in the current context when f_Ψ/f_Φ is not uniformly bounded.

The current research is partially inspired by the "escape theorems" for radial basis function approximation established by Brownlee and Light (2004); Narcowich et al. (2005); Narcowich (2005); Narcowich and Ward (2002, 2004); Narcowich et al. (2006). These works show that, some radial basis functions interpolants still provide effective approximation, even if the underlying functions are too rough to lie in the corresponding RKHS.

Our results on interpolation of Gaussian processes with oversmoothed kernels are based on an escape theorem, given by Lemma 11. Given Condition 3, it is known that the RKHS generated by Φ is equivalent to $W_2^{\nu+d/2}(\mathbb{R}^d)$ (see Lemma 13 in Section 7.1), which is a proper subset of $W_2^{\nu_0+d/2}(\mathbb{R}^d)$ when $\nu_0 < \nu$. Lemma 11 shows that the radial basis function approximation may still give reasonable error bounds even if the underlying function does not lie in the RKHS.

Lemma 11 *Let Φ be a kernel with a spectral density f_Φ satisfying Condition 3, and g be a function in $W_2^{\nu_0+d/2}(\mathbb{R}^d)$ with $\nu \geq \nu_0 > 0$. Suppose $\Omega \subset \mathbb{R}^d$ is a domain satisfying Condition 1. Then there exist constants $C > 0$ and $h_0 \in (0, 1]$ such that for any design X with $h_{X,\Omega} \leq h_0$, we have*

$$\sup_{x\in\Omega} |g(x) - \mathcal{I}_{\Phi,X} g(x)| \leq C h_{X,\Omega}^{\nu_0} \rho_{X,\Omega}^{\nu-\nu_0} \|g\|_{W_2^{\nu_0+d/2}(\mathbb{R}^d)}. \tag{11}$$

Here the constant C depends only on Ω, Φ and Ψ, including ν and ν_0.

Theorem 4.2 of Narcowich et al. (2006) states that under the conditions of Lemma 11 in addition to

$$\lfloor \beta \rfloor > d/2, \tag{12}$$

we have

$$\|g - \mathcal{I}_{\Phi,X} g\|_{W_2^\mu(\Omega)} \leq C h_{X,\Omega}^{\beta-\mu} \rho_{X,\Omega}^{\tau-\beta} \|g\|_{W_2^\beta(\mathbb{R}^d)}, \tag{13}$$

for $0 \leq \mu \leq \beta$. As commented by a reviewer, condition (12) can be removed by using Theorem 4.1 of Arcangéli et al. (2007) in the proof of Theorem 4.2 of Narcowich et al. (2006); also see Theorem 10 of Wynne et al. (2020). Having (13), Lemma 11 is an immediate consequence. Specifically, combining (13) with the real interpolation theory for Sobolev spaces (See, e.g., Theorem 5.8 and Chapter 7 of Adams and Fournier (2003)), yields (11). An alterative proof of Lemma 11, also suggested by a reviewer, is given in Section 7.1.1.

Next we make a remark on the notion of smoothness and the settings of smoothness misspecification. For a deterministic function g, we say g has smoothness $\nu_0+d/2$ if $g \in W^{\nu_0+d/2}(\Omega)$.

The smoothness misspecification in Lemma 11 is stated as: the smoothness associated with the RKHS is higher than the true smoothness of the function when $\nu_0 < \nu$.

Now we turn to the role of ν_0 for a stationary Gaussian process $Z(x)$ with spectral density f_Ψ satisfying Condition 2. Unlike the usual perception on the smoothness of deterministic functions, here ν_0 should be interpreted as the mean squared differentiability (Stein, 1999) of the Gaussian process, which is related to the smoothness of the correlation function Ψ.

On the other hand, we can also consider the smoothness of sample paths of $Z(x)$, under the usual definition of smoothness for deterministic functions. It turns out that the sample path smoothness is lower than ν_0 with probability one (Driscoll, 1973; Steinwart, 2019; Kanagawa et al., 2018). In view of this, Theorem 2 implies that the sample paths of Gaussian processes can escape the $d/2$ smoothness misspecification in terms of the L_∞ norm, disregarding the logarithmic factor. In other words, there exist functions with smoothness less than ν_0 that can be approximated at the rate $O(n^{-\nu_0/d}\sqrt{\log n})$, and the set of such functions is large under the probability measure of a certain Gaussian process.

5. Simulation studies

The objective of this section is to verify whether the rate of convergence given by Corollary 7 is accurate. We consider the settings in Example 1. We have shown that under a random sampling over the experimental region $\Omega = [0, 1]$, the kriging prediction error has the rate $O_{\mathbb{P}}(n^{\nu-2\nu_0} \log^{\nu+1/2} n)$ for $\nu > \nu_0$. If grid sampling is used, Corollaries 7 and 9 show that the error has the order of magnitude $n^{-\nu_0} \log^{1/2} n$ for $\nu > \nu_0$.

We denote the expectation of (9) with random sampling and grid sampling by $\mathcal{E}_{\mathrm{rand}}$ and $\mathcal{E}_{\mathrm{grid}}$, respectively. Our idea of assessing the rate of convergence is described as follows. If the error rates are sharp, we have the approximations

$$\log \mathcal{E}_{\mathrm{rand}} \approx (\nu - 2\nu_0) \log n + (\nu + \frac{1}{2}) \log \log n + \log c_1,$$
$$\log \mathcal{E}_{\mathrm{grid}} \approx -\nu_0 \log n + \frac{1}{2} \log \log n + \log c_2,$$

for random samplings and grid samplings, respectively, where c_1, c_2 are constants. Since $\log \log n$ grows much slower than $\log n$, we can regard the $\log \log n$ term as a constant and get the second approximations

$$\log \mathcal{E}_{\mathrm{rand}} \approx (2\nu_0 - \nu) \log(1/n) + C_1, \tag{14}$$
$$\log \mathcal{E}_{\mathrm{grid}} \approx \nu_0 \log(1/n) + C_2. \tag{15}$$

To verify the above formulas via numerical simulations, we can regress $\log \mathcal{E}_{\mathrm{rand}}$ and $\log \mathcal{E}_{\mathrm{grid}}$ on $\log(1/n)$ and examine the estimated slopes. If the bounds are sharp, the estimated slopes should be close to the theoretical assertions $2\nu_0 - \nu$ and ν_0, respectively.

In our simulation studies, we consider the sample sizes $n = 10k$, for $k = 2, 3, ..., 15$. For each k, we simulate 100 realizations of a Gaussian process. For a specific realization of a Gaussian process, we generate $10k$ independent and uniformly distributed random points as X, and

use $\sup_{x\in\Omega_1}|Z(x)-\mathcal{I}_{\Phi,X}Z(x)|$ to approximate the uniform error $\sup_{x\in\Omega}|Z(x)-\mathcal{I}_{\Phi,X}Z(x)|$, where Ω_1 is the first 200 points of the Halton sequence (Niederreiter, 1992). We believe that the points are dense enough so that the approximation can be accurate. Then the regression coefficient is estimated using the least squares method. For grid sampling, we adopt a similar approach with the same number of design points X. The results are presented in Table 2. The first two columns of Table 2 show the true and imposed smoothness parameters of the Matérn correlation functions. The fourth and the fifth columns show the convergence rates obtained from the simulation studies and the theoretical analysis, respectively. The sixth column shows the relative difference between the fourth and the fifth columns, given by |estimated slope-theoretical slope|/(theoretical slope). The last column gives the R-squared values of the linear regression of the simulated data.

ν_0	ν	Design	ES	TS	RD	R^2
1.1	1.3	RS	0.9011	0.9	0.0012	0.8579
		GS	1.0670	1.1	0.0300	0.9992
1.1	2.8	RS	0.1653	-0.6 (No convergence)	-	0.0308
		GS	1.0968	1.1	0.0030	0.9995
2.1	2.8	RS	1.523	1.4	0.088	0.9834
		GS	2.0953	2.1	0.0022	0.9992
1.5	3.5	RS	0.1083	-0.5 (No convergence)	-	0.0991
		GS	1.4982	1.5	0.0012	0.9989

Table 2: Numerical studies on the convergence rates of kriging prediction with oversmoothed correlation functions. The following abbreviations are used: RS=Random sampling, GS= Grid sampling, ES=Estimated slope, TS=Theoretical slope, RD=relative difference. The relative differences are not computed when the corresponding theoretical slopes are negative.

In the setting of Rows 2, 3, 5-7 and 9 of Table 2, our theory suggests the prediction consistency, i.e., $h_n^{\nu_0}\rho_n^{\nu-\nu_0}$ tends to zero. It can be seen that the estimated slopes coincide with our theoretical assertions for these cases. Also, the R-squared values for these rows are high, which implies a good model fitting of (14)-(15). When $h_n^{\nu_0}\rho_n^{\nu-\nu_0}$ goes to infinity, our simulation results suggest a very slow rate of convergence. Specifically, under the random sampling scheme and $(\nu_0,\nu)=(1.1,2.8)$ and $(\nu_0,\nu)=(1.5,3.5)$, the estimated rates of convergence are 0.1653 and 0.1083, respectively. Also, the R-squared values are very low. These slow rates and poor model fitting imply that the kriging predictor could be inconsistent. Figure 2 shows the scattered plots of the raw data and the regression lines under the four combinations of (ν_0,ν) in Table 2.

6. Concluding remarks

The error bounds presented in this work are not only valuable in mathematics. They can also provide guidelines for practitioners of kriging. Especially, our work confirms the importance of the design of experiments for kriging: if the design is quasi-uniform, the use of an oversmoothed correlation would not be an issue.

