On the inference of applying Gaussian process modeling to a deterministic function

Wenjia Wang

Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
e-mail: wenjiawang@ust.hk

Abstract: Gaussian process modeling is a standard tool for building emulators for computer experiments, which are usually used to study deterministic functions, for example, a solution to a given system of partial differential equations. This work investigates applying Gaussian process modeling to a deterministic function from prediction and uncertainty quantification perspectives, where the Gaussian process model is misspecified. Specifically, we consider the case where the underlying function is fixed and from a reproducing kernel Hilbert space generated by some kernel function, and the same kernel function is used in the Gaussian process modeling as the correlation function for prediction and uncertainty quantification. While upper bounds and the optimal convergence rate of prediction in the Gaussian process modeling have been extensively studied in the literature, a comprehensive exploration of convergence rates and theoretical study of uncertainty quantification is lacking. We prove that, if one uses maximum likelihood estimation to estimate the variance in Gaussian process modeling, under different choices of the regularization parameter value, the predictor is not optimal and/or the confidence interval is not reliable. In particular, lower bounds of the prediction error under different choices of the regularization parameter value are obtained. The results indicate that, if one directly applies Gaussian process modeling to a fixed function, the reliability of the confidence interval and the optimality of the predictor cannot be achieved at the same time.

Keywords and phrases: Gaussian process modeling, computer experiments, uncertainty quantification, Matérn correlation functions.

1. Introduction

Computer experiments are often used to study a system of interest. For example, [36] studies a complex simulation model for turbulent flows in swirl injectors. Other examples include [12], who estimates sexual transmissibility of human papillomavirus infection, and [40], who uses the Cardiovascular Disease Policy Model to project cost-effectiveness of treating hypertension in the U.S. according to 2014 guidelines. In these examples, the simulators are expensive, and the inputs/responses pairs are often not available for an extensive exploration of the underlying function. One well-established approach for solving this problem is the use of an *emulator*, which is an inexpensive approximation for the simulator.

In computer experiments, the two major problems are prediction and uncertainty quantification. Gaussian process modeling, which is a widely used method in computer experiments, naturally enables prediction and statistical uncertainty quantification. In the Gaussian process modeling, the underlying function is assumed to be a realization of a Gaussian process. Based on the Gaussian process assumption, the conditional distribution can be constructed at each unobserved point in a region, which provides a natural predictor via conditional expectation and a pointwise confidence interval. The pointwise confidence intervals can be used for statistical uncertainty quantification.

However, in practice, the Gaussian process is usually misspecified. The responses of a computer model usually come from a deterministic function which may not be a sample path of the Gaussian process used in Gaussian process modeling, or may come from a smaller function space that has probability zero [4, 11, 25, 55, 68]. For example, a function in an infinite-dimensional reproducing kernel Hilbert space is typically smoother than a sample path of the corresponding Gaussian process [52]. Here, the word "*corresponding*" means that the covariance function of the Gaussian process and the kernel function of the reproducing kernel Hilbert space are the same, up to a constant multiplier. Therefore, if one applies Gaussian process modeling to a function in the corresponding reproducing kernel Hilbert space, a model misspecification issue occurs.

Despite this model misspecification, maximum likelihood estimation is commonly used to estimate unknown parameters in the covariance function within the Gaussian process model [46]. Applying maximum likelihood estimation when a model is misspecified may be problematic because the estimated parameter can diverge as the sample size goes to infinity. For example, [67] shows that if the underlying function is $f(x) = x^\gamma$ on $[0, 1]$ and a Gaussian correlation function is used in the Gaussian process modeling, the estimated variance can either go to zero or infinity as the sample size increases to infinity. Another question is: when the Gaussian process model is misspecified, are the confidence intervals with estimated parameters reliable? In practice, it is often observed that Gaussian process models have poor coverage of their confidence intervals [23, 30, 69]. One possible reason is that the Gaussian process model may be misspecified; thus the confidence intervals may be inadequate for quantifying the uncertainty of predictions.

In this work, we investigate the prediction and confidence intervals in misspecified Gaussian process models used to recover deterministic functions from a frequentist view, i.e., assuming the underlying function is fixed but unknown. This is different from the Bayesian perspective, where a Gaussian process prior on the function space is induced. Specifically, we consider the following settings.

Settings 1.1. *The underlying deterministic function f is fixed and lies in a reproducing kernel Hilbert space. A Gaussian process model is applied for prediction and uncertainty quantification. The kernel function of the reproducing kernel Hilbert space and the correlation function in the Gaussian process model are the same.*

In Settings 1.1, we assume that the variance in the Gaussian process model

is unknown and needs to be estimated. For more on the formal settings considered in this work and more discussion, see Section 2.5. As stated before, the model misspecification occurs under Settings 1.1. This model misspecification does not change the form of the predictor (which is one reason that the Gaussian process model is typically misspecified), but significantly changes the uncertainty quantification results. We consider two cases: one case is that the observations have no noise; the other is that the observations have noise. When the observations have no noise, we show that if an estimated variance obtained by maximum likelihood estimation is used in the confidence intervals, then the confidence intervals are not reliable. Here, the reliability is used in the sense that is to be introduced later; see Section 2.3. This suggests that, if the Gaussian process model is misspecified, the confidence interval needs to be carefully constructed to quantify the uncertainties—not merely derived by the corresponding Gaussian process model.

In many cases, computer experiments are stochastic, in the sense that stochastic errors are introduced to simulate the randomness in real systems. A recent overview of stochastic emulators is [8]. In stochastic computer experiments, a regularization parameter is used to counteract the noise's influence. The value of this regularization parameter is usually a constant [3, 8, 14]. It is known that the optimal convergence rate for non-parametric regression is determined by the smoothness of the underlying function, denoted by ν. The optimal convergence rate is $n^{-\frac{\nu}{2\nu+d}}$, where n is the sample size, and d is the dimension of the input space [53]. In this work, we show that if the regularization parameter value is chosen to be a constant, the corresponding predictor is not optimal, in the sense that the convergence rate of the prediction error is of a higher order than the optimal rate. Furthermore, with an estimated variance obtained by maximum likelihood estimation, we show that under different choices of the regularization parameter value (not restricted to be a constant), the corresponding predictor is not optimal, or the confidence interval is not reliable. We also derive some lower bounds on the convergence rates of the prediction error of Gaussian process modeling. These results suggest that we may lose the prediction efficiency or reliability of uncertainty quantification if the Gaussian process model is misspecified and maximum likelihood estimation is used.

The rest of this paper is arranged as follows. In Section 2, we introduce Gaussian process modeling, maximum likelihood estimation, and reproducing kernel Hilbert spaces, as well as our definition of reliability of confidence intervals. The main results of this work are also summarized in Section 2. In Sections 3 and 4, we present the main results of this work, under the case where observations have no noise and the case where observations have noise, respectively. Simulation studies are reported in Section 5. Conclusions and discussion are made in Section 6. The technical proofs are given in Appendix.

2. Preliminaries

This section provides a brief introduction to Gaussian process modeling, maximum likelihood estimation, and reproducing kernel Hilbert spaces, which are

used in developing the main results. We also provide our definition of the reliability of confidence intervals. Problem settings and a summary of the main results are presented at the end of this section.

2.1. Gaussian process modeling

In this work, we consider applying Gaussian process modeling to a fixed function f, defined on a convex and compact set[1] $\Omega \subset \mathbb{R}^d$ with a positive Lebesgue measure. Because the domain Ω is fixed, the corresponding asymptotic framework is called fixed-domain asymptotics [50, 51]. Suppose we have n observed pairs $(x_k, y_k), k = 1, \ldots, n$, given by

$$y_k = f(x_k) + \epsilon_k, \tag{2.1}$$

where $x_k \in \Omega$ are distinct measurement locations (i.e., $x_k \neq x_j$ for $k \neq j$) and $\epsilon_k \sim N(0, \sigma_\epsilon^2)$ are i.i.d. normally distributed random errors with variance $\sigma_\epsilon^2 \geq 0$. If the observations are not corrupted by noise, we have $\sigma_\epsilon^2 = 0$, otherwise $\sigma_\epsilon^2 > 0$. One popular method to recover the function f is stationary Gaussian process modeling. Let Z be a stationary Gaussian process defined on $\mathbb{R}^d$. For the ease of mathematical treatment, we consider simple kriging. Therefore, we assume Z has mean zero, variance σ^2 and correlation function Ψ, denoted by $Z \sim GP(0, \sigma^2\Psi)$. The correlation function Ψ is stationary, i.e., the function value of $\Psi(x, x')$ only depends on the difference $x - x'$; thus we can write $\Psi(x - x') := \Psi(x, x')$. We also assume Ψ is strictly positive definite and integrable on $\mathbb{R}^d$, and $\Psi(0) = 1$. By Bochner's theorem (Page 208 of [19]; Theorem 6.6 of [64]) and Theorem 6.11 of [64], there exists a function f_Ψ such that

$$\Psi(h) = \int_{\mathbb{R}^d} e^{i\omega^T h} f_\Psi(\omega) d\omega$$

for any $h \in \mathbb{R}^d$. The function f_Ψ is known as the *spectral density* of Z or Ψ. In this work, we suppose that f_Ψ decays algebraically, i.e., satisfies the following condition.

Condition 2.1. *There exist constants $c_2 \geq c_1 > 0$ and $\nu > d/2$ such that, for all $\omega \in \mathbb{R}^d$,*

$$c_1(1 + \|\omega\|_2^2)^{-\nu} \leq f_\Psi(\omega) \leq c_2(1 + \|\omega\|_2^2)^{-\nu},$$

where $\|\cdot\|_2$ denotes the Euclidean metric.

One example of correlation functions satisfying Condition 2.1 is the isotropic Matérn correlation function [51], given by

$$\Psi_M(h) = \frac{1}{\Gamma(\nu)2^{\tilde{\nu}-1}} (2\sqrt{\tilde{\nu}}\phi\|h\|_2)^{\tilde{\nu}} K_{\tilde{\nu}}(2\sqrt{\tilde{\nu}}\phi\|h\|_2), \tag{2.2}$$

---

[1] This condition can be relaxed to a compact set satisfying interior cone condition and with Lipschitz boundary; see [1, 64] for discussion of these conditions. In fact, the compactness and convexity imply the interior cone condition and Lipschitz boundary; see [26, 41].

with the spectral density [57]

$$f_\Psi(\omega; \tilde{\nu}, \phi) = \pi^{-d/2} \frac{\Gamma(\tilde{\nu} + d/2)}{\Gamma(\tilde{\nu})} (4\tilde{\nu}\phi^2)^{\tilde{\nu}} (4\tilde{\nu}\phi^2 + \|\omega\|_2^2)^{-(\tilde{\nu}+d/2)},$$

where $\phi, \tilde{\nu} > 0$, and $K_{\tilde{\nu}}$ is the modified Bessel function of the second kind. By setting $\nu = \tilde{\nu} + d/2$, we can see Ψ_M satisfies Condition 2.1.

Another example of correlation functions satisfying Condition 2.1 is the generalized Wendland correlation function [9, 13, 21], given by

$$\Psi_{GW}(h) = \begin{cases} \frac{1}{B(2\kappa, \mu+1)} \int_h^1 u(u^2 - h^2)^{\kappa-1}(1-u)^\mu du, & 0 \leq h < 1, \\ 0, & h \geq 1, \end{cases}$$

where $\kappa > 0$ and $\mu \geq (d+1)/2 + \kappa$, and B is the beta function. It can be shown that Ψ_{GW} also satisfies Condition 2.1; see Theorem 1 of [9].

Suppose we are interested in the value of $Z(x)$ on Ω and we observe data

$$z_j = Z(x_j) + \epsilon'_j, j = 1, ..., n, \tag{2.3}$$

where $x_k \in \Omega$ are distinct measurement locations (i.e., $x_k \neq x_j$ for $k \neq j$) and $\epsilon'_k \sim N(0, \sigma_\epsilon^2)$ are i.i.d. normally distributed random errors with variance $\sigma_\epsilon^2 \geq 0$. Conditional on $\mathcal{Z} = (z_1, ..., z_n)^T$, $Z(x)$ is normally distributed at point x. Note that $\mathcal{Z}$ is a random vector, where the randomness is induced by $Z(x_j)$ and ϵ'_j. The conditional expectation of $Z(x)$ is given by

$$\mathbb{E}[Z(x)|\mathcal{Z}] = r(x)^T(R + \mu I_n)^{-1}\mathcal{Z}, \tag{2.4}$$

where $r(x) = (\Psi(x - x_1), \ldots, \Psi(x - x_n))^T$, $R = (\Psi(x_j - x_k))_{jk}$, I_n is an identity matrix, and $\mu = \sigma_\epsilon^2/\sigma^2$. The conditional expectation is a nature predictor of $Z(x)$, and it can be shown that the conditional expectation (2.4) is the best linear unbiased predictor [46, 51]. A predictor given by Gaussian process modeling is then the conditional expectation of $Z(x)$.

In addition to prediction, uncertainty quantification plays an essential role in statistics. Gaussian process modeling enables statistical uncertainty quantification via confidence intervals [45, 46]. Conditional on $\mathcal{Z}$, the conditional variance of $Z(x)$ is given by

$$\mathrm{Var}[Z(x)|\mathcal{Z}] = \sigma^2(1 - r(x)^T(R + \mu I)^{-1}r(x)),$$

where R, $r(x)$ and μ are as in (2.4). Let Φ denote the cumulative distribution function of the standard normal distribution $N(0, 1)$ and let $q_\beta = \Phi^{-1}(1 - \beta/2)$ denote the $(1 - \beta/2)$th quantile, where $\beta \in (0, 1)$. A level $(1 - \beta)100\%$ pointwise confidence interval on point $x \in \Omega$ can be constructed by

$$CI_{n,\beta}(x) = [\mathbb{E}[Z(x)|\mathcal{Z}] - c_{n,\beta}(x), \mathbb{E}[Z(x)|\mathcal{Z}] + c_{n,\beta}(x)],$$

where

$$c_{n,\beta}(x) = q_{1-\beta/2}\sqrt{\mathrm{Var}[Z(x)|\mathcal{Z}]} = q_{1-\beta/2}\sqrt{\sigma^2(1 - r(x)^T(R + \mu I_n)^{-1}r(x))}, \tag{2.5}$$

and $r(x)$ and R are as in (2.4). The confidence interval is often used in the numerical simulations to show the uncertainty quantification results in Gaussian process modeling. A few examples are [4, 22, 28, 65].

Recall that we apply Gaussian process modeling to the deterministic function f. Therefore, we *treat* the observations $Y = (y_1, ..., y_n)^T$ as the observations from the Gaussian process Z. The predictor of $f(x)$ becomes

$$f_n(x) := \mathbb{E}[Z(x)|y_1, \ldots, y_n] = r(x)^T (R + \mu I_n)^{-1} Y, x \in \Omega, \qquad (2.6)$$

and the confidence interval becomes

$$CI_{n,\beta}(x) = [f_n(x) - c_{n,\beta}(x), f_n(x) + c_{n,\beta}(x)], \qquad (2.7)$$

where $c_{n,\beta}(x)$ is as in (2.5).

2.2. Maximum likelihood estimation

Recall that we consider that the underlying function is deterministic and lies in some reproducing kernel Hilbert space (see Settings 1.1). The Gaussian process modeling is used for prediction and uncertainty quantification, where the correlation function is the same as the kernel function of the reproducing kernel Hilbert space. Despite the model misspecification issue, i.e., the underlying function is typically smoother than the sample path in the corresponding Gaussian process, maximum likelihood estimation [45, 46, 51] is often used to specify the parameters values that are not pre-determined in the covariance function $\sigma^2 \Psi(\cdot - \cdot)$. Let Ψ_θ be a family of correlation functions indexed by $\theta = (\theta_1, \ldots, \theta_q)^T \in \Theta \subset \mathbb{R}^q$, and $R(\theta) = (\Psi_\theta(x_j - x_k))_{jk}$. By direct calculation and reparametrization, it can be shown that, up to an additive constant, the log-likelihood function is (Page 169 of [51]; Page 66 of [46])

$$\ell(\theta, \sigma^2, \mu; X, Y) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2} \log \det(R(\theta) + \mu I_n) - \frac{Y^T (R(\theta) + \mu I_n)^{-1} Y}{2\sigma^2}. \tag{2.8}$$

The maximum likelihood estimate of the unknown parameters θ, σ^2, μ can be found by maximizing the log-likelihood function. In practice, it is often assumed that the scale parameter, which is ϕ in (2.2) if a Matérn correlation function is used, and the variance σ^2 are unknown and need to be estimated [5, 10, 27, 29, 34, 35, 45, 46, 51]. The smoothness parameter $\tilde{\nu}$ in the Matérn correlation function (2.2) is usually pre-determined. We will discuss the parameter μ later. The consistency of parameter estimation has been studied in literature under the assumption that the underlying truth is a realization of a Gaussian process [2, 9, 32, 37, 61, 72, 73, 75]. In particular, [9, 32, 61, 72, 75] show that the parameter estimation can be inconsistent and only the microergodic parameter can be estimated. For details of microergodic parameters, see [39]; also see pages 163-165 of [51].

If the underlying truth is a fixed function that is not a sample path of the Gaussian process used in the Gaussian process modeling, to the best of our knowledge, the only related work is [67]. In [67], it is shown that if the underlying function is $f(x) = x^\gamma$ with $\gamma \geq 0$ defined on $[0, 1]$ and a Gaussian correlation function is used in the Gaussian process modeling, the estimated variance can either go to infinity or converge to zero as the sample size increases. Other works related to the parameter estimation under model misspecification include [6, 7].

In this work, we do not consider the consistency of parameter estimation, but we investigate the influence of parameter estimation on the prediction and uncertainty quantification of Gaussian process modeling under model misspecification. For the ease of mathematical treatment, we assume Ψ is known. It follows the standard arguments that the maximizer of (2.8) with respect to σ^2 is

$$\hat{\sigma}^2 = Y^T (R + \mu I_n)^{-1} Y / n. \tag{2.9}$$

Similar settings have also been considered by [31], where they call σ a scale parameter and consider only estimating σ. In practice, μ is usually imposed as a constant [14], estimated by the sample average if there are replicates on each measurement location [3], or estimated via maximum likelihood estimation [62]. We mainly focus on the first approach (using an imposed $\hat{\mu}_n$) and note that the results can be easily generalized to the case of using the second approach. The third approach is much more complicated, and we do not consider it in the present work. Here we use $\hat{\mu}_n$ to stress the difference, because it may not be the true value of μ. Recall that the underlying truth is a deterministic function from a frequentist point of view. Thus there is no definition of the "true" value of σ^2. It is also not clear what is the "true" $\mu = \sigma_\epsilon^2 / \sigma^2$. In this work, we call $\hat{\mu}_n$ a *regularization parameter*, because this parameter is *imposed*. Nevertheless, we consider that the value of the regularization parameter $\hat{\mu}_n$ can increase or decrease with the sample size and is not restricted to be a constant.

By plugging in estimated (and imposed) parameters, we can obtain the corresponding predictor and confidence interval. We use $\hat{f}_n(x)$ to denote the predictor with imposed $\hat{\mu}_n$ on an unobserved point x, i.e.,

$$\hat{f}_n(x) = r(x)^T (R + \hat{\mu}_n I_n)^{-1} Y, \tag{2.10}$$

where $r(x)$ and R are as in (2.4). By plugging $\hat{\sigma}^2$ as in (2.9) and $\hat{\mu}_n$ into (2.7) and (2.5), we obtain the estimated pointwise confidence interval on point $x \in \Omega$

$$\widehat{CI}_{n,\beta}(x) = [\hat{f}_n(x) - \hat{c}_{n,\beta}(x), \hat{f}_n(x) + \hat{c}_{n,\beta}(x)], \tag{2.11}$$

where

$$\hat{c}_{n,\beta}(x) = q_{1-\beta/2} \sqrt{\hat{\sigma}^2 (1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x))}, \tag{2.12}$$

and $\hat{f}_n(x)$ is as in (2.10). In (2.12), we impose a regularization parameter $\hat{\mu}_n > 0$ if $\sigma_\epsilon^2 > 0$, and set $\hat{\mu}_n = 0$ if $\sigma_\epsilon^2 = 0$. Note that the estimated variance $\hat{\sigma}^2$ is not present in (2.10); thus it does not influence the predictor. However,

$\hat{\sigma}^2$ appears in (2.11) and as we will see later, it influences the reliability of the confidence interval; thus the uncertainty quantification results of Gaussian process modeling.