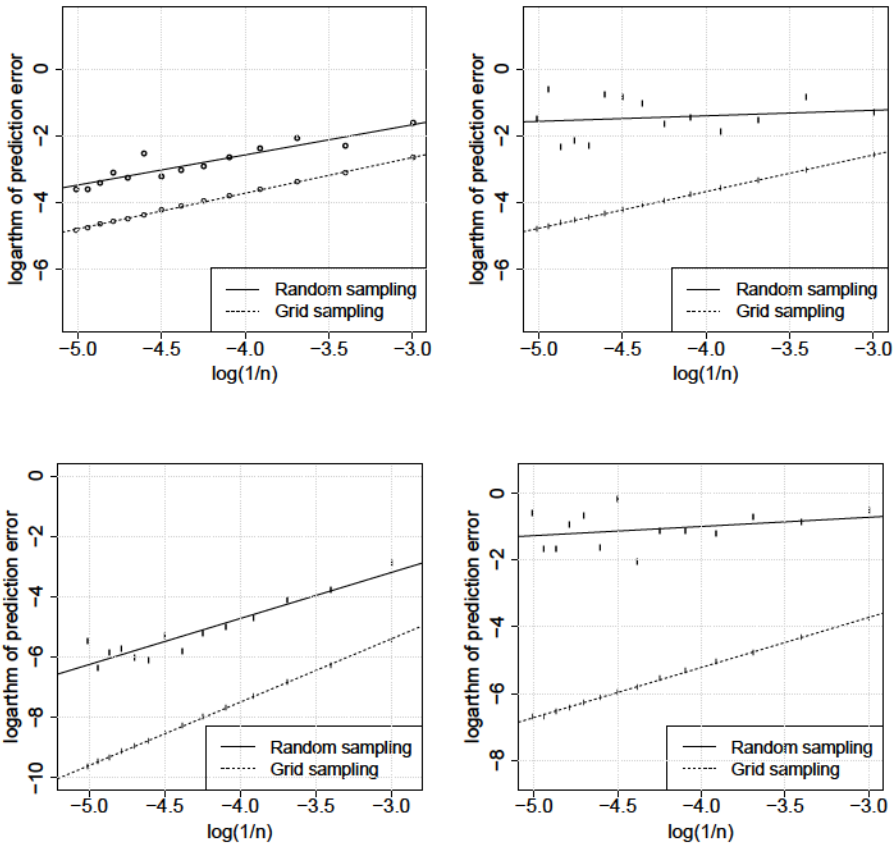


Figure 2: The regression line of $\log \mathcal{E}_{\mathrm{unif}}$ and $\log \mathcal{E}_{\mathrm{grid}}$ on $\log(1/n)$, under the four combinations of (ν_0, ν) in Table 2. Each point denotes one average prediction error for each n.

It has been known for a while that using quasi-uniform sampling is helpful for deterministic function approximation. From an approximation theory perspective, one of the main contributions of this work is the discovery that sample paths of Gaussian processes escapes the $d/2$ smoothness misspecification (in the scattered data approximation sense (Kanagawa et al., 2018)).

As a final remark, we compare the rates in this work with the ones in radial basis function approximation (Edmunds and Triebel, 2008; Wendland, 2004). For the radial basis function approximation problems, we adopt the standard framework so that the underlying function lies in the reproducing kernel Hilbert space generated by the correlation function. For the L_∞ norm, the obtained optimal rate of convergence for kriging is $O_{\mathbb{P}}(n^{-\nu_0/d}\sqrt{\log n})$; while that for the radial basis function approximation is $O(n^{-\nu_0/d})$. So there is a difference in the $\sqrt{\log n}$ factor. For L_p norms with $1 \leq p < \infty$, the difference is more dramatic. While the optimal rate of convergence for kriging is $O_{\mathbb{P}}(n^{-\nu_0/d})$, that for radial basis function approximation is $O(n^{-\nu_0/d-\min(1/2,1/p)})$. This gap between the optimal rates can be explained, as

the support of a Gaussian process is essentially larger than the corresponding reproducing kernel Hilbert space (van der Vaart and van Zanten, 2008b).

7. Proofs

This section comprises our technical proofs. The proofs rely on some results in scattered data approximation of functions in reproducing kernel Hilbert spaces. We introduce these results in Section 7.1. The proofs of the theorems in Sections 3.1 and 3.2 are given in Sections 7.2 and 7.3, respectively.

Before introducing the details, we first note that in the proofs of all results in Sections 3.1 and 3.2, it suffices to consider only the case with $\sigma^2 = 1$. This should not affect the general result because otherwise we can consider the Gaussian process Z/σ instead of Z. Thus for notational simplicity, we assume $\sigma^2 = 1$ throughout this section.

7.1 Reproducing kernel Hilbert spaces and scattered data approximation

We adopt one reviewer's suggestions to prove our main results using techniques from reproducing kernel Hilbert spaces and recent developments in scattered data approximation, in lieu of our original technique of Fourier transform calculations in the previous version. The current treatment can streamline the proofs, and better show how the intermediate quantities toward the error analysis for Gaussian process regression are linked to those studied in scattered data approximation. Reproducing kernel Hilbert spaces is a common mathematical tool in Gaussian processes and scattered data approximation.

Definition 3 *Given a positive definite kernel $K(\cdot)$, the reproducing kernel Hilbert space (RKHS) $\mathcal{N}_K(\mathbb{R}^d)$ is defined as the completion of the function space*

$$\left\{ \sum_{j=1}^{N} \beta_j K(\cdot - x_j) : N \in \mathbb{N}, \beta_j \in \mathbb{R}, x_j \in \mathbb{R}^d \right\}$$

under the inner product

$$\left\langle \sum_{j=1}^{N} \beta_j K(\cdot - x_j), \sum_{k=1}^{N'} \beta'_k K(\cdot - x'_k) \right\rangle_K = \sum_{j=1}^{N} \sum_{k=1}^{N'} \beta_j \beta'_k K(x_j - x'_k). \tag{16}$$

Denote the RKHS norm by $\|\cdot\|_K$.

7.1.1 Interpolation in RKHSs

We first consider the interpolation of a function $f \in \mathcal{N}_\Psi(\mathbb{R}^d)$, by $\mathcal{I}_{\Psi,X} f$. We have the following known results. Lemmas 12 and 13 are Corollaries 10.25 and 10.48 of Wendland (2004), respectively.

Lemma 12 *For any $f \in \mathcal{N}_\Psi(\mathbb{R}^d)$, $\|f - \mathcal{I}_{\Psi,X} f\|_\Psi \le \|f\|_\Psi$.*

Lemma 13 *Under Condition 2, $\mathcal{N}_\Psi(\mathbb{R}^d) = W_2^{\nu_0+d/2}(\mathbb{R}^d)$ with equivalent norms.*

A reviewer suggested an alternative proof of Lemma 11, by leveraging the following Lemma 14 from Narcowich et al. (2006). It is worth presenting the proof here, because we will later employ Lemma 14 again.

Lemma 14 (Theorem 3.4 of Narcowich et al. (2006)) *Suppose $\nu \ge \nu_0 > 0$. Then for each $g \in W_2^{\nu_0+d/2}(\mathbb{R}^d)$, there exists $g_\gamma \in W_2^{\nu+d/2}(\mathbb{R}^d)$, so that $g|_X = g_\gamma|_X$ and*

$$\begin{aligned}\|g_\gamma\|_{W_2^{\nu+d/2}(\mathbb{R}^d)} &\le C q_X^{-(\nu-\nu_0)} \|g\|_{W_2^{\nu_0+d/2}(\mathbb{R}^d)},\\ \|g_\gamma\|_{W_2^{\nu_0+d/2}(\mathbb{R}^d)} &\le C \|g\|_{W_2^{\nu_0+d/2}(\mathbb{R}^d)},\end{aligned}$$

for a constant C depending only on d and ν_0.

Proof of Lemma 11. For each $g \in W_2^{\nu_0+d/2}(\mathbb{R}^d)$, let g_γ be the function given in Lemma 14. The condition $g|_X = g_\gamma|_X$ implies that $\mathcal{I}_{\Phi,X} g = \mathcal{I}_{\Phi,X} g_\gamma$ and $\mathcal{I}_{\Psi,X} g = \mathcal{I}_{\Psi,X} g_\gamma$.

Corollary 11.33 of Wendland (2004) asserts that

$$\|f - \mathcal{I}_{\Phi,X} f\|_{L_\infty(\mathbb{R}^d)} \le C h_{X,\Omega}^\nu \|f\|_{W_2^{\nu+d/2}(\mathbb{R}^d)}. \tag{17}$$

Now by triangle inequality,

$$\begin{aligned}\|g - \mathcal{I}_{\Phi,X} g\|_{L_\infty(\mathbb{R}^d)} \le& \|g - \mathcal{I}_{\Psi,X} g\|_{L_\infty(\mathbb{R}^d)} + \|g_\gamma - \mathcal{I}_{\Psi,X} g_\gamma\|_{L_\infty(\mathbb{R}^d)} + \|g_\gamma - \mathcal{I}_{\Phi,X} g_\gamma\|_{L_\infty(\mathbb{R}^d)}\\ \le& C_1 h_{X,\Omega}^{\nu_0} \left(\|g\|_{W_2^{\nu_0+d/2}(\mathbb{R}^d)} + \|g_\gamma\|_{W_2^{\nu_0+d/2}(\mathbb{R}^d)} \right) + C_2 h_{X,\Omega}^{\nu} \|g_\gamma\|_{W_2^{\nu+d/2}(\mathbb{R}^d)}\\ \le& C_3 (h_{X,\Omega}^{\nu_0} + \rho_{X,\Omega}^{\nu-\nu_0} h_{X,\Omega}^{\nu_0}) \|g\|_{W_2^{\nu_0+d/2}(\mathbb{R}^d)}\\ \le& C_4 h_{X,\Omega}^{\nu_0} \rho_{X,\Omega}^{\nu-\nu_0} \|g\|_{W_2^{\nu_0+d/2}(\mathbb{R}^d)},\end{aligned}$$

where the second inequality follows from (17) and an equivalent form of (17) by replacing Φ with Ψ and ν with ν_0. Hence the proof is completed. ■

7.1.2 Quasi-power functions

Lemma 15 states a simple connection between Gaussian processes and RKHSs.