2.3. Reliability of confidence intervals

In practice, the Gaussian process model is often misspecified. The underlying fixed function may lie in a subspace of the support of the corresponding Gaussian process and the subspace may have probability zero, or may not even be in the support. This model misspecification may influence the reliability of the confidence interval thus the quality of uncertainty quantification. However, it is not possible to quantify the reliability of confidence intervals without having a clear definition of the term "reliability". In this section, we first review some possible ways to define the reliability, and propose our definition of the reliability of confidence intervals.

Recall that in this work, we assume that the underlying truth is a deterministic function. Therefore, we mainly consider the reliability of confidence intervals for a fixed function. Let $g \in \mathcal{G}$ be a fixed function, where $\mathcal{G}$ is a Hilbert space of functions equipped with norm $\|\cdot\|_{\mathcal{G}}$. Let $I_X g$ be a linear predictor for a function $g \in \mathcal{G}$, where $X = \{x_1, ..., x_n\} \subset \Omega$ is the set of measurement locations. The predictor $I_X g$ depends on X and observations. Suppose the observations are $y_j^{(g)}$ for $j = 1, ..., n$, given by

$$y_j^{(g)} = g(x_j) + \epsilon_j, \tag{2.13}$$

where ϵ_j's are i.i.d. noise realizations of a random variable with mean zero and variance $\sigma_\epsilon^2 \in [0, \infty)$. Typically, $I_X g$ is a linear combination of the observations, i.e., has the form

$$I_X g(x) = \sum_{j=1}^{n} b_j(x) y_j^{(g)} \tag{2.14}$$

for point $x \in \Omega$, where b_j's are functions not depending on g but can depend on X. Let $CI_{X,\beta}(x) = [I_X g(x) - a_\beta(x), I_X g(x) + a_\beta(x)]$ be an *imposed* level $(1 - \beta)100\%$ pointwise confidence interval on point $x \in \Omega$ (determined by the uncertainty quantification method that a user applies), where $\beta \in (0, 1)$ and a_β is a non-negative function. Clearly, this imposed confidence interval may not have confidence level $(1 - \beta)100\%$. We want to define the reliability of this imposed confidence interval. Note that $CI_{X,\beta}$ and a_β can depend on X, but we suppress the dependency for notational simplicity. Also note the confidence interval on point x is centered at $I_X g(x)$.

Probably the most natural way to define the reliability is by the definition of confidence intervals. This approach is considered by [48]. Consider the probability $\mathbb{P}_g(g(x) \in CI_{X,\beta}(x))$ for point $x \in \Omega$, where $\mathbb{P}_g$ refers to the distribution of $y_1^{(g)}, ..., y_n^{(g)}$ as in (2.13), where the "true" g is given. If confidence intervals

are reliable, the probability $\mathbb{P}_g(g(x) \in CI_{X,\beta}(x))$ should be close to the nominal level $(1 - \beta)100\%$, or at least larger than $(1 - \beta)100\%$ (conservative). In [48], a function defined on $[0, 1]$ and a Brownian motion prior are considered. The measurement locations are equally spaced. Under these settings, [48] shows that $CI_{X,\beta}$ can be conservative or unreliable, depending on the smoothness of g. However, as stated in [48], the exact formulas strongly depend on the equally spaced measurement locations and cannot be easily extended to a more general choice of measurement locations.

Another probability-based definition of reliability of confidence intervals is by the average coverage probability (ACP) [43]. In [43], confidence intervals are considered to be reliable if the ACP for the function g and the confidence interval $CI_{X,\beta}$

$$\frac{1}{n} \sum_{j=1}^{n} \mathbb{P}(g(x_j) \in CI_{X,\beta}(x_j))$$

is close to the nominal level $(1 - \beta)100\%$, where x_j and $g(x_j)$ are as in (2.13). This definition only quantifies the reliability of the confidence interval on the measurement locations and does not count the confidence interval $CI_{X,\beta}(x)$ at any *unobserved* point $x \in \Omega$. Therefore, the ACP is not suitable to be used for quantifying the uncertainties because if the observations are noiseless and an interpolant is used, the ACP is always equal to one.

Coverage rates are often used to assess the reliability of the confidence interval in the field of computer experiments [30, 34, 54]. The coverage rate is defined by

$$\frac{\mathrm{Vol}(\{x|g(x) \in CI_{X,\beta}(x)\})}{\mathrm{Vol}(\Omega)}, \tag{2.15}$$

where $\mathrm{Vol}(A)$ denotes the volume of a set $A \subset \Omega$ with respect to the Lebesgue measure. A practical way to compute the coverage rate is by random sampling. Suppose $x'_1, ..., x'_N$ are N uniformly distributed points in Ω. Then the coverage rate can be approximated by

$$\frac{\mathrm{card}(\{x'_j|g(x'_j) \in CI_{X,\beta}(x'_j)\})}{N},$$

where $\mathrm{card}(B)$ denotes the cardinality of a set B. However, we find it is hard to theoretically investigate the quantity (2.15), because $\{x|g(x) \in CI_{X,\beta}(x)\}$ can be irregular and hard to characterize.

In this work, we consider the ratio of the prediction error and the width of the confidence interval, given by $(g - I_X g)/|CI_{X,\beta}|$, where $|CI_{X,\beta}| = 2a_\beta$ denotes the width of $CI_{X,\beta}$. We use the convention $0/0 = 0$ if $|CI_{X,\beta}(x)| = 0$ for some $x \in \Omega$. If the confidence interval $CI_{X,\beta}$ is reliable, the width of the confidence interval should be large enough to cover the difference between the predictor $I_X g$ and the true function g with high probability such that the ratio $|g(x) - I_X g(x)|/|CI_{X,\beta}(x)|$ is small for $x \in \Omega$. In particular, we consider the

expectation $\left(\mathbb{E}\|(g - I_X g)/|CI_{X,\beta}|\|_{L_p(\Omega)}^p\right)^{1/p}$ for $2 \leq p \leq \infty$ (we assume it exists; if it does not exist, then the confidence interval is thought to be not reliable), where the expectation is taken with respect to the noise and the set of measurement locations X, and $\|f\|_{L_p(\Omega)}$ is the L_p-norm of $f \in L_p(\Omega)$, defined by

$$\|f\|_{L_p(\Omega)}^p = \int_\Omega |f(x)|^p dx.$$

The expectation $\left(\mathbb{E}\|(g - I_X g)/|CI_{X,\beta}|\|_{L_p(\Omega)}^p\right)^{1/p}$ is the L_p-norm on the probability space $(\mathcal{A}, \mathcal{B}, P)$, where $\mathcal{A}$ is the sample space, $\mathcal{B}$ is the Borel algebra, and P is the probability measure induced by the noise ϵ and X. Note that the randomness in $(g - I_X g)/|CI_{X,\beta}|$ does not come from the function g, because g is fixed from a frequentist perspective. Because we are interested in the scenario when the number of measurement locations increases, we consider an infinite sequence of the set of measurement locations, denoted by $\mathcal{X} = \{X_1, X_2, ..., X_n, ...\}$. Without loss of generality, we assume that $\mathrm{card}(X_n) = n$, where n takes its value in an infinite subset of $\mathbb{N}_+$. We call $\mathcal{X}$ a *sampling scheme*, as in [56]. In the rest of this work, we suppress the dependency of X on n for notational simplicity. If the confidence interval $CI_{X,\beta}$ is reliable, $\left(\mathbb{E}\|(g - I_X g)/|CI_{X,\beta}|\|_{L_p(\Omega)}^p\right)^{1/p}$ should be small, at least should be less than a constant that does not depend on the sample size. From a standard frequentist perspective, we consider the *minimax* setting, i.e., we consider the *worst* case.

According to the prior knowledge on the function g, we consider two subcases. Recall that $g \in \mathcal{G}$, where $\mathcal{G}$ is a Hilbert space of functions equipped with norm $\|\cdot\|_\mathcal{G}$. The first subcase is that $\|g\|_\mathcal{G}$ is upper bounded by some known constant. Without loss of generality, assume this known constant is one, i.e., the function g lies in the unit ball of $\mathcal{G}$. We say the confidence interval $CI_{X,\beta}$ is L_p-*weakly-reliable*, if

$$\sup_{g \in \mathcal{G}, \|g\|_\mathcal{G} \leq 1} \left(\mathbb{E}\|(g - I_X g)/|CI_{X,\beta}|\|_{L_p(\Omega)}^p\right)^{1/p} \leq C \qquad (2.16)$$

holds for all $X \in \mathcal{X}$ and all n, where C is a constant not depending on n. In other words, the confidence interval is weakly-reliable if it is reliable in a ball of $\mathcal{G}$ with certain radius. However, in practice we cannot always expect $\|g\|_\mathcal{G}$ to be bounded by a *known* constant. Since g is a fixed function in $\mathcal{G}$, we know $\|g\|_\mathcal{G}$ is finite. Therefore, for any increasing sequence $\{a_n\}_{n \geq 1}$ not depending on g and $\lim_{n\to\infty} a_n = \infty$, there exists an N such that for all $n \geq N$, $\|g\|_\mathcal{G} \leq a_n$. We say the confidence interval $CI_{X,\beta}$ is L_p-*strongly-reliable*, if there exists an increasing sequence $\{a_n\}_{n \geq 1}$ not depending on g such that $\lim_{n\to\infty} a_n = \infty$ and

$$\sup_{g \in \mathcal{G}, \|g\|_\mathcal{G} \leq a_n} \left(\mathbb{E}\|(g - I_X g)/|CI_{X,\beta}|\|_{L_p(\Omega)}^p\right)^{1/p} \leq C' \qquad (2.17)$$

holds for all $X \in \mathcal{X}$ and all n, where C' is a constant not depending on n. Here we note that the constants C and C' can depend on $\mathcal{X}$. Roughly speaking, a confidence interval is strongly-reliable if it is eventually reliable in the entire space $\mathcal{G}$ as the sample size increases to infinity. We summarize the above arguments in the following definition. Note in Definition 2.1, we suppress the dependency of X on n for notational simplicity.