Lemma 15 *Let $G(\cdot)$ be a stationary Gaussian process on Ω with a unit variance and a positive definite correlation function K. Then for $x_1, \ldots, x_N \in \Omega$ and $\beta_1, \ldots, \beta_N \in \mathbb{R}$,*

$$Var\left(\sum_{j=1}^{N} \beta_j G(x_j)\right) = \left\|\sum_{j=1}^{N} \beta_j K(\cdot - x_j)\right\|_K^2 \tag{18}$$

$$= \sup_{\|f\|_K \le 1} \left|\sum_{j=1}^{N} \beta_j f(x_j)\right|^2. \tag{19}$$

Proof. Equation (18) follows from direct calculations using (16); equation (19) is Lemma 3.9 of Kanagawa et al. (2018). ■

Recall that a kriging interpolant is defined as $\mathcal{I}_{\Phi,X} Z(x) = r_\Phi^T(x) K_\Phi^{-1} Y$; see (8). Lemma 15 will be employed by partially choosing β_j's as the coefficients of a kriging interpolant, i.e., $(\beta_1, \ldots, \beta_n) = r_\Phi^T(x) K_\Phi^{-1}$, which is indeed a constant vector given x and X. For example, Lemma 15 implies

$$\begin{aligned}\mathbb{E}[Z(x) - \mathcal{I}_{\Phi,X} Z(x)]^2 &= \|\Psi(\cdot - x) - \mathcal{I}_{\Phi,X}\Psi(\cdot - x)\|_\Psi^2 \\ &= \sup_{\|f\|_\Psi \le 1} |f(x) - \mathcal{I}_{\Phi,X} f(x)|^2. \end{aligned} \tag{20}$$

We shall call the quantity in (20) the *quasi-power function*, denoted as $Q^2(x)$. Note that $Q^2(x)$ should also depend on Φ, Ψ and X, but we suppress this dependence for notational simplicity, and this will cause no ambiguity. A related quantity is the *power function* (Wendland, 2004), defined as

$$P_{\Psi,X}^2(x) := \mathbb{E}[Z(x) - \mathcal{I}_{\Psi,X} Z(x)]^2,$$

which is the conditional variance of $Z(x)$ given $Z(x_1), \ldots, Z(x_n)$. A simple relationship between $Q(x)$ and $P_{\Psi,X}(x)$ is

$$Q(x) \ge P_{\Psi,X}(x). \tag{21}$$

This inequality can be proven via elementary calculations by showing that $\mathcal{I}_{\Psi,X} Z$ has the smallest mean squared prediction error among all predictors in terms of linear combinations of $Z(x_j)$, and $\mathcal{I}_{\Phi,X} Z$ is one of such. This result is also known as the *best linear prediction* property of $\mathcal{I}_{\Psi,X} Z$ (Stein, 1999; Santner et al., 2003).

The interest here lies in bounding $Q(x)$ in different manners. We state the results in Sections 7.1.3 and 7.1.4.

7.1.3 Upper bounds of the quasi-power function

Lemma 16 can be proven immediately by putting together Lemmas 11, 13 and 15. Lemma 17 is a counterpart of Lemma 16 under the condition $\nu \le \nu_0$, which follows directly from Lemmas 11, 13 and (17).

Lemma 16 *Suppose Conditions 1-3 are met. If $\nu \geq \nu_0$, then there exist constants $C > 0$ and $h_0 \in (0, 1]$ independent of X and x such that*

$$Q(x) \leq C h_{X,\Omega}^{\nu_0} \rho_{X,\Omega}^{\nu-\nu_0}$$

holds for all $x \in \Omega$ and all X satisfying $h_{X,\Omega} \leq h_0$.

Lemma 17 *Suppose Conditions 1-3 are met. If $\nu \leq \nu_0$, then there exist constants $C > 0$ and $h_0 \in (0, 1]$ independent of X and x such that*

$$Q(x) \leq C h_{X,\Omega}^{\nu}$$

holds for all $x \in \Omega$ and all X satisfying $h_{X,\Omega} \leq h_0$.

7.1.4 A lower bound of the quasi-power function

The goal of this section is to prove a lower bound of the quasi-power function under the $L_2(\Omega)$ norm, given by Lemma 18.

Lemma 18 *Suppose Conditions 1-2 are met. Then we have*

$$\|P_{\Psi,X}\|_{L_2(\Omega)} \geq C n^{-\nu_0/d},$$

where $n = \mathrm{card}(X)$. Here the constant C depends only on Ω and Ψ, including ν_0.

Because $\|P_{\Psi,X}\|_{L_2(\Omega)} \leq \sqrt{\mathrm{Vol}(\Omega)} \sup_{x\in\Omega} P_{\Psi,X}(x)$, where $\mathrm{Vol}(\Omega)$ denotes the volume of Ω, we obtain Corollary 19. Corollary 19 is a standard result in scattered data approximation; see, for example, Theorem 11 of Wenzel et al. (2019).

Corollary 19 *Suppose Conditions 1-2 are met. Then we have*

$$\sup_{x\in\Omega} P_{\Psi,X}(x) \geq C n^{-\nu_0/d}.$$

Here the constant C depends only on Ω and Ψ, including ν_0.

To prove Lemma 18, we need a result from the average-case analysis of numerical problems, given by Lemma 20, which is a direct consequence of Theorem 1.2 of Papageorgiou and Wasilkowski (1990). It states lower bounds of $\|Q\|_{L_2(\Omega)}$ in terms of the eigenvalues.

Because Ψ is a positive definite function, by Mercer's theorem (see Pogorzelski (1966) for example), there exists a countable set of positive eigenvalues $\lambda_1 \geq \lambda_2 \geq ... > 0$ and an orthonormal basis for $L_2(\Omega)$, denoted as $\{\varphi_k\}_{k\in\mathbb{N}}$, such that

$$\Psi(x - y) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(x) \varphi_k(y), \tag{22}$$

where the summation is uniformly and absolutely convergent.

Lemma 20 *Let λ_k's be eigenvalues of Ψ. Then we have*

$$\|P_{\Psi,X}\|^2_{L_2(\Omega)} \geq \sum_{k=n+1}^{\infty} \lambda_k.$$

Proof of Lemma 18. Define the kth approximation number of the embedding $id : W_2^{\nu_0+\frac{d}{2}}(\Omega) \to L_p(\Omega)$, denoted by a_k, by

$$a_k = \inf\{\|id - L\|, H \in \mathcal{H}(W_2^{\nu_0+\frac{d}{2}}(\Omega), L_2(\Omega)), \text{rank}(H) < k\},$$

where $\mathcal{H}(W_2^{\nu_0+\frac{d}{2}}(\Omega), L_2(\Omega))$ is the family of all bounded linear mappings $W_2^{\nu_0+\frac{d}{2}}(\Omega) \to L_2(\Omega)$, $\|\cdot\|$ is the operator norm, and $\text{rank}(H)$ is the dimension of the range of H (Edmunds and Triebel, 2008). The approximation number measures the approximation properties by affine (linear) k-dimensional mappings. Let T be the embedding operator of $\mathcal{N}_\Psi(\Omega)$ into $L_2(\Omega)$, and T^* be the adjoint of T. By Proposition 10.28 in Wendland (2004),

$$T^*v(x) = \int_\Omega \Psi(x-y)v(y)dy, \quad v \in L_2(\Omega), x \in \Omega.$$

By Lemma 13, $W_2^{\nu_0+\frac{d}{2}}(\Omega)$ coincides with $\mathcal{N}_\Psi(\Omega)$. By Theorem 5.7 in Edmunds and Evans (2018), T and T^* have the same singular values. By Theorem 5.10 in Edmunds and Evans (2018), for all $k \in \mathbb{N}$, $a_k(T) = \mu_k(T)$, where $a_k(T)$ denotes the approximation number for the embedding operator (as well as the integral operator), and μ_k denotes the singular value of T. By the theorem in Section 3.3.4 in Edmunds and Triebel (2008), the embedding operator T has approximation numbers satisfying

$$C_3 k^{-\nu_0/d-1/2} \leqslant a_k \leqslant C_4 k^{-\nu_0/d-1/2}, \forall k \in \mathbb{N}, \tag{23}$$

where C_3 and C_4 are two positive numbers. By Theorem 5.7 in Edmunds and Evans (2018), $T^*T\varphi_k = \mu_k^2\varphi_k$, and $T^*T\varphi_k = T^*\varphi_k = \lambda_k\varphi_k$, we have $\lambda_k = \mu_k^2$. By (23), $\lambda_k \asymp k^{-2\nu_0/d-1}$ holds. Then the desired result follows from Lemma 20. ■

7.2 L_∞ results

In this section, we prove Theorems 2 and 3. The natural distances of Gaussian processes play a crucial role in establishing these L_∞ results.

Definition 4 *The natural distance $d(x, x')$ of a zero-mean Gaussian process $G(x)$ with $x \in \Omega$ is defined as*

$$d_G^2(x, x') = \mathbb{E}[G(x) - G(x')]^2,$$

for $x, x' \in \Omega$. Once equipped with d_G, Ω becomes a metric space.

The ϵ-*covering number* of the metric space (Ω, d_G), denoted as $N(\epsilon, \Omega, d_G)$, is the minimum integer N so that there exist N distinct balls in (Ω, d_G) with radius ϵ, and the union of these balls covers Ω. The natural distance and the associated covering number, are closely tied to the L_∞ norm of the Gaussian process, say $\sup_{x\in\Omega}|G(x)|$. The needed results are collected in Lemmas 21-23. Lemma 21 is a version of the Borell-TIS inequality for the L_∞ norm of a Gaussian process. Its proof can be found in Pisier (1999).

Lemma 21 (Borell-TIS inequality) *Let $G(x)$ be a separable zero-mean Gaussian process with continuous sample paths almost surely and x lying in a d_G-compact set Ω. Let $\sigma_G^2 = \sup_{x\in\Omega}\mathbb{E}G(x)^2$. Then, we have $\mathbb{E}\sup_{x\in\Omega}|G(x)| < \infty$ and for all $t > 0$,*

$$\mathbb{P}\left(\mathbb{E}\sup_{x\in\Omega}|G(x)| - \sup_{x\in\Omega}|G(x)| \geq t\right) \leq e^{-t^2/2\sigma_G^2}, \tag{24}$$

$$\mathbb{P}\left(\mathbb{E}\sup_{x\in\Omega}|G(x)| - \sup_{x\in\Omega}|G(x)| \leq -t\right) \leq e^{-t^2/2\sigma_G^2}. \tag{25}$$

Lemma 22 (Corollary 2.2.8 of van der Vaart and Wellner (1996)) *Let $G(x)$ be as in Lemma 21. For some universal constant C, we have*

$$\mathbb{E}\sup_{x,x'\in\Omega}|G(x) - G(x')| \leq C\int_0^D \sqrt{\log N(\epsilon, \Omega, d_G)}d\epsilon,$$

where $D = \sup_{x,x'\in\Omega} d_G(x, x')$ is the diameter of Ω under d_G.