Definition 2.1. *Let $\beta \in (0,1)$ be fixed, and $\mathcal{X}$ be a sampling scheme. Let $I_X g$ be a linear predictor as in (2.14) and $n = \mathrm{card}(X)$ be the sample size. Let $CI_{X,\beta}$ be an imposed level $(1-\beta)100\%$ confidence interval centered at $I_X g$ with $\lim_{n \to \infty} \sup_{x \in \Omega} |CI_{X,\beta}(x)| = 0$, where $|CI_{X,\beta}(x)|$ is the width of $CI_{X,\beta}(x)$.*

For $2 \leq p \leq \infty$, $CI_{X,\beta}$ is said to be L_p-weakly-reliable if (2.16) holds for all n, and is said to be L_p-strongly-reliable if there exists an increasing sequence $\{a_n\}_{n \geq 1}$ not depending on g, and $\lim_{n \to \infty} a_n = \infty$ such that for all n, (2.17) holds, where C and C' are constants not depending on n but possibly depending on p, β, and $\mathcal{X}$. The expectation is taken with respect to noise and X.

Remark 2.1. *In Definition 2.1, the reason we require the width of the confidence interval satisfies $\lim_{n \to \infty} \sup_{x \in \Omega} |CI_{X,\beta}(x)| = 0$ because we want the confidence interval to provide some information, otherwise we can select a wide confidence interval (for example, $CI_{X,\beta}(x) = [I_X g(x) - n, I_X g(x) + n]$ for all $x \in \Omega$) which can cover $g(x)$ and does not provide any useful information.*

Definition 2.1 is motivated by the properties of confidence intervals of Gaussian process. Let $Z \sim GP(0, \sigma^2 \Psi)$ be a Gaussian process defined on Ω. On point $x \in \Omega$, let $I_X^{(1)} Z(x) = \mathbb{E}[Z(x) | \mathcal{Z}]$, where $\mathbb{E}[Z(x) | \mathcal{Z}]$ is as in (2.4). It can be seen that $I_X^{(1)} Z$ is a linear predictor and has the form as in (2.14). Let $CI_{n,\beta}(x)$ be the confidence interval as in (2.7). Furthermore, assume the observations are not corrupted by noise, which implies $\sigma_\epsilon^2 = 0$ and $\mu = 0$. Consider $\left(\mathbb{E} \| (Z - I_X^{(1)} Z) / |CI_{n,\beta}| \|_{L_p(\Omega)}^p \right)^{1/p}$. We have the following proposition.

Proposition 2.1. *Let $Z, I_X^{(1)} Z, CI_{n,\beta}$ be described above, and $\beta \in (0,1)$. Then we have*

$$\left(\mathbb{E} \| (Z - I_X^{(1)} Z) / |CI_{n,\beta}| \|_{L_p(\Omega)}^p \right)^{1/p} = C \tag{2.18}$$

holds for all n and any $2 \leq p < \infty$, where C is a constant only depending on p, β and Ω.

It can be seen that our definition of reliability stated in Definition 2.1 is analogous to (2.18). According to Definition 2.1, if a confidence interval $CI_{X,\beta}$ is reliable, then for any fixed constant $c > 0$, $cCI_{X,\beta} := [I_X g(x) - ca_\beta(x), I_X g(x) + ca_\beta(x)]$ is also reliable. Furthermore, the "less than or equal to" relationship in Definition 2.1 encourages a wider confidence interval. Therefore, our definition of the reliability is more like a *necessary* condition rather than a sufficient condition. One way to specify the constant in Definition 2.1 is by using the constant C in Proposition 2.1. However, one can argue that this constant may not be

appropriate because unlike the unbiased predictor $I_X^{(1)} Z$, $I_X g$ is usually a biased predictor, and the constant in Proposition 2.1 may not be large enough to cover the bias. Practitioners may also consider other constants to counteract the model misspecification. How to choose an appropriate constant is out of the scope of this work, and we do not make any further discussion.

2.4. *Reproducing kernel Hilbert spaces and power functions*

In this subsection, we review reproducing kernel Hilbert spaces and power functions, which are closely related to the Gaussian process model. Under the settings of computer experiments, if $\mu = 0$ in the Gaussian process model, the right-hand side of (2.6) is called a kriging interpolant [63], denoted by

$$\mathcal{I}_{\Psi,X} f(x) = r(x)^T R^{-1} Y, \tag{2.19}$$

where $X = \{x_1, ..., x_n\}$ denotes the set of measurement locations. Note that $x_k \in \Omega$ are distinct measurement locations and Ψ is strictly positive definite, thus R is invertible. In the area of scattered data approximation, the interpolation using operator $\mathcal{I}_{\Psi,X}$ is also called radial basis function approximation. A standard theory of radial basis function approximation works by employing *reproducing kernel Hilbert spaces*. One way to define the reproducing kernel Hilbert space generated by a stationary correlation function is via the Fourier transform, defined by

$$\mathcal{F}(f)(\omega) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x) e^{-ix^T \omega} dx$$

for $f \in L_1(\mathbb{R}^d)$. The definition of the reproducing kernel Hilbert space can be generalized to $f \in L_2(\mathbb{R}^d) \cap C(\mathbb{R}^d)$. See [20] and Theorem 10.12 of [64].

Definition 2.2. *Let Ψ be a stationary correlation function that is integrable on $\mathbb{R}^d$. Define the reproducing kernel Hilbert space $\mathcal{N}_\Psi(\mathbb{R}^d)$ generated by Ψ as*

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

with the inner product

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

For a positive number $\nu > d/2$, the Sobolev space on $\mathbb{R}^d$ with smoothness ν can be defined as

$$H^\nu(\mathbb{R}^d) = \{f \in L_2(\mathbb{R}^d) : |\mathcal{F}(f)(\cdot)|(1 + \|\cdot\|_2^2)^{\nu/2} \in L_2(\mathbb{R}^d)\},$$

equipped with an inner product

$$\langle f, g \rangle_{H^\nu(\mathbb{R}^d)} = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \mathcal{F}(f)(\omega) \overline{\mathcal{F}(g)(\omega)} (1 + \|\omega\|_2^2)^\nu d\omega.$$

It can be shown that $H^\nu(\mathbb{R}^d)$ coincides with the reproducing kernel Hilbert space $\mathcal{N}_\Psi(\mathbb{R}^d)$, if Ψ satisfies Condition 2.1 ([64], Corollary 10.13, also see Lemma C.3). This equivalence allows us to evaluate whether a predictor in a reproducing kernel Hilbert space is optimal; see Section 4 for more details.

Reproducing kernel Hilbert spaces can also be defined on a suitable subset (for example, convex and compact) $\Omega \subset \mathbb{R}^d$, denoted by $\mathcal{N}_\Psi(\Omega)$, with norm

$$\|f\|_{\mathcal{N}_\Psi(\Omega)} = \inf\{\|f_E\|_{\mathcal{N}_\Psi(\mathbb{R}^d)} : f_E \in \mathcal{N}_\Psi(\mathbb{R}^d), f_E|_\Omega = f\},$$

where $f_E|_\Omega$ denotes the restriction of f_E to Ω. Sobolev spaces on Ω can be defined in a similar way.

If $f \in \mathcal{N}_\Psi(\Omega)$, there is a simple error bound ([64], Theorem 11.4):

$$|f(x) - \mathcal{I}_{\Psi,X} f(x)| \le P_{\Psi,X}(x)\|f\|_{\mathcal{N}_\Psi(\Omega)}, \tag{2.20}$$

for each $x \in \Omega$, where $P_{\Psi,X}$ is a function independent of f. The square of $P_{\Psi,X}$ is called the *power function*, given by

$$P_{\Psi,X}^2(x) = 1 - r(x)^T R^{-1} r(x)$$

for each $x \in \Omega$, where $r(x)$ and R are as in (2.4). In addition, we define

$$\mathcal{P}_{\Psi,X} := \sup_{x\in\Omega} P_{\Psi,X}(x). \tag{2.21}$$

Note that the power function $P_{\Psi,X}$ and its supremum $\mathcal{P}_{\Psi,X}$ only depend on X, Ω and Ψ, and does not depend on the observations.

2.5. Problem settings and summary of results

In this work, we consider the inference of misspecified Gaussian process models. Specifically, we consider prediction and uncertainty quantification when applying Gaussian process modeling to a fixed function $f \in \mathcal{N}_\Psi(\Omega)$, under the following misspecified model assumption.

Assumption* 2.1 (Misspecified model assumption)**.** *The function f is a realization of a Gaussian process with mean zero and covariance function $\sigma^2\Psi$ with a finite $\sigma > 0$.*

We use an asterisk "*" to denote that Assumption* 2.1 is a *misspecified* assumption, and is not true. After the earliest version of this work was submitted, Assumption* 2.1 was also considered by [31]. Under Assumption* 2.1, we incorrectly assume f is a realization of $Z \sim GP(0, \sigma^2\Psi)$ for $f \in \mathcal{N}_\Psi(\Omega)$. Assumption* 2.1 is a misspecified model assumption because if $Z \sim GP(0, \sigma^2\Psi)$, then $\mathbb{P}(Z \in \mathcal{N}_\Psi(\Omega)) = 0$ since Ψ satisfies Condition 2.1 and Ω is convex and compact with positive Lebesgue measure [16]. In fact, the smoothness of the sample paths are at least $d/2$ different from the smoothness of the correlation function, if ν in Condition 2.1 is larger than d [52]. The different assumptions of $f \in \mathcal{N}_\Psi(\Omega)$ and f is a realization of $Z \sim GP(0, \sigma^2\Psi)$ yield the same predictor,

but the prediction error analysis methodologies are completely different. For discussion of these two different assumptions, see [47].

Under the misspecified model assumption Assumption* 2.1, one can use maximum likelihood estimation to "estimate" the unknown parameters and impose confidence intervals. Of course, this is questionable, but it is widely used in practice as stated in Section 1, and also in numerical examples showing in research papers. In these synthetic numerical examples, the test function is typically chosen to be a fixed function with closed form (and usually infinitely differentiable), which naturally satisfies the condition $f \in \mathcal{N}_{\Psi}(\Omega)$. Under Assumption* 2.1, we show the following results:

(i) If the observations are not corrupted by noise, then the confidence interval is not L_p-weakly-reliable for $p \in (2, \infty]$, and is not L_2-strongly-reliable.

(ii) If the observations are corrupted by noise, then the confidence interval is not L_2-strongly-reliable, or the predictor is not optimal, in the sense that the predictor does not achieve the optimal convergence rate under L_2 metric.