Lemma 23 (Theorem 6.5 of van Handel (2014)) *Let $G(x)$ be as in Lemma 21. For some universal constant C, we have*

$$\mathbb{E}\sup_{x\in\Omega}|G(x)| \geq C\sup_{\eta>0}\eta\sqrt{\log N(\eta, \Omega, d_G)}.$$

To utilize the above lemmas to bound $\sup_{x\in\Omega}|Z(x) - \mathcal{I}_{\Phi,X}Z(x)|$, the main idea is to note that

$$g_Z(x) := Z(x) - \mathcal{I}_{\Phi,X}Z(x)$$

is also a Gaussian process. So Lemma 21 can be applied directly. The remainder is to bound $\mathbb{E}\sup_{x\in\Omega}|Z(x) - \mathcal{I}_{\Phi,X}Z(x)|$. According to Lemmas 22 and 23, it is crucial to understand the natural distance, given by

$$d_{g_Z}^2(x, x') = \mathbb{E}[Z(x) - \mathcal{I}_{\Phi,X}Z(x) - Z(x') + \mathcal{I}_{\Phi,X}Z(x')]^2.$$

7.2.1 Proof of Theorem 2

The main steps of proving Theorem 2 are: 1) bounding the diameter D; 2) connecting the natural distance d_{g_Z} with the Euclidean distance; 3) bounding the covering integral and establishing the desired result.

Step 1. An upper bound of D is given by

$$\begin{aligned} D^2 &= \sup_{x,x' \in \Omega} \mathbb{E}[Z(x) - \mathcal{I}_{\Phi,X} Z(x) - Z(x') + \mathcal{I}_{\Phi,X} Z(x')]^2 \\ &\leq 4 \sup_{x \in \Omega} \mathbb{E}[Z(x) - \mathcal{I}_{\Phi,X} Z(x)]^2 \\ &= 4 \sup_{x \in \Omega} Q^2(x) \leq C_1^2 h_{X,\Omega}^{2\nu_0} \rho_{X,\Omega}^{2(\nu-\nu_0)}, \end{aligned} \tag{26}$$

where the first inequality follows from the basic inequality $(x+y)^2 \leq 2x^2 + 2y^2$; the last inequality follows from Lemma 16.

Step 2. By Lemma 15,

$$d_{g_Z}(x,x') = \sup_{\|f\|_\Psi \leq 1} |f(x) - \mathcal{I}_{\Phi,X} f(x) - f(x') + \mathcal{I}_{\Phi,X} f(x')|. \tag{27}$$

The Hölder space $C_b^{0,\alpha}(\mathbb{R}^d)$ for $0 < \alpha \leq 1$ consists of continuous bounded function on $\mathbb{R}^d$, with its norm defined as

$$\|f\|_{C_b^{0,\alpha}(\mathbb{R}^d)} := \sup_{x,x' \in \mathbb{R}^d, x \neq x'} \frac{|f(x) - f(x')|}{\|x - x'\|^\alpha}.$$

For $f \in \mathcal{N}_\Psi(\mathbb{R}^d)$, Lemma 13 implies that $f - \mathcal{I}_{\Phi,X} f \in W_2^{\nu_0 + d/2}(\mathbb{R}^d)$. The Sobolev embedding theorem (see, for example, Theorem 4.47 of Demengel et al. (2012)) implies the embedding relationship $W_2^{\nu_0+d/2}(\mathbb{R}^d) \subset C_b^{0,\tau}(\mathbb{R}^d)$ with $\tau = \min(\nu_0, 1)$, and

$$\|h\|_{C_b^{0,\tau}(\mathbb{R}^d)} \leq C_2 \|h\|_{W_2^{\nu_0+d/2}(\mathbb{R}^d)}, \tag{28}$$

for all $h \in W_2^{\nu_0+d/2}(\mathbb{R}^d)$ and a constant C_2. Therefore, we have $f - \mathcal{I}_{\Phi,X} f \in C_b^{0,\tau}(\mathbb{R}^d)$.

Thus by (27), we have

$$\begin{aligned} d_{g_Z}(x,x') &\leq \sup_{\|f\|_\Psi \leq 1} \|f - \mathcal{I}_{\Phi,X} f\|_{C_b^{0,\tau}(\mathbb{R}^d)} \|x - x'\|^\tau \\ &\leq \sup_{\|f\|_\Psi \leq 1} C_2 \|f - \mathcal{I}_{\Phi,X} f\|_{W_2^{\nu_0+d/2}(\mathbb{R}^d)} \|x - x'\|^\tau \\ &\leq \sup_{\|f\|_\Psi \leq 1} C_3 \|f - \mathcal{I}_{\Phi,X} f\|_\Psi \|x - x'\|^\tau, \end{aligned} \tag{29}$$

where the second inequality follows from (28); the third inequality follows from Lemma 13.

Now we employ Lemma 14 again. Let f_γ be the function asserted by Lemma 14 with $f|_X = f_\gamma|_X$. Similar to the proof of Lemma 11, we have

$$\begin{aligned} &\|f - \mathcal{I}_{\Phi,X} f\|_\Psi \\ \leq &\|f - \mathcal{I}_{\Psi,X} f\|_\Psi + \|f_\gamma - \mathcal{I}_{\Psi,X} f_\gamma\|_\Psi + \|f_\gamma - \mathcal{I}_{\Phi,X} f_\gamma\|_\Psi \\ \leq &\|f\|_\Psi + \|f_\gamma\|_\Psi + \|f_\gamma - \mathcal{I}_{\Phi,X} f_\gamma\|_\Psi, \\ \leq &C_4 \|f\|_\Psi + \|f_\gamma - \mathcal{I}_{\Phi,X} f_\gamma\|_\Psi \end{aligned} \tag{30}$$

where the second inequality follows from Lemma 12; the last inequality follows from Lemmas 13 and 14. Similarly, we have

$$\begin{aligned}
&\|f_\gamma - \mathcal{I}_{\Phi,X} f_\gamma\|_\Psi \le C_5 \|f_\gamma - \mathcal{I}_{\Phi,X} f_\gamma\|_{W_2^{\nu_0+d/2}(\mathbb{R}^d)} \\
\le& C_5 \|f_\gamma - \mathcal{I}_{\Phi,X} f_\gamma\|_{W_2^{\nu+d/2}(\mathbb{R}^d)} \le C_6 \|f_\gamma - \mathcal{I}_{\Phi,X} f_\gamma\|_\Phi \\
\le& C_6 \|f_\gamma\|_\Phi \le C_7 \|f_\gamma\|_{W_2^{\nu+d/2}(\mathbb{R}^d)} \le C_8 q_X^{-(\nu-\nu_0)} \|f\|_{W_2^{\nu_0+d/2}(\mathbb{R}^d)} \\
\le& C_9 q_X^{-(\nu-\nu_0)} \|f\|_\Psi,
\end{aligned}$$

which, together with (29) and (30), yields

$$d_{g_Z}(x,x') \le C_{10} q_X^{-(\nu-\nu_0)} \|x-x'\|^\tau = C_{10} h_{X,\Omega}^{\nu_0-\nu} \rho_{X,\Omega}^{\nu-\nu_0} \|x-x'\|^\tau.$$

Therefore, by the definition of the covering number, we have

$$N(\epsilon, \Omega, d_{g_Z}) \le N((\epsilon / C_{10} h_{X,\Omega}^{\nu_0-\nu} \rho_{X,\Omega}^{\nu-\nu_0})^{1/\tau}, \Omega, \|\cdot\|). \tag{31}$$

The right side of (31) involves the covering number of a Euclidean ball, which is studied in the literature; see Lemma 2.5 of van de Geer (2000). This result leads to the bound

$$\log N(\epsilon, \Omega, d_{g_Z}) \le C_{11} \log\left(1 + C_{12}\left(\frac{h_{X,\Omega}^{\nu_0-\nu} \rho_{X,\Omega}^{\nu-\nu_0}}{\epsilon}\right)^{1/\tau}\right). \tag{32}$$

Step 3. For any $x_1 \in X$, the interpolation property implies $g_Z(x_1) = 0$. Using our findings in Steps 1 and 2, together with Lemma 22, we have

$$\begin{aligned}
&\mathbb{E} \sup_{x\in\Omega} |g_Z(x)| = \mathbb{E} \sup_{x\in\Omega} |g_Z(x) - g_Z(x_1)| \\
\le& \mathbb{E} \sup_{x,x'\in\Omega} |g_Z(x) - g_Z(x')| \\
\le& C_{13} \int_0^{C_1 h_{X,\Omega}^{\nu_0} \rho_{X,\Omega}^{\nu-\nu_0}} \sqrt{\log\left(1 + C_{12}\left(\frac{h_{X,\Omega}^{\nu_0-\nu} \rho_{X,\Omega}^{\nu-\nu_0}}{\epsilon}\right)^{1/\tau}\right)} d\epsilon \\
=& C_{13} h_{X,\Omega}^{\nu_0} \rho_{X,\Omega}^{\nu-\nu_0} \int_0^{C_1} \sqrt{\log\left(1 + C_{12}\left(\frac{h_{X,\Omega}^{-\nu}}{t}\right)^{1/\tau}\right)} dt,
\end{aligned} \tag{33}$$

where the second equality is obtained by the change of variables. Note that for any $b > \frac{1}{C_1}$ and $a > 0$, taking $C' = \max\{C_1, 1\}$ leads to $1 + b^a \le (1+C'b)^a + C'b(1+C'b)^a \le (1+C'b)^{a+1}$. Thus we have

$$\log\left(1 + C_{12}\left(\frac{h_{X,\Omega}^{-\nu}}{t}\right)^{1/\tau}\right) \le \left(1 + \frac{1}{\tau}\right) \log\left(1 + C_{14} \frac{h_{X,\Omega}^{-\nu}}{t}\right)$$

for $t \in (0, C_1]$.

Therefore, the integral (33) can be further bounded by

$$C_{15}h_{X,\Omega}^{\nu_0}\rho_{X,\Omega}^{\nu-\nu_0}\int_0^{C_1}\sqrt{\log\left(1+\frac{C_{14}}{h_{X,\Omega}^{\nu}t}\right)}dt. \tag{34}$$

We then apply the Cauchy-Schwarz inequality to get

$$\begin{aligned}
&C_{15}h_{X,\Omega}^{\nu_0}\rho_{X,\Omega}^{\nu-\nu_0}\int_0^{C_1}\sqrt{\log\left(1+\frac{C_{14}}{h_{X,\Omega}^{\nu}t}\right)}dt\\
\leq&C_{15}C_1^{1/2}h_{X,\Omega}^{\nu_0}\rho_{X,\Omega}^{\nu-\nu_0}\left(\int_0^{C_1}\log\left(1+\frac{C_{14}}{h_{X,\Omega}^{\nu}t}\right)dt\right)^{1/2}\\
=&C_{15}C_1^{1/2}h_{X,\Omega}^{\nu_0}\rho_{X,\Omega}^{\nu-\nu_0}\left(C_{14}h_{X,\Omega}^{-\nu}\log\left(1+\frac{C_1h_{X,\Omega}^{\nu}}{C_{14}}\right)+C_1\log\left(1+\frac{C_{14}}{C_1h_{X,\Omega}^{\nu}}\right)\right)^{1/2}.
\end{aligned} \tag{35}$$

By the basic inequality $\log(1+x)\leq x$, we conclude that

$$C_{14}h_{X,\Omega}^{-\nu}\log\left(1+\frac{C_1h_{X,\Omega}^{\nu}}{C_{14}}\right)\leq C_1.$$

Consequently, by incorporating the condition $h_{X,\Omega}\leq 1$, we get

$$\begin{aligned}
&C_{14}h_{X,\Omega}^{-\nu}\log\left(1+\frac{C_1h_{X,\Omega}^{\nu}}{C_{14}}\right)+C_1\log\left(1+\frac{C_{14}}{C_1h_{X,\Omega}^{\nu}}\right)\\
\leq&C_1+C_1\log\left(1+\frac{C_{14}}{C_1h_{X,\Omega}^{\nu}}\right)\leq C_{16}\log\left(1+\frac{C_{17}}{h_{X,\Omega}}\right),
\end{aligned} \tag{36}$$

where in the last equality, we utilize $1+b^a\leq(1+C'b)^{a+1}$ again. Combining (33)-(36), we have shown that

$$\mathbb{E}\sup_{x\in\Omega}|g_Z(x)|\leq C_{18}h_{X,\Omega}^{\nu_0}\rho_{X,\Omega}^{\nu-\nu_0}\sqrt{\log\left(1+\frac{C_{17}}{h_{X,\Omega}}\right)}.$$

By Lemma 16, we have the following upper bound for the maximum variance

$$\sigma_{g_Z}^2=\max_{x\in\Omega}Q^2(x)\leq Ch_{X,\Omega}^{2\nu_0}\rho_{X,\Omega}^{2(\nu-\nu_0)}.$$

Then we complete the proof of Theorem 2 by invoking (24) of Lemma 21.

7.2.2 Proof of Theorem 3

According to Lemma 23, the key is to find a lower bound of $N(\epsilon,\Omega,d_{g_Z})$. The idea is as follows. Suppose for any n-point set $\{y_1,\ldots,y_n\}\subset\Omega$, we can find $y_0\in\Omega$ such that $\min_{1\leq j\leq n}d_{g_Z}(y_0,y_j)\geq\epsilon_n$ for some number $\epsilon_n>0$. Then Ω can not be covered by n (ϵ_n,d_{g_Z})-balls, and thus $N(\epsilon_n,\Omega,d_{g_Z})\geq n$.

Now take an arbitrary n-point set $Y = \{y_1, \ldots, y_n\} \subset \Omega$. For each y_j,

$$\begin{aligned} d_{g_Z}^2(y, y_j) =& \mathbb{E}(Z(y) - \mathcal{I}_{\Phi,X} Z(y) - Z(y_j) + \mathcal{I}_{\Phi,X} Z(y_j)))^2 \\ \geq& \mathbb{E}(Z(y) - \mathcal{I}_{\Psi,X\cup Y} Z(y))^2, \end{aligned}$$

because $\mathcal{I}_{\Psi,X\cup Y} Z(y)$ is the best linear predictor of $Z(y)$ given $Z(x_1), \ldots, Z(x_n), Z(y_1), \ldots, Z(y_n)$, and $\mathcal{I}_{\Phi,X} Z(y) + Z(y_j) - \mathcal{I}_{\Phi,X} Z(y_j)$ is a linear predictor and thus should have a greater mean squared prediction error. Corollary 19 implies that

$$\sup_{y\in\Omega} \mathbb{E}(Z(y) - \mathcal{I}_{\Psi,X\cup Y} Z(y))^2 = \sup_{y\in\Omega} P_{\Psi,X\cup Y}^2(y) \geq C_1^2 (2n)^{-\frac{2\nu_0}{d}}.$$

Therefore, there exists $y_0 \in \Omega$ such that $d_{g_Z}^2(y_0, y_j) \geq C_1^2 (2n)^{-\frac{2\nu_0}{d}}/4$ for each y_j, which implies $N(C_1(2n)^{-\frac{\nu_0}{d}}/2, \Omega, d_{g_Z}) \geq n$. Now we invoke Lemma 23 with $\eta = C_1(2n)^{-\frac{\nu_0}{d}}/2$ to obtain that

$$\mathbb{E} \sup_{x\in\Omega} g_Z(x) \geq C_2 n^{-\frac{\nu_0}{d}} \sqrt{\log n}. \tag{37}$$

The desired result then follows from (25) of Lemma 21.

7.3 L_p results with $1 \leq p < \infty$

Our results for the L_p norms with $1 \leq p < \infty$ replies on a counterpart of the Borell-TIS inequality (Lemma 21) under the L_p norms. Such a result is given by Lemma 24; its proof is presented in Section 7.3.1.

Lemma 24 *Suppose Ω satisfies Condition 1. Let G be a zero-mean Gaussian process on $\Omega \subset \mathbb{R}^d$ with continuous sample paths almost surely and with a finite maximum pointwise variance $\sigma_G^2 = \sup_{x\in\Omega} \mathbb{E}G(x)^2 < \infty$. Then for all $u > 0$ and $1 \leq p < \infty$, we have*

$$\begin{aligned} \mathbb{P}\left(\|G\|_{L_p(\Omega)} - \mathbb{E}\|G\|_{L_p(\Omega)} > u\right) &\leq e^{-u^2/(2C_p\sigma_G^2)}, \\ \mathbb{P}\left(\|G\|_{L_p(\Omega)} - \mathbb{E}\|G\|_{L_p(\Omega)} < -u\right) &\leq e^{-u^2/(2C_p\sigma_G^2)}, \end{aligned}$$

with $C_p = \mathrm{Vol}(\Omega)^{2/p}$. Here $\mathrm{Vol}(\Omega)$ denotes the volume of Ω.

Remark 25 *Similar to the Borell-TIS inequality (Lemma 21), the variation of L_p norm of G in Lemma 24 can be controlled by its pointwise fluctuations. In fact, by letting $p \to \infty$, Lemma 24 becomes Lemma 21.*

As before, let $g_Z(x) = Z(x) - \mathcal{I}_{\Phi,X} Z(x)$, which is still a zero-mean Gaussian process; let $\sigma_{g_Z}^2 = \sup_{x\in\Omega} \mathbb{E}g_Z(x)^2$. In view of Lemma 24, the remaining task is to bound $\mathbb{E}\|g_Z\|_{L_p(\Omega)}$. This will be done by employing the known bounds of $\mathbb{E}\|g_Z\|_{L_2(\Omega)}^2$, as in Lemmas 16 and 18, together with Jensen's inequality and some other basic inequalities.

7.3.1 Proof of Lemma 24

We will use the Gaussian concentration inequality given by Lemma 26. Its proof can be found in Adler and Taylor (2009, Lemma 2.1.6). We say that L is a Lipchitz constant of the function $h : \mathbb{R}^k \to \mathbb{R}$, if $|h(x) - h(y)| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^k$.

Lemma 26 (Gaussian concentration inequality) *Let G_k be a k-dimensional vector of centered, unit-variance, independent Gaussian variables. If $h : \mathbb{R}^k \to \mathbb{R}$ has Lipschitz constant L, then for all $u > 0$.*

$$\mathbb{P}(h(G_k) - \mathbb{E}h(G_k) > u) \leq e^{-u^2/(2L^2)}.$$

The proof proceeds by approximating of the integral $\|G\|^p_{L_p(\Omega)} = \int_\Omega G(x)^p dx$ by a Riemann sum. For each $n = 1, 2, \ldots$, let $\{\Omega_{nj}\}_{j=1}^n$ be a partition of Ω such that

$$\max_{1\leq j\leq n} \mathrm{Diam}(\Omega_{nj}) \to 0, \text{ as } n \to \infty, \tag{38}$$

where $\mathrm{Diam}(\Omega_{nj})$ denotes the (Euclidean) diameter of Ω_{nj}. We have $\Omega = \cup_{j=1}^n \Omega_{nj}$ and $\sum_{j=1}^n \mathrm{Vol}(\Omega_{nj}) = \mathrm{Vol}(\Omega)$. Let $w_{nj} = \mathrm{Vol}(\Omega_{nj})$ and define

$$\|a\|_w = \left(\sum_{j=1}^n w_{nj} |a_j|^p \right)^{1/p}$$

for a vector $a = (a_1, ..., a_n)^T$. Let $G_n = (G_{n1}, ..., G_{nn})^T$ with $G_{nj} = G(x_{nj})$ for some $x_{nj} \in \Omega_{nj}$. Therefore, $\|G_n\|_w$ is an approximate of $\|G\|_{L_p(\Omega)}$.