In the rest of this work, we will use the following definitions. For two positive sequences a_n and b_n, we write $a_n \asymp b_n$ if, for some constants $C, C' > 0$, $C \leq a_n/b_n \leq C'$. Similarly, we write $a_n \gtrsim b_n$ and $b_n \lesssim a_n$ if $a_n \geq C b_n$ for some constant $C > 0$. For notational simplicity, we will use $C, C', C_1, C_2, ...$ and $\eta, \eta_0, \eta_1, ...$ to denote the constants, of which the values can change from line to line.

3. When the observations are noiseless

In this section, we consider the case that the observations have no noise. We call this case *deterministic case*, because several measurements at the same location will always lead to the same response.

3.1. The unreliability of the confidence interval

We focus on the Matérn correlation function, defined in (2.2). Since ϕ and $\tilde{\nu}$ are known, we can let $\phi = 1/(2\sqrt{\tilde{\nu}})$, because otherwise we can stretch the region Ω to adjust the scale parameter ϕ. After a proper reparametrization, we can rewrite (2.2) as

$$\Psi_M(h) = \frac{1}{\Gamma(\nu - d/2)2^{\nu - d/2 - 1}} \|h\|_2^{\nu - d/2} K_{\nu - d/2}(\|h\|_2) \qquad (3.1)$$

for $h \in \mathbb{R}^d$, where $\nu > d/2$. We set $\Psi = \Psi_M$ in this section.

Recall that in the deterministic case, $\sigma_\epsilon^2 = 0$, thus $\epsilon_k = 0$, $k = 1, ..., n$, $\mu = 0$ and $\hat{\mu}_n = 0$. The predictor $\hat{f}_n(x)$ in (2.10) becomes a kriging interpolant (2.19), i.e.,

$$\hat{f}_n(x) = \mathcal{I}_{\Psi, X} f(x) = r(x)^T R^{-1} Y \qquad (3.2)$$

for any point $x \in \Omega$, where $r(x)$ and R are as in (2.4), and $Y = (y_1, ..., y_n)^T$. Because the observations are not corrupted by noise, we have $y_k = f(x_k)$, for $k = 1, ..., n$. Note that in (3.2), the variance is not present and there is no estimated or imposed parameter.

As stated in Section 2.3, for $\beta \in (0, 1)$, an imposed confidence interval with estimated variance at point $x \in \Omega$ can be constructed by plugging $\hat{\mu}_n = 0$ in (2.11). The confidence interval is given by

$$\widehat{CI}_{n,\beta}(x) = [\hat{f}_n(x) - \hat{c}_{n,\beta}(x), \hat{f}_n(x) + \hat{c}_{n,\beta}(x)], \tag{3.3}$$

where

$$\hat{c}_{n,\beta}(x) = q_{1-\beta/2}\sqrt{\hat{\sigma}^2(1 - r(x)^T R^{-1} r(x))}, \tag{3.4}$$
$$\text{and} \qquad \hat{\sigma}^2 = Y^T R^{-1} Y / n.$$

Since the underlying function f is fixed and $f \in \mathcal{N}_\Psi(\Omega)$, we can apply (2.20) to derive an upper bound on the prediction error $|f(x) - \hat{f}_n(x)|$ for $x \in \Omega$. By (3.3), at point x, if $f(x) \in \widehat{CI}_{n,\beta}(x)$, we have $|f(x) - \hat{f}_n(x)| \leq \hat{c}_{n,\beta}(x)$. Comparing this inequality with (2.20), and noting that $\hat{c}_{n,\beta}(x) \asymp \hat{\sigma} P_{\Psi,X}(x)$, if the confidence interval is reliable, intuitively, it can be expected that $\hat{\sigma}^2$ should be close to $\|f\|^2_{\mathcal{N}_\Psi(\Omega)}$. However, this is not true. From the identity [64]

$$\|f - \mathcal{I}_{\Psi,X} f\|^2_{\mathcal{N}_\Psi(\Omega)} + \|\mathcal{I}_{\Psi,X} f\|^2_{\mathcal{N}_\Psi(\Omega)} = \|f\|^2_{\mathcal{N}_\Psi(\Omega)}, \tag{3.5}$$

it can be seen that $\hat{\sigma}^2 = \|\mathcal{I}_{\Psi,X} f\|^2_{\mathcal{N}_\Psi(\Omega)}/n \leq \|f\|^2_{\mathcal{N}_\Psi(\Omega)}/n = O(n^{-1})$, which is not close to $\|f\|^2_{\mathcal{N}_\Psi(\Omega)}$ as n becomes larger. This indicates that $\hat{c}_{n,\beta}$ is too small to be used in constructing confidence intervals. Following this intuition, we show that the confidence interval is not reliable, as stated in Theorem 3.1. We need the following condition. Recall that we suppress the dependency of X on n for notational simplification.

Condition 3.1. *Let $X = \{x_1, ..., x_n\}$, thus $n = \text{card}(X)$. The fill distance of X, defined as*

$$h_{X,\Omega} := \sup_{x \in \Omega} \inf_{x_j \in X} \|x - x_j\|_2,$$

satisfies $h_{X,\Omega} \asymp n^{-1/d}$, for all $X \in \mathcal{X}$, where $\mathcal{X}$ is a sampling scheme.

Condition 3.1 can be easily fulfilled. For example, sampling schemes with grid points satisfy Condition 3.1. In fact, any quasi-uniform sampling scheme satisfies Condition 3.1, as shown in the following proposition.

Proposition 3.1 (Proposition 14.1 of [64])**.** *Let $\mathcal{X}$ be a sampling scheme. Suppose there exists a constant $C > 0$ such that for all $X \in \mathcal{X}$, $h_{X,\Omega} \leq Cq_n$, where*

$$q_n := \min_{x_j, x_k \in X, 1 \leq j \neq k \leq n} \|x_j - x_k\|_2/2.$$

Then we have $h_{X,\Omega} \asymp n^{-1/d}$. Such sequence $\mathcal{X}$ is said quasi-uniform.

By the definition of fill distance, it can be seen that

$$\Omega \subset \bigcup_{k=1}^{n} \mathbf{B}(x_k, h_{X,\Omega}),$$

where $\mathbf{B}(x_k, h_{X,\Omega})$ denotes the Euclidean ball centered at x_k with radius $h_{X,\Omega}$. Therefore, a comparison of volumes yields

$$\text{Vol}(\Omega) \leq n\text{Vol}(\mathbf{B}(0, h_{X,\Omega})) = nh_{X,\Omega}^d \frac{\pi^{d/2}}{\Gamma(d/2+1)}.$$

Hence, for any set of measurement locations X with $\text{card}(X) = n$, $h_{X,\Omega} \gtrsim n^{-1/d}$. By (2.20), (2.21) and Lemma C.2 in Appendix C, a set of measurement locations with small fill distance is desired, because we want the measurement locations to be spread in Ω as much as possible. Because quasi-uniform sampling schemes achieve the optimal rate of fill distance, they are widely used in computer experiments. Thus we believe Condition 3.1 is satisfied in many practical situations.

The following proposition provides an upper bound on $|\widehat{CI}_{n,\beta}(x)|$, which implies $\lim_{n\to\infty} \sup_{x\in\Omega} |\widehat{CI}_{n,\beta}(x)| = 0$. Proposition 3.2 is a direct result of Lemma C.2 and the relationship $\hat{\sigma}^2 \leq \|f\|_{\mathcal{N}_\Psi(\Omega)}^2/n$, thus the proof is omitted.

Proposition 3.2. *For any fixed sampling scheme $\mathcal{X}$ satisfying Condition 3.1, we have that $|\widehat{CI}_{n,\beta}(x)| \leq Cn^{-\frac{\nu}{d}}$.*

Under Condition 3.1, we have the following theorem. Note that $f \in \mathcal{N}_\Psi(\Omega)$ and Ψ is a Matérn correlation function defined in (3.1) with $\nu > d/2$ imply $f \in H^\nu(\Omega)$; thus $f \in L_\infty(\Omega)$.

Theorem 3.1. *Suppose $2 < p \leq \infty$, $\beta \in (0,1)$ are fixed, and $\sigma_\epsilon = 0$. For any fixed sampling scheme $\mathcal{X}$ satisfying Condition 3.1, we have that*

$$\sup_{\|f\|_{\mathcal{N}_\Psi(\Omega)} \leq 1} \|(f - \hat{f}_n)/|\widehat{CI}_{n,\beta}|\|_{L_p(\Omega)} \geq Cn^{1/2-1/p} \tag{3.6}$$

holds for all n. For any increasing sequence $\{a_n\}_{n\geq 0}$ satisfying $\lim_{n\to\infty} a_n = \infty$, we have that

$$\sup_{\|f\|_{\mathcal{N}_\Psi(\Omega)} \leq a_n} \|(f - \hat{f}_n)/|\widehat{CI}_{n,\beta}|\|_{L_2(\Omega)} \geq C'a_n \tag{3.7}$$

holds for all n. In (3.6) and (3.7), $\hat{f}_n$ is as in (3.2), $\widehat{CI}_{n,\beta}$ is as in (3.3), C and C' are positive constants depending on p, β, Ω, Ψ and the constants in Condition 3.1, and do not depend on n.

Remark 3.1. *After the earliest version of this work was submitted, a related result has appeared as Theorem 3.2 of [31], which showed that for any function $f \in \mathcal{N}_\Psi(\Omega)$, $\|(f - \hat{f}_n)/|\widehat{CI}_{n,\beta}|\|_{L_\infty(\Omega)} \leq Cn^{1/2}$.*

Theorem 3.1 states that if one uses the estimated variance $\hat{\sigma}^2$ derived by maximum likelihood estimation to construct a pointwise confidence interval, the

confidence interval can be unreliable. The confidence interval is not L_p-weakly-reliable for $2 < p \leq \infty$ as in (3.6), i.e., for any $M > 0$ and sufficient large n, there exists a function f in the unit ball of $\mathcal{N}_\Psi(\Omega)$ such that $\|(f - \hat{f}_n)/|\widehat{CI}_{n,\beta}|\|_{L_p(\Omega)} \geq M$. Furthermore, the confidence interval is not L_2-strongly-reliable as in (3.7), i.e., for $M > 0$ and sufficient large n, there exists a function $f \in N_\Psi(\Omega)$ such that $\|(f - \hat{f}_n)/|\widehat{CI}_{n,\beta}|\|_{L_2(\Omega)} \geq M$. Therefore, it may not be appropriate to quantify the uncertainties by using the confidence interval derived by Gaussian process modeling for a deterministic function lying in the corresponding reproducing kernel Hilbert space if there is no noise.