We first prove a similar result for G_n, and then arrive at the desired results by letting $n \to \infty$. Let K be the $n \times n$ covariance matrix of G_n on Ω, with components $K_{ij} = \mathbb{E}(G_{ni} G_{nj})$. Define $\sigma^2_{\Omega_n} = \max_{1\leq j\leq n} \mathbb{E}G^2_{nj}$. Let W be a vector of independent, standard Gaussian variables, and A be a matrix such that $A^T A = K$. Thus G_n has the same distribution as AW.

Consider the function $h(x) = \|Ax\|_w$. Let e_j be the vector with one in the jth entry and zeros in other entries. Denote the jth entry of a vector v by $[v]_j$. Then we have

$$\begin{aligned}
|h(x) - h(y)| =& |\|Ax\|_w - \|Ay\|_w| \leq \|A(x - y)\|_w \\
=& \left(\sum_{j=1}^n w_{nj} \left| [A(x - y)]_j \right|^p \right)^{1/p} = \left(\sum_{j=1}^n w_{nj} |e_j^T A(x - y)|^p \right)^{1/p} \\
\leq& \left(\sum_{j=1}^n w_{nj} \|e_j^T A\|^p \|x - y\|^p \right)^{1/p},
\end{aligned}$$

where the first inequality follows from the triangle inequality (i.e., the Minkowski inequality); and the last inequality follows from the Cauchy-Schwarz inequality. Noting that for each j,

$$\|e_j^T A\|^2 = e_j^T A^T A e_j = \mathbb{E}(G^2_{nj}) \leq \sigma^2_{\Omega_n},$$

we have

$$|h(x) - h(y)| \leq \mathrm{Vol}(\Omega)^{1/p} \sigma_{\Omega_n} \|x - y\|,$$

which implies h is a Lipschitz continuous function with Lipschitz constant $\mathrm{Vol}(\Omega)^{1/p}\sigma_{\Omega_n}$. Because G_n and AW have the same distribution, and together with Lemma 26, we obtain

$$\mathbb{P}\left(\|G_n\|_w - \mathbb{E}\|G_n\|_w > u\right) \leq e^{-u^2/(2C_p\sigma^2_{\Omega_n})}, \tag{39}$$

where $C_p = \mathrm{Vol}(\Omega)^{2/p}$. Similarly, by considering $h(x) = -\|Ax\|_w$, we can obtain

$$\mathbb{P}\left(\|G_n\|_w - \mathbb{E}\|G_n\|_w < -u\right) \leq e^{-u^2/(2C_p\sigma^2_{\Omega_n})}. \tag{40}$$

To prove the desired results, we let $n \to \infty$ in (39) and (40). First we show that the left-hand sides of (39) and (40) tend to $\mathbb{P}\left(\|G\|_{L_p(\Omega)} - \mathbb{E}\|G\|_{L_p(\Omega)} > u\right)$ and $\mathbb{P}\left(\|G\|_{L_p(\Omega)} - \mathbb{E}\|G\|_{L_p(\Omega)} < -u\right)$, respectively. According to Lebesgue's dominated convergence theorem, it suffices to prove that

$$\|G_n\|_w \to \|G\|_{L_p(\Omega)}, \quad a.s., \text{ as } n \to \infty, \tag{41}$$

and

$$\mathbb{E}\|G_n\|_w \to \mathbb{E}\|G\|_{L_p(\Omega)}, \text{ as } n \to \infty, \tag{42}$$

as the indicator function is dominated by one. Since G has continuous sample paths with probability one, (41) is an immediate consequence of the convergence of Riemann integrals. Now we prove (42). Note that $\|G_n\|_w \leq \mathrm{Vol}(\Omega)^{1/p} \sup_{x\in\Omega} |G(x)|$ and Lemma 21 suggests that $\mathbb{E}\sup_{x\in\Omega}|G(x)| < \infty$. Thus Lebesgue's dominated convergence theorem implies (42). Now we consider the right-hand sides of (39) and (40). To prove the desired results, it remains to prove $\sigma^2_{\Omega_n} \to \sigma^2_G$. By the definition of $\sigma^2_{\Omega_n}$, we have

$$\max_{1\leq j\leq n} \inf_{x\in\Omega_{nj}} \mathbb{E}G(x)^2 \leq \sigma^2_{\Omega_n} \leq \sigma^2_G. \tag{43}$$

The almost sure continuity of G implies that $\mathbb{E}G(x)^2$ is continuous in x. Since Ω is compact, $\mathbb{E}G(x)^2$ is also uniformly continuous. Therefore, the condition of the partitions in (38) implies

$$\max_{1\leq j\leq n} \inf_{x\in\Omega_{nj}} \mathbb{E}G^2(x) \to \sigma^2_G, \text{ as } n \to \infty. \tag{44}$$

Combining (43) and (44) proves $\sigma^2_{\Omega_n} \to \sigma^2_G$, which completes the proof.

7.3.2 Proof of Theorem 4

By Fubini's theorem,

$$\begin{aligned}
\mathbb{E}\|Z - \mathcal{I}_{\Phi,X}Z\|^p_{L_p(\Omega)} &= \int_{x\in\Omega} \mathbb{E}|Z(x) - \mathcal{I}_{\Phi,X}Z(x)|^p dx \\
&= \int_{x\in\Omega} \frac{2^{p/2}\Gamma(\frac{p+1}{2})}{\sqrt{\pi}} \left(\mathbb{E}(Z(x) - \mathcal{I}_{\Phi,X}Z(x))^2\right)^{p/2} dx \\
&\leq C_1 \sigma^p_{g_Z}.
\end{aligned} \tag{45}$$

The second equality of (45) is true because $Z(x) - \mathcal{I}_{\Phi,X} Z(x)$ follows a normal distribution with mean zero, and the absolute moments of a normal random variable $X_\sigma \sim N(0, \sigma^2)$ can be expressed by its variance as

$$\mathbb{E}|X_\sigma|^p = \sigma^p \cdot \frac{2^{p/2}\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}; \tag{46}$$

see Walck (1996). By combining Lemma 24 and (45), we have

$$\begin{aligned} e^{-u^2/2C_p\sigma_{g_Z}^2} &\geq \mathbb{P}\left(\|g_Z\|_{L_p(\Omega)} > \mathbb{E}\|g_Z\|_{L_p(\Omega)} + u\right) \\ &\geq \mathbb{P}\left(\|g_Z\|_{L_p(\Omega)}^p > 2^{p-1}(\mathbb{E}\|g_Z\|_{L_p(\Omega)}^p + u^p)\right) \\ &\geq \mathbb{P}\left(\|g_Z\|_{L_p(\Omega)}^p > 2^{p-1}(C_1\sigma_{g_Z}^p + u^p)\right) \\ &\geq \mathbb{P}\left(\|g_Z\|_{L_p(\Omega)} > 2^{1-1/p}(C_1^{1/p}\sigma_{g_Z} + u)\right), \end{aligned} \tag{47}$$

where the second inequality follows from the Jensen's inequality and the c_r-inequality. Combining Lemma 16 and (47) completes the proof.

7.3.3 Proof of Theorem 5

The proof of Theorem 5 is similar to that of Theorem 4. The only difference here is that at the last step we employ Lemma 17 instead of Lemma 16.

7.3.4 Proof of Theorem 6

Take a quasi-uniform design $X' \subset \Omega$ with $\mathrm{card}(X') = n$. Obviously $h_{X\cup X',\Omega} \leq h_{X',\Omega}$. By Proposition 14.1 of Wendland (2004), $h_{X',\Omega} \leq Cn^{-1/d}$. By Hölder's inequality, we have $\|f\|_{L_2(\Omega)} \leq \|f\|_{L_1(\Omega)}^{1/4}\|f\|_{L_3(\Omega)}^{3/4}$ for any continuous function f, which implies

$$\begin{aligned} &\left(\int_{x\in\Omega} \mathbb{E}(Z(x) - \mathcal{I}_{\Psi,X\cup X'} Z(x))^2 dx\right)^{1/2} \\ \leq &\left(\int_{x\in\Omega} \left(\mathbb{E}(Z(x) - \mathcal{I}_{\Psi,X\cup X'} Z(x))^2\right)^{1/2} dx\right)^{1/4} \left(\int_{x\in\Omega} \left(\mathbb{E}(Z(x) - \mathcal{I}_{\Psi,X\cup X'} Z(x))^2\right)^{3/2} dx\right)^{1/4}. \end{aligned} \tag{48}$$

Applying Lemma 17 to $\sup_{x\in\Omega} \mathbb{E}(Z(x) - \mathcal{I}_{\Psi,X\cup X'} Z(x))^2$ with $\nu = \nu_0$ yields

$$\begin{aligned} &\left(\int_{x\in\Omega} \left(\mathbb{E}(Z(x) - \mathcal{I}_{\Psi,X\cup X'} Z(x))^2\right)^{3/2} dx\right)^{1/4} \\ \leq &C_1\left(\sup_{x\in\Omega} \left(\mathbb{E}(Z(x) - \mathcal{I}_{\Psi,X\cup X'} Z(x))^2\right)^{3/2}\right)^{1/4} \\ \leq &C_2 h_{X\cup X',\Omega}^{\frac{3\nu_0}{4}} \leq C_2 h_{X',\Omega}^{\frac{3\nu_0}{4}} \leq C_3 n^{-\frac{3\nu_0}{4d}}. \end{aligned} \tag{49}$$

The left hand side of (48) can be bounded from below by using Lemma 18, which yields

$$\begin{aligned}\Big(\int_{x\in\Omega}\mathbb{E}(Z(x)-\mathcal{I}_{\Psi,X\cup X'}Z(x))^2dx\Big)^{1/2} &= \Big(\mathbb{E}\|Z-\mathcal{I}_{\Psi,X\cup X'}Z\|^2_{L_2(\Omega)}\Big)^{1/2}\\ &= \|P_{\Psi,X\cup X'}\|_{L_2(\Omega)}\geq C_4(2n)^{-\nu_0/d},\end{aligned}\tag{50}$$

where the equality follows from Fubini's theorem. Plugging (49) and (50) into (48), we have

$$\int_{x\in\Omega}\big(\mathbb{E}(Z(x)-\mathcal{I}_{\Psi,X\cup X'}Z(x))^2\big)^{1/2}dx\geq C_5n^{\frac{3\nu_0}{d}}n^{-\frac{4\nu_0}{d}}=C_5n^{-\frac{\nu_0}{d}}.\tag{51}$$

By Fubini's theorem and (51), it can be seen that

$$\begin{aligned}\mathbb{E}\|Z-\mathcal{I}_{\Phi,X}Z\|_{L_1(\Omega)} &= \int_{x\in\Omega}\mathbb{E}|Z(x)-\mathcal{I}_{\Phi,X}Z(x)|dx\\ &= \int_{x\in\Omega}\frac{2^{1/2}}{\sqrt{\pi}}\big(\mathbb{E}(Z(x)-\mathcal{I}_{\Phi,X}Z(x))^2\big)^{1/2}dx\\ &\geq C_6\int_{x\in\Omega}\big(\mathbb{E}(Z(x)-\mathcal{I}_{\Psi,X\cup X'}Z(x))^2\big)^{1/2}dx\\ &\geq C_7n^{-\frac{\nu_0}{d}},\end{aligned}\tag{52}$$

where the second equality follows from (46) with $p=1$; the first inequality is because $\mathcal{I}_{\Psi,X\cup X'}Z(x)$ is the best linear predictor of $Z(x)$. For $1\leq p<\infty$ and any $u>0$, applying Lemma 24 yields

$$\begin{aligned}e^{-u^2/(2C_8\sigma_\Omega^2)} &\geq \mathbb{P}\left(\|g_Z\|_{L_p(\Omega)}<\mathbb{E}\|g_Z\|_{L_p(\Omega)}-u\right)\\ &\geq \mathbb{P}\left(\|g_Z\|^p_{L_p(\Omega)}<2^{1-p}(\mathbb{E}\|g_Z\|_{L_p(\Omega)})^p-u^p\right)\\ &\geq \mathbb{P}\left(\|g_Z\|^p_{L_p(\Omega)}<2^{1-p}(C_9\mathbb{E}\|g_Z\|_{L_1(\Omega)})^p-u^p\right)\\ &\geq \mathbb{P}\left(\|g_Z\|^p_{L_p(\Omega)}<C_{10}n^{-\nu_0p/d}-u^p\right)\\ &\geq \mathbb{P}\left(\|g_Z\|_{L_p(\Omega)}<C_{11}n^{-\nu_0/d}-u\right).\end{aligned}\tag{53}$$

In (53), the second inequality is because of Jensen's inequality; the third inequality is because of the fact $\|g_Z\|_{L_p(\Omega)}\geq C_9\|g_Z\|_{L_1(\Omega)}$ for some constant $C_9>0$ depending on p and Ω; the fourth inequality is by (52); and the last inequality is true because of the elementary inequality $(a+b)^p\geq a^p+b^p$ for $a,b>0$. Thus, we finish the proof of Theorem 6.

Acknowledgements

The authors are grateful to the AE and three reviewers for very helpful comments and suggestions.

Appendix A. Distributions and Asymptotic Orders in Example 1

Proposition 27 *Let $x_1, \ldots, x_n$ be mutually independent random variables following the uniform distribution on $[0,1]$. Denote their order statistics as*

$$0 = x_{(0)} \leq x_{(1)} \leq \cdots \leq x_{(n)} \leq x_{(n+1)} = 1.$$

Let $y_1, \ldots, y_n$ be mutually independent random variables following the exponential distribution with mean one. Therefore, $(x_{(1)}, \ldots, x_{(n)})$ has the same distribution as

$$\left(\frac{y_1}{\sum_{j=1}^{n+1} y_j}, \ldots, \frac{\sum_{j=1}^{n} y_j}{\sum_{j=1}^{n+1} y_j} \right).$$

The proof of Proposition 27 relies on the following lemma.

Lemma 28 (Lemma 4.5.1 of Resnick (1992)) *Let $y_1, \ldots, y_n, y_{n+1}$ be mutually independent random variables following the exponential distribution with mean one. Define $E_k = \sum_{i=1}^{k} y_i$ for $k = 1, ..., n+1$. Then conditional on $E_{n+1} = t$, the joint density of $E_1, ..., E_n$ is*

$$f_{E_1,\ldots,E_n|E_{n+1}=t}(u_1, ..., u_n) = \begin{cases} \frac{n!}{t^n}, & \text{if } 0 < u_1 < \cdots < u_n < t, \\ 0, & \text{otherwise.} \end{cases}$$

Proof of Proposition 27. By Lemma 28, it can be shown that

$$\begin{aligned} f_{\frac{E_1}{E_{n+1}},\ldots,\frac{E_n}{E_{n+1}}|E_{n+1}=t}(u_1, ..., u_n) &= f_{E_1,\ldots,E_n|E_{n+1}=t}(u_1 t, ..., u_n t) \\ &= \begin{cases} n!, & \text{if } 0 < u_1 < \cdots < u_n < 1, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

which implies

$$f_{\frac{E_1}{E_{n+1}},\ldots,\frac{E_n}{E_{n+1}}}(u_1, ..., u_n) = \begin{cases} n!, & \text{if } 0 < u_1 < ... < u_n < 1, \\ 0, & \text{otherwise,} \end{cases} \tag{54}$$

by taking the expectation with respect to E_{n+1}. Note that (54) is the same as the joint density of order statistics $(x_{(1)}, \ldots, x_{(n)})$, which completes the proof. ■

Proposition 29 *Let $y_1, \ldots, y_n$ be mutually independent random variables following the exponential distribution with mean one. Then $\max y_j \asymp \log n$, $\min y_j \asymp 1/n$, $\max y_j / \min y_j \asymp n \log n$ and $\max y_j / \sum_{j=1}^{n} y_j = O_{\mathbb{P}}(n^{-1} \log n)$.*

Remark 30 *For positive sequences of random variables a_n, b_n, we write $a_n \asymp b_n$ if $a_n = O_{\mathbb{P}}(b_n)$ and $b_n = O_{\mathbb{P}}(a_n)$.*

Proof of Proposition 29. We first show that for any $\epsilon > 0$, there exists an M and N such that

$$\sup_{n>N} \mathbb{P}\left(\frac{\max y_j}{\log n} > M\right) \leq \epsilon, \sup_{n>N} \mathbb{P}\left(\frac{\log n}{\max y_j} > M\right) \leq \epsilon, \tag{55}$$

$$\sup_{n>N} \mathbb{P}\left(n \min y_j > M\right) \leq \epsilon, \sup_{n>N} \mathbb{P}\left(\frac{1}{n \min y_j} > M\right) \leq \epsilon. \tag{56}$$

For any $x > 0$, it can be checked that

$$\mathbb{P}(\max y_j \leq x) = (1 - e^{-x})^n,$$

which, for $n > N$, by Bernoulli's inequality and the basic inequality $\log(1 + x) < x$, implies

$$\begin{aligned}
\mathbb{P}(\max y_j \geq M \log n) =& 1 - (1 - n^{-M})^n \leq n^{-M+1} \leq N^{-M+1} \to 0, \\
\mathbb{P}\left(\max y_j \leq \frac{\log n}{M}\right) =& (1 - n^{-\frac{1}{M}})^n = e^{n \log(1-n^{-\frac{1}{M}})} \leq e^{-n^{1-\frac{1}{M}}} \leq e^{-N^{1-\frac{1}{M}}} \to 0,
\end{aligned}$$

as $N, M \to \infty$. This finishes the proof of (55).

For any $x > 0$, $\min y_j$ has the cumulative distribution function

$$\mathbb{P}(\min y_j \leq x) = 1 - (e^{-x})^n = 1 - e^{-nx}.$$

Therefore, we have

$$\mathbb{P}\bigg(\min y_j \geq \frac{M}{n}\bigg) = e^{-M}, \text{ and } \mathbb{P}\bigg(\min y_j \leq \frac{1}{nM}\bigg) = 1 - e^{-\frac{1}{M}} \to 0,$$

as $M \to \infty$, which finishes the proof of (56). Note (55) and (56) imply $\max y_j \asymp \log n$ and $\min y_j \asymp 1/n$, respectively. Because for positive sequences a_n, b_n, c_n, $a_n = O_{\mathbb{P}}(b_n)$ and $b_n = O_{\mathbb{P}}(c_n)$ implies $a_n = O_{\mathbb{P}}(c_n)$, we have $\max y_j / \min y_j \asymp n \log n$.

Next we show $\max y_j / \sum_{j=1}^n y_j = O_{\mathbb{P}}(n^{-1} \log n)$. Because we have shown that $\max y_j \asymp \log n$, it suffices to show $n = O_{\mathbb{P}}(\sum_{j=1}^n y_j)$, which is equivalent to show that for any $\epsilon > 0$, there exists an M and N such that

$$\sup_{n>N} \mathbb{P}\left(\sum_{j=1}^n y_j < \frac{n}{M}\right) \leq \epsilon. \tag{57}$$

By Chebyshev's inequality, for $n > N$

$$\begin{aligned}
&\mathbb{P}\left(\sum_{j=1}^n y_j < \frac{n}{M}\right) = \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n y_j - 1 < \frac{1}{M} - 1\right) \\
&\leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^n y_j - 1\right| > \left|\frac{1}{M} - 1\right|\right) \leq \frac{1}{n(1-1/M)^2} \leq \frac{1}{N(1-1/M)^2} \to 0,
\end{aligned}$$

as $N, M \to \infty$. This shows $n = O_{\mathbb{P}}(\sum_{j=1}^n y_j)$, and finishes the proof. ■

References

Robert A Adams and John JF Fournier. *Sobolev Spaces*, volume 140. Academic Press, 2003.