3.2. Some reliable confidence intervals under Assumption* 2.1

We adopt a reviewer's suggestion and consider two other approaches to imposing $\hat{\sigma}^2$: (1) setting it equal to a constant; and (2) removing the $1/n$ factor from the maximum likelihood estimate. Note that in both cases, Lemma C.2 and Condition 3.1 imply that $|\widehat{CI}_{n,\beta}(x)| \leq Cn^{-\nu/d+1/2}$; thus $\lim_{n \to \infty} \sup_{x \in \Omega} |\widehat{CI}_{n,\beta}(x)| = 0$. If we set $\hat{\sigma}^2$ to be a positive constant, the corresponding confidence interval is L_∞-weakly-reliable (thus is L_p-weakly-reliable for $2 \leq p < \infty$) but not L_∞-strongly-reliable, as stated in the following theorem.

Theorem 3.2. *Suppose $\beta \in (0,1)$ is fixed, and $\sigma_\epsilon = 0$. Let $\Psi = \Psi_M$, where Ψ_M is as in (3.1). Let $\widehat{CI}_{n,\beta}$ be as in (3.3) but with $\hat{\sigma}^2 = 1$. For any fixed sampling scheme $\mathcal{X}$ satisfying Condition 3.1, we have that*

$$\sup_{\|f\|_{\mathcal{N}_\Psi(\Omega)} \leq 1} \|(f - \hat{f}_n)/|\widehat{CI}_{n,\beta}|\|_{L_\infty(\Omega)} \leq 1/(2q_{1-\beta/2}) \tag{3.8}$$

holds for all n. For any increasing sequence $\{a_n\}_{n \geq 0}$ satisfying $\lim_{n \to \infty} a_n = \infty$, we have that

$$\sup_{\|f\|_{\mathcal{N}_\Psi(\Omega)} \leq a_n} \|(f - \hat{f}_n)/|\widehat{CI}_{n,\beta}|\|_{L_\infty(\Omega)} \geq Ca_n \tag{3.9}$$

holds for all n. The constant C is positive and depends on p, β, Ω, Ψ and the constants in Condition 3.1, but does not depend on n.

Next we discuss the second approach, removing the $1/n$ factor from the maximum likelihood estimate. By this approach, the constructed confidence interval $\widehat{CI}_{n,\beta}$ is as in (3.3) with $\hat{\sigma}^2 = Y^T R^{-1} Y = \|\mathcal{I}_{\Psi,X} f\|^2_{\mathcal{N}_\Psi(\Omega)}$. From the identity (3.5), it can be seen that if $\|f - \mathcal{I}_{\Psi,X} f\|^2_{\mathcal{N}_\Psi(\Omega)}$ converges to zero, $Y^T R^{-1} Y$ converges to $\|f\|^2_{\mathcal{N}_\Psi(\Omega)}$. However, in general $\|f - \mathcal{I}_{\Psi,X} f\|^2_{\mathcal{N}_\Psi(\Omega)}$ is not $o(1)$ [18]. Therefore, we need to impose a stronger condition on f such that $\|f - \mathcal{I}_{\Psi,X} f\|^2_{\mathcal{N}_\Psi(\Omega)}$ converges to zero and the corresponding confidence interval is reliable. Define an integral operator $T : L_2(\Omega) \to L_2(\Omega)$ by

$$Tv(x) = \int_\Omega \Psi(x - y)v(y)dy, \quad v \in L_2(\Omega), \quad x \in \Omega,$$

and

$$T(L_2(\Omega)) = \{f | f = Tv, v \in L_2(\Omega)\}.$$

If $f \in T(L_2(\Omega))$, the following lemma states that $\|f - \mathcal{I}_{\Psi,X}f\|_{\mathcal{N}_\Psi(\Omega)} \lesssim \mathcal{P}_{\Psi,X}$. Note that by Lemma C.2, $\mathcal{P}_{\Psi,X} = o(1)$; thus $\|f - \mathcal{I}_{\Psi,X}f\|^2_{\mathcal{N}_\Psi(\Omega)} = o(1)$.

Lemma 3.1. *Suppose $f \in T(L_2(\Omega))$. Then we have*

$$\|f - \mathcal{I}_{\Psi,X}f\|_{\mathcal{N}_\Psi(\Omega)} \leq C\mathcal{P}_{\Psi,X}\|T^{-1}f\|_{L_2(\Omega)},$$

where $\mathcal{P}_{\Psi,X}$ is as in (2.21), and C only depends on Ω.

Lemma 3.1 can be derived directly by the proof of Theorem 11.23 in [64] and the fact $\mathcal{P}_{\Psi,X} \leq 1$; thus the proof is omitted here. We have the following theorem, which states that the confidence interval constructed by the second approach is asymptotically reliable for a fixed function f.

Theorem 3.3. *Suppose $\beta \in (0,1)$ and $f \in T(L_2(\Omega))$ are fixed, and $\sigma_\epsilon = 0$. Let $\Psi = \Psi_M$, where Ψ_M is as in (3.1). Let $\widehat{CI}_{n,\beta}$ be as in (3.3) but with $\hat{\sigma}^2 = Y^T R^{-1} Y$. For any fixed sampling scheme $\mathcal{X}$ satisfying Condition 3.1, there exists $N > 0$ depending on Ψ, Ω, f and the constants in Condition 3.1, such that for all $n \geq N$,*

$$\|(f - \hat{f}_n)/|\widehat{CI}_{n,\beta}|\|_{L_\infty(\Omega)} \leq C,$$

where C is a positive constant only depending on f, Ψ, Ω, and β.

Although Theorem 3.3 does not imply that the confidence interval is L_∞-strongly-reliable because the sample size N depends on f, it can provide a guideline for practitioners to construct confidence intervals for deterministic functions. Whether the confidence interval with $\hat{\sigma}^2 = Y^T R^{-1} Y$ is L_∞-strongly-reliable and the confidence interval with constant $\hat{\sigma}^2$ is L_p-strongly-reliable ($p < \infty$) will be pursued in future works.

4. When the observations are noisy

In this section, we consider the case that the observations are corrupted by noise. We call it *stochastic case*, because multiple evaluations of the function on the same measurement location may have different observations. The observations y_k's are given by (2.1). In the stochastic case, the variance of noise $\sigma_\epsilon^2 > 0$. In this section, we still assume $f \in \mathcal{N}_\Psi(\Omega)$ in (2.1) is a fixed function. Under the misspecified assumption Assumption* 2.1, we use $\hat{f}_n(x)$ defined by

$$\hat{f}_n(x) = r(x)^T(R + \hat{\mu}_n I_n)^{-1} Y \tag{4.1}$$

to predict $f(x)$ on a point $x \in \Omega$, where $r(x)$ and R are as in (2.4), and $Y = (y_1, ..., y_n)^T$. Through this section, we assume that the measurement locations $x_1, ..., x_n$ are drawn uniformly from the input space Ω, and $\hat{\mu}_n \asymp n^\alpha$ with

$\alpha \in \mathbb{R}$. It is obvious that α should be less than one in order to make meaningful predictions. In particular, if $\alpha = 0$, then $\hat{\mu}_n$ is at a constant rate, which is widely used in computer experiments [8, 14]. If replicates on the same measurement location are available, then [3] sets $\hat{\mu}_n$ to be the sample variance of these replicates, which also converges to a constant as the number of replicates on each measurement location goes to infinity. It is well-known that if $\alpha = d/(2\nu+d)$, $\hat{f}_n$ achieves the optimal convergence rate $n^{-\frac{\nu}{2\nu+d}}$ under L_2 metric [53, 58]. In the following theorem, we show that if $\alpha \neq d/(2\nu+d)$, the optimal convergence rate is not achieved. Recall that we use $C, C', C_1, C_2, ...$ and $\eta, \eta_0, \eta_1, ...$ to denote the constants, of which the values can change from line to line, and x_k's are drawn uniformly from Ω.

Theorem 4.1. *Suppose $\hat{\mu}_n \asymp n^\alpha$ with $\alpha < 1$, and the correlation function Ψ satisfies Condition 2.1. Let $\hat{f}_n$ be given by (4.1). Let $X = \{x_1, ..., x_n\}$, where $x_1, ..., x_n$ are uniformly distributed on $\Omega \subset \mathbb{R}^d$. Under the stochastic case ($\sigma_\epsilon > 0$), the following statements are true for all n.*
(i) Suppose $1 > \alpha > \frac{d}{2\nu+d}$. With probability at least $1 - C_1 \exp(-C_2 n^{\eta_1})$,

$$\sup_{f \in \mathcal{N}_\Psi(\Omega), \|f\|_{\mathcal{N}_\Psi(\Omega)} \leq 1} \|f - \hat{f}_n\|_{L_2(\Omega)}^2 \geq C_3 n^{-(1-2\eta)\frac{2\nu}{2\nu+d}}, \tag{4.2}$$

where $\eta = \left(\frac{(\alpha-1)(2\nu+d)}{2\nu} + 1\right)/4 \in (0, 1/4)$. In particular, with probability at least $1 - C_4 \exp(-C_5 n^{\eta_2})$,

$$\sup_{f \in \mathcal{N}_\Psi(\Omega), \|f\|_{\mathcal{N}_\Psi(\Omega)} \leq 1} \mathbb{E}_\epsilon \|f - \hat{f}_n\|_{L_2(\Omega)}^2 \geq C_6 n^{-(1-2\eta)\frac{2\nu}{2\nu+d}}. \tag{4.3}$$

(ii) Suppose $\alpha < \frac{d}{2\nu+d}$. With probability at least $1 - C_7 \exp(-C_8 n^{\eta_3})$,

$$\sup_{f \in \mathcal{N}_\Psi(\Omega), \|f\|_{\mathcal{N}_\Psi(\Omega)} \leq 1} \mathbb{E}_\epsilon \|f - \hat{f}_n\|_{L_2(\Omega)}^2 \geq C_9 n^{\eta_4}, \tag{4.4}$$

where $\eta_4 = \min\left(-\frac{1}{2}, (1-\alpha)\frac{d}{2\nu} - 1\right) > -\frac{2\nu}{2\nu+d}$. In (i) and (ii), the constants C_i's and η_1, η_2, η_3 are positive and depending on Ψ, Ω and σ_ϵ^2 but not depending on n, and the expectation is taken with respect to ϵ_k's. The probability of (4.2) is with respect to X and ϵ_k's, and the probabilities of (4.3) and (4.4) are with respect to X.