Robert J Adler and Jonathan E Taylor. *Random Fields and Geometry*. Springer Science & Business Media, 2009.

Rémi Arcangéli, María Cruz López de Silanes, and Juan José Torrens. An extension of a bound for functions in Sobolev spaces, with applications to (m, s)-spline interpolation and smoothing. *Numerische Mathematik*, 107(2):181–211, 2007.

M.J Bayarri, James O Berger, Rui Paulo, Jerry Sacks, John Cafeo, James Cavendish, Chin-Hsu Lin, and Jian Tu. A framework for validation of computer models. *Technometrics*, 49(138-154), 2007.

Moreno Bevilacqua, Tarik Faouzi, Reinhard Furrer, Emilio Porcu, et al. Estimation and prediction using generalized Wendland covariance functions under fixed domain asymptotics. *The Annals of Statistics*, 47(2):828–856, 2019.

S Borodachov, D Hardin, and E Saff. Asymptotics of best-packing on rectifiable sets. *Proceedings of the American Mathematical Society*, 135(8):2369–2380, 2007.

Rob Brownlee and Will Light. Approximation orders for interpolation by surface splines to rough functions. *IMA Journal of Numerical Analysis*, 24(2):179–192, 2004.

Adam D Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(Oct):2879–2904, 2011.

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

Ismaël Castillo. Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics*, 2:1281–1299, 2008.

Ismaël Castillo. On Bayesian supremum norm contraction rates. *The Annals of Statistics*, 42(5):2058–2091, 2014.

Jia Chen and Heping Wang. Average case tractability of multivariate approximation with Gaussian kernels. *Journal of Approximation Theory*, 239:51–71, 2019.

Andrew Chernih and Simon Hubbert. Closed form representations and properties of the generalised Wendland functions. *Journal of Approximation Theory*, 177:17–33, 2014.

Noel Cressie. *Statistics for Spatial Data*. John Wiley & Sons, 1993.

Felipe Cucker and Ding Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*, volume 24. Cambridge University Press, 2007.

Françoise Demengel, Gilbert Demengel, and Reinie Erné. *Functional Spaces for the Theory of Elliptic Partial Differential Equations*. Springer, 2012.

Michael F Driscoll. The reproducing kernel Hilbert space structure of the sample paths of a Gaussian process. *Probability Theory and Related Fields*, 26(4):309–316, 1973.

David Eric Edmunds and W Desmond Evans. *Spectral Theory and Differential Operators*. Oxford University Press, 2018.

David Eric Edmunds and Hans Triebel. *Function Spaces, Entropy Numbers, Differential Operators*, volume 120. Cambridge University Press, 2008.

GE Fasshauer, Fred J Hickernell, and H Woźniakowski. Average case approximation: convergence and tractability of Gaussian kernels. In *Monte Carlo and Quasi-Monte Carlo Methods 2010*, pages 329–344. Springer, 2012.

Gregory E Fasshauer and Michael J McCourt. *Kernel-based approximation methods using MATLAB*, volume 19. World Scientific Publishing Company, 2015.

Peter I Frazier. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Iosif I Gihman and Anatoli V Skorokhod. *The Theory of Stochastic Processes I*. Springer, 1974.

Matteo Giordano and Richard Nickl. Consistency of Bayesian inference with Gaussian process priors in an elliptic inverse problem. *arXiv preprint arXiv:1910.07343*, 2019.

T Gneiting. Stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97:590–600, 2002.

Chong Gu. *Smoothing Spline ANOVA Models*. Springer, 2013.

Philipp Hennig, Michael A Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150142, 2015.

V V Ivanov. On optimal minimization algorithms in classes of differentiable functions. *Dokl. Akad. Nauk SSSR*, 201:527–530, 1971.

V Roshan Joseph, Tirthankar Dasgupta, Rui Tuo, and C. F. Jeff Wu. Sequential exploration of complex surfaces using minimum energy designs. *Technometrics*, 57(1):64–74, 2015.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Alexey A Khartov and Marguerite Zani. Asymptotic analysis of average case approximation complexity of additive random fields. *Journal of Complexity*, 52:24–44, 2019.

Aaron Klein, Stefan Falkner, Simon Bartels, Philipp Hennig, and Frank Hutter. Fast Bayesian optimization of machine learning hyperparameters on large datasets. In *Artificial Intelligence and Statistics*, pages 528–536, 2017.

Mikhail Lifshits and Marguerite Zani. Approximation of additive random fields based on standard information: Average case and probabilistic settings. *Journal of Complexity*, 31 (5):659–674, 2015.

Wei-Liem Loh. Fixed-domain asymptotics for a subclass of Matérn-type Gaussian random fields. *The Annals of Statistics*, 33(5):2344–2394, 2005.

Harald Luschgy and Gilles Pagès. Sharp asymptotics of the functional quantization problem for Gaussian processes. *The Annals of Probability*, 32(2):1574–1599, 2004.

Harald Luschgy and Gilles Pagès. High-resolution product quantization for Gaussian processes under sup-norm distortion. *Bernoulli*, 13(3):653–671, 2007.

Georges Matheron. Principles of geostatistics. *Economic geology*, 58(8):1246–1266, 1963.

Stefan Müller. *Komplexität und Stabilität von kernbasierten Rekonstruktionsmethoden*. PhD thesis, Niedersächsische Staats-und Universitätsbibliothek Göttingen, 2009.

Thomas Müller-Gronbach and Klaus Ritter. Uniform reconstruction of Gaussian processes. *Stochastic Processes and Their Applications*, 69(1):55–70, 1997.

Francis Narcowich, Joseph Ward, and Holger Wendland. Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting. *Mathematics of Computation*, 74(250):743–763, 2005.

Francis J Narcowich. Recent developments in error estimates for scattered-data interpolation via radial basis functions. *Numerical Algorithms*, 39(1):307–315, 2005.

Francis J Narcowich and Joseph D Ward. Scattered data interpolation on spheres: error estimates and locally supported basis functions. *SIAM Journal on Mathematical Analysis*, 33(6):1393–1410, 2002.

Francis J Narcowich and Joseph D Ward. Scattered-data interpolation on $\mathbb{R}^n$: Error estimates for radial basis and band-limited functions. *SIAM Journal on Mathematical Analysis*, 36(1):284–300, 2004.

Francis J Narcowich, Joseph D Ward, and Holger Wendland. Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions. *Constructive Approximation*, 24(2):175–186, 2006.

Richard Nickl and Jakob Söhl. Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *The Annals of Statistics*, 45(4):1664–1693, 2017.

Harald Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*, volume 63. SIAM, 1992.

Erich Novak. *Deterministic and Stochastic Error Bounds in Numerical Analysis*, volume 1349. Springer, 2006.

Anargyros Papageorgiou and Grzegorz W Wasilkowski. On the average complexity of multivariate problems. *Journal of Complexity*, 6(1):1–23, 1990.

Debdeep Pati, Anirban Bhattacharya, and Guang Cheng. Optimal Bayesian estimation in random covariate design with a rescaled Gaussian process prior. *The Journal of Machine Learning Research*, 16(1):2837–2851, 2015.

Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press, 1999.

W Pogorzelski. *Integral Equations and Their Applications*, volume 1. Pergamon Press, 1966.

Carl Edward Rasmussen. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Sidney I Resnick. *Adventures in Stochastic Processes*. Springer Science & Business Media, 1992.

Klaus Ritter. *Average-Case Analysis of Numerical Problems*. Springer, 2007.

Klaus Ritter, Grzegorz W Wasilkowski, and Henryk Wozniakowski. Multivariate integration and approximation for random fields satisfying Sacks-Ylvisaker conditions. *The Annals of Applied Probability*, 5(2):518–540, 1995.

Dilip N Ghosh Roy and Louise S Couchman. *Inverse Problems and Inverse Scattering of Plane Waves*. Academic Press, 2001.

Jerome Sacks and Donald Ylvisaker. Designs for regression problems with correlated errors. *The Annals of Mathematical Statistics*, 37(1):66–89, 1966.

Jerome Sacks and Donald Ylvisaker. Designs for regression problems with correlated errors: many parameters. *The Annals of Mathematical Statistics*, 39(1):49–69, 1968.

Jerome Sacks and Donald Ylvisaker. Designs for regression problems with correlated errors iii. *The Annals of Mathematical Statistics*, 41(6):2057–2074, 1970.

Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.

Thomas J Santner, Brian J Williams, and William I Notz. *The Design and Analysis of Computer Experiments*. Springer Science & Business Media, 2003.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

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

Ingo Steinwart. Convergence types and rates in generic Karhunen-Loève expansions with applications to sample path properties. *Potential Analysis*, 51(3):361–395, 2019.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

Rui Tuo and C. F. Jeff Wu. A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):767–795, 2016.

Sara A van de Geer. *Empirical Processes in M-Estimation*, volume 6. Cambridge University Press, 2000.

Aad W van der Vaart and Harry van Zanten. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12(Jun):2095–2119, 2011.

Aad W van der Vaart and J Harry van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008a.

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

Aad W van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

Ramon van Handel. Probability in high dimension. Technical report, Princeton Univ. NJ, 2014.

Jan van Waaij and Harry van Zanten. Gaussian process methods for one-dimensional diffusions: Optimal rates and adaptation. *Electronic Journal of Statistics*, 10(1):628–645, 2016.

Christian Walck. Hand-book on statistical distributions for experimentalists. Technical report, 1996.

Wenjia Wang, Rui Tuo, and C. F. Jeff Wu. On prediction properties of kriging: Uniform error bounds and robustness. *Journal of the American Statistical Association*, 115(530): 920–930, 2020.

Holger Wendland. *Scattered Data Approximation*, volume 17. Cambridge University Press, 2004.

Tizian Wenzel, Gabriele Santin, and Bernard Haasdonk. A novel class of stabilized greedy kernel approximation algorithms: Convergence, stability & uniform point distribution. *arXiv preprint arXiv:1911.04352*, 2019.

George Wynne, François-Xavier Briol, and Mark Girolami. Convergence guarantees for Gaussian process approximations under several observation models. *arXiv preprint arXiv:2001.10818*, 2020.