Theorem 4.1 provides non-asymptotic lower bounds on the mean squared prediction error under different choices of the regularization parameter value. In particular, it shows that if $\alpha \neq d/(2\nu+d)$, the optimal convergence rate cannot be achieved with high probability, as summarized in the following corollary.

Corollary 4.1. *Suppose $\hat{\mu}_n \asymp n^\alpha$, the correlation function Ψ satisfies Condition 2.1, and $\alpha \in (-\infty, d/(2\nu+d)) \cup (d/(2\nu+d), 1)$. Let $\hat{f}_n$ be given by (4.1). Under the stochastic case ($\sigma_\epsilon > 0$), we have that*

$$\sup_{f \in \mathcal{N}_\Psi(\Omega), \|f\|_{\mathcal{N}_\Psi(\Omega)} \leq 1} \mathbb{E}\left(\|f - \hat{f}_n\|_{L_2(\Omega)}^2\right) \geq C n^\eta$$

holds for all n, where $\eta > -\frac{2\nu}{2\nu+d}$ and $C > 0$ are constants depending on Ψ, Ω, and σ_ϵ^2.

Corollary 4.1 states that with the uniformly distributed measurement locations, the value of α other than $d/(2\nu+d)$ cannot lead to the optimal predictor. This is intuitively true because the regularization parameter $\hat{\mu}_n$ determines the trade-off between the bias and variance. It is well-known in the literature that by choosing $\alpha = d/(2\nu+d)$, we can achieve the best trade-off between the bias and variance. If $\alpha > d/(2\nu+d)$, the bias becomes large, and the variance is small. On the other hand, if $\alpha < d/(2\nu+d)$, the variance is large, and the bias is small. In both cases, we cannot achieve the best trade-off between the bias and variance, and have a slower convergence rate of the prediction error. To the best of our knowledge, it has not been presented in the literature that by choosing the regularization parameter value other than the rate $n^{\frac{d}{2\nu+d}}$, the optimal rate cannot be achieved.

Next, we consider the uncertainty quantification for $\hat{f}_n$. Recall that for $\beta \in (0,1)$, at point $x \in \Omega$, the confidence interval constructed by Gaussian process modeling is given by

$$\widehat{CI}_{n,\beta}(x) = [\hat{f}_n(x) - \tilde{c}_{n,\beta}(x;\hat{\mu}_n), \hat{f}_n(x) + \tilde{c}_{n,\beta}(x;\hat{\mu}_n)],$$

where

$$\tilde{c}_{n,\beta}(x;\hat{\mu}_n) = q_{1-\beta/2}\sqrt{\hat{\sigma}^2(1 - r(x)^T(R + \hat{\mu}_n I_n)^{-1}r(x))}, \qquad (4.5)$$
$$\text{and} \qquad \hat{\sigma}^2 = Y^T(R + \hat{\mu}_n I_n)^{-1}Y/n.$$

The following proposition states that $|\widehat{CI}_{n,\beta}(x)| = 2\tilde{c}_{n,\beta}(x;\hat{\mu}_n)$ converges to zero with $0 \leq \alpha < 1$.

Proposition 4.1. *Under the conditions of Theorem 4.1 and $0 \leq \alpha < 1$, it holds that with probability at least $1 - C_1 \exp(-C_2 n^\eta)$, $c_{n,\beta}(x;\hat{\mu}_n)^2 \leq C_3 n^{-\frac{2\nu+(\alpha-1)d}{2\nu}}$ for all n, where C_1, C_2, C_3 and η are positive constants depending on Ψ, Ω, and σ_ϵ^2.*

Proposition 4.1 is a direct result of Lemmas F.6 and F.8 in Appendix F, and implies $\lim_{n\to\infty} \sup_{x\in\Omega} |\widehat{CI}_{n,\beta}(x)| = 0$ if $0 \leq \alpha < 1$. Intuitively, when α is large, the bias is large. Therefore, the confidence interval, which is a reflection of the variance, is not wide enough to capture the bias. As a consequence, the confidence interval $\widehat{CI}_{n,\beta}$ is not reliable. On the other hand, a smaller regularization parameter value lets the variance dominate. Therefore, the variance dominates the confidence interval; thus the confidence interval should be reliable. The results related to the reliability of the confidence interval $\widehat{CI}_{n,\beta}$ are presented in the following theorem. In Theorem 4.2, recall that $x_1, ..., x_n$ are drawn uniformly from Ω, and $|\widehat{CI}_{n,\beta}(x)| = 2\tilde{c}_{n,\beta}(x;\hat{\mu}_n)$.

Theorem 4.2. *Suppose $\hat{\mu}_n \asymp n^\alpha$ with $\alpha < 1$, and the correlation function Ψ satisfies Condition 2.1. Fix $\beta \in (0,1)$. Under the stochastic case ($\sigma_\epsilon > 0$), the following statements are true for all n.*

(i) Suppose $\alpha > \frac{d}{2\nu+d}$. With probability at least $1 - C_1 \exp(-C_2 n^{\eta_1})$,

$$\sup_{f \in \mathcal{N}_\Psi(\Omega), \|f\|_{\mathcal{N}_\Psi(\Omega)} \leq 1} \mathbb{E}_\epsilon \|(f - \hat{f}_n)/\tilde{c}_{n,\beta}(\cdot; \hat{\mu}_n)\|_{L_2(\Omega)}^2 \geq C n^{\eta_2}. \qquad (4.6)$$

(ii) Suppose $0 \leq \alpha < \frac{d}{2\nu+d}$. With probability at least $1 - C_3 \exp(-C_4 n^{\eta_3})$,

$$\sup_{f \in \mathcal{N}_\Psi(\Omega), \|f\|_{\mathcal{N}_\Psi(\Omega)} \leq \sqrt{\log n}} \mathbb{E}_\epsilon \|f - \hat{f}_n\|_{L_2(\Omega)}^2 \leq C' \mathbb{E}_\epsilon \|\tilde{c}_{n,\beta}(\cdot; \hat{\mu}_n)\|_{L_2(\Omega)}^2. \qquad (4.7)$$

Suppose $\alpha = \frac{d}{2\nu+d}$. Then for any increasing positive sequence $\{a_n\}_{n \geq 0}$ satisfying $\lim_{n \to \infty} a_n = \infty$,

$$\sup_{f \in \mathcal{N}_\Psi(\Omega), \|f\|_{\mathcal{N}_\Psi(\Omega)}^2 \leq a_n} \mathbb{E}_\epsilon \|(f - \hat{f}_n)/\tilde{c}_{n,\beta}(\cdot; \hat{\mu}_n)\|_{L_2(\Omega)}^2 \geq C'' a_n. \qquad (4.8)$$

(iii) Suppose $\alpha < 0$. With probability at least $1 - C_5 \exp(-C_6 n^{\eta_5})$,

$$\sup_{f \in \mathcal{N}_\Psi(\Omega), \|f\|_{\mathcal{N}_\Psi(\Omega)} \leq \log n} \mathbb{E}_\epsilon \|(f - \hat{f}_n)/\tilde{c}_{n,\beta}(\cdot; \hat{\mu}_n)\|_{L_2(\Omega)}^2 \leq C'''. \qquad (4.9)$$

In (4.6)-(4.9), $\hat{f}_n$ is as in (4.1) and $\tilde{c}_{n,\beta}(\cdot; \hat{\mu}_n)$ is as in (4.5). In all statements, the expectation is taken with respect to $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$, and the constants C', C'', C''', C_i's and η_j's are positive, and depend on Ψ, Ω, β and σ_ϵ^2 but not depending on n. The probabilities are with respect to X.

As direct results of Theorem 4.2, we have the following corollary, which states the results related to the L_2-reliability of the confidence interval $\widetilde{CI}_{n,\beta}(x)$.

Corollary 4.2. *Suppose $\hat{\mu}_n \asymp n^\alpha$ with $\alpha < 1$, and the correlation function Ψ satisfies Condition 2.1. Fix $\beta \in (0, 1)$. Under the stochastic case $(\sigma_\epsilon > 0)$, the following statements are true for all n.*
(i) Suppose $\alpha > \frac{d}{2\nu+d}$. We have

$$\sup_{f \in \mathcal{N}_\Psi(\Omega), \|f\|_{\mathcal{N}_\Psi(\Omega)} \leq 1} \mathbb{E} \left(\|(f - \hat{f}_n)/\tilde{c}_{n,\beta}(\cdot; \hat{\mu}_n)\|_{L_2(\Omega)}^2 \right) \geq C n^{\eta_1}.$$

(ii) Suppose $0 \leq \alpha < \frac{d}{2\nu+d}$. We have

$$\sup_{f \in \mathcal{N}_\Psi(\Omega), \|f\|_{\mathcal{N}_\Psi(\Omega)} \leq \sqrt{\log n}} \mathbb{E} \|f - \hat{f}_n\|_{L_2(\Omega)}^2 \leq C' \mathbb{E} \|\tilde{c}_{n,\beta}(\cdot; \hat{\mu}_n)\|_{L_2(\Omega)}^2.$$

Suppose $\alpha = \frac{d}{2\nu+d}$. Then for any increasing positive sequence $\{a_n\}_{n \geq 0}$ satisfying $\lim_{n \to \infty} a_n = \infty$,

$$\sup_{f \in \mathcal{N}_\Psi(\Omega), \|f\|_{\mathcal{N}_\Psi(\Omega)}^2 \leq a_n} \mathbb{E} \left(\|(f - \hat{f}_n)/\tilde{c}_{n,\beta}(\cdot; \hat{\mu}_n)\|_{L_2(\Omega)}^2 \right) \geq C'' a_n.$$

(iii) Suppose $\alpha < 0$. We have

$$\sup_{f \in \mathcal{N}_\Psi(\Omega), \|f\|_{\mathcal{N}_\Psi(\Omega)} \leq \log n} \mathbb{E}\left(\|(f - \hat{f}_n)/\tilde{c}_{n,\beta}(\cdot; \hat{\mu}_n)\|_{L_2(\Omega)}^2 \right) \leq C'''.$$

In all statements, the constants C', C'', C''', C_i's and η_1 are positive, and only depend on Ψ, Ω, β and σ_ϵ^2 but not depending on n.

Note that Theorem 4.2 and Corollary 4.2 do not make any theoretical assertion about L_2-weak-reliability under the case $0 \leq \alpha \leq \frac{d}{2\nu+d}$, and L_2-strong-reliability under the case $0 \leq \alpha < \frac{d}{2\nu+d}$. As (4.7) indicates, we conjecture that the constructed confidence interval under the stochastic case is L_2-weakly-reliable if $\alpha = \frac{d}{2\nu+d}$, and is L_2-strongly-reliable if $0 \leq \alpha < \frac{d}{2\nu+d}$. We also note that *(iii)* in Corollary 4.2 does not imply the L_2-strong-reliability since we do not confirm the width of the confidence interval converges to zero when $\alpha < 0$.

Combining Corollary 4.1 and Corollary 4.2 suggests that if one applies the prediction and uncertainty quantification procedure from Gaussian process modeling to a deterministic function with noisy observations, the optimality of the predictor and the L_2-strong-reliability of the confidence interval cannot be achieved at the same time.

As a by-product of Theorem 4.1, we show that if the observations are not corrupted by noise, and a regularization parameter is used as a counteract of the potential numerical instability [44], then with uniformly distributed measurement locations, the prediction error can be controlled.

Theorem 4.3. *Suppose $\sigma_\epsilon = 0$, $\hat{\mu}_n \asymp n^\alpha$ with $\alpha < 1$, and the correlation function Ψ satisfies Condition 2.1. Then for all n, with probability at least $1 - C_1 \exp(-C_2 n^{\eta_1})$, we have*

$$\|f - \hat{f}_n\|_{L_\infty(\Omega)}^2 \leq C_3 \max\left(n^{(\alpha-1)(1-\frac{d}{2\nu})}, n^{-(1-\frac{d}{2\nu})} \right),$$

where C_1, C_2, C_3 and η_1 are positive constants depending on Ψ, Ω, and f.

Theorem 4.3 is a direct result of Lemma F.8 in Appendix F. Theorem 4.3 states that $\hat{f}_n$ defined in (4.1) converges to the true underlying function f, even if the observations are noiseless. Note that in this theorem, it is allowed that the regularization parameter value increases as the sample size increases.

5. A numerical example

We present a numerical example to illustrate the results in Section 3, where we show that the confidence interval is not reliable in the deterministic case.

Recall that Theorem 3.1 states that there exists a function with smoothness (at least) ν such that the confidence interval is not L_p-reliable. However, such a function is generally not straightforward to find. In this section, we numerically illustrate that there exists a function such that the confidence interval is not L_p-reliable.

Consider the following function [23],

$$f(x) = \sin(4x) - 0.02t_1(x; 1.57, 0.05) \qquad (5.1)$$

for $x \in [0, 1]$, where $t_1(\cdot; J_\mu, J_\sigma)$ is a Cauchy density with mean J_μ and spread J_σ. In [23], it is shown that using a Gaussian correlation function yields a poor coverage rate. In this section we use a Matérn correlation function. The numerical results suggest that with a Matérn correlation function, the Gaussian process model does not provide an L_p-reliable confidence interval. As in Section 3, we compute

$$\|(f - \hat{f}_n)/|\widehat{CI}_{n,\beta}|\|_{L_p(\Omega)},$$

where f is as in (5.1), $\hat{f}_n$ is as in (3.2), $\widehat{CI}_{n,\beta}$ is as in (3.3), and $p = 4$. We use a Matérn correlation function as in (3.1) with $\nu = 3.5$. It can be checked that $f \in \mathcal{N}_{\Psi_M}([0, 1])$. We consider the 95% confidence interval constructed by the Gaussian process modeling. Thus, $\beta = 0.05$. We set the number of measurement locations as $n = 20k$, $k = 2, ..., 20$, and the measurement locations are grid points. We use

$$\mathcal{E} = \frac{1}{500} \sum_{j=1}^{500} \frac{|f(x_j) - \hat{f}_n(x_j)|^4}{|\widehat{CI}_{n,\beta}(x_j)|^4} \qquad (5.2)$$

to approximate $\|(f - \hat{f}_n)/|\widehat{CI}_{n,\beta}|\|_{L_4(\Omega)}^4$, where x_j's are the first 500 points in the Halton sequence [42]. This should give a good approximation since the points are dense enough. We add a jitter 10^{-8} to stabilize the computation of the matrix inverse in (3.2) and (3.3). The results are shown in Panel 1 of Figure 1.

We use the following approach to numerically show the rate of divergence of the ratio of the prediction error divided by the width of the confidence interval. By Theorem 3.1, we have

$$\sup_{\|g\|_{\mathcal{N}_\Psi(\Omega)} \leq 1} \|(g - \hat{g}_n)/|\widehat{CI}_{n,\beta}|\|_{L_4(\Omega)}^4 \geq Cn. \qquad (5.3)$$

Taking logarithm on both sides of (5.3), we have

$$\log \left(\sup_{\|g\|_{\mathcal{N}_\Psi(\Omega)} \leq 1} \|(g - \hat{g}_n)/|\widehat{CI}_{n,\beta}|\|_{L_4(\Omega)}^4 \right) \geq \log n + \log C. \qquad (5.4)$$

We regress

$$\log \left(\frac{1}{500} \sum_{j=1}^{500} \frac{|f(x_j) - \hat{f}_n(x_j)|^4}{|\widehat{CI}_{n,\beta}(x_j)|^4} \right)$$

on $\log n$. If the regression coefficient is larger than one, it indicates that the results in Theorem 3.1 hold. The results are shown in Panel 2 of Figure 1.
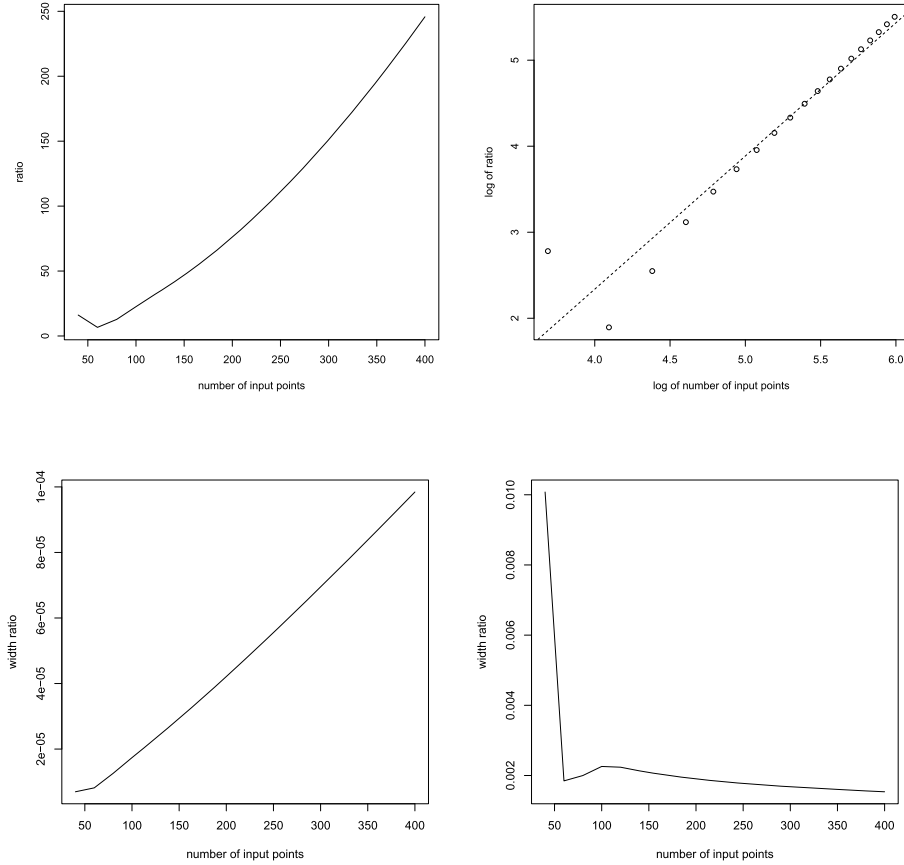
FIG 1. **Panel 1:** *Plot of $\mathcal{E}$ in* (5.2) *with* $\hat{\sigma}^2 = Y^T R^{-1} Y / n$. **Panel 2:** *The regression line of* $\log \mathcal{E}$ *on* $\log n$. *The dashed line shows the regression line. Each point denotes* $\log \mathcal{E}$ *for each number of measurement locations. The regression coefficient is 1.548.* **Panel 3:** *Plot of* $\log \mathcal{E}$ *with* $\hat{\sigma}^2 = C$. **Panel 4:** *Plot of* $\log \mathcal{E}$ *with* $\hat{\sigma}^2 = Y^T R^{-1} Y$.

Next, we consider two approaches for constructing confidence intervals described in Section 3.2. Similarly, we compute

$$\frac{1}{500} \sum_{j=1}^{500} \frac{|f(x_j) - \hat{f}_n(x_j)|^4}{|\widehat{CI}_{n,\beta}(x_j)|^4}, \tag{5.5}$$

with $\hat{\sigma}^2 = C$ and $\hat{\sigma}^2 = Y^T R^{-1} Y$ in $\widehat{CI}_{n,\beta}$, respectively, where C is a constant satisfying $\|f\|^2_{\mathcal{N}_\Psi([0,1])} \leq C$. In this example, we set $C = 2\tilde{Y}^T \tilde{R}^{-1} \tilde{Y}$, where $\tilde{Y} = (f(\tilde{x}_1), ..., f(\tilde{x}_{1000}))^T$, $\tilde{R} = (\Psi(\tilde{x}_j - \tilde{x}_k))_{jk}$, and $\tilde{x}_j$'s are 1000 grid points. Since $\tilde{Y}^T \tilde{R}^{-1} \tilde{Y}$ should be a good approximation of $\|f\|^2_{\mathcal{N}_\Psi([0,1])}$, C should be a valid upper bound of $\|f\|^2_{\mathcal{N}_\Psi([0,1])}$. The results are shown in Panels 3 and 4 of Figure 1.

It can be seen from Panel 1 of Figure 1 that $\mathcal{E}$ in (5.2) increases as the number of measurement locations increases. From Panel 2 of Figure 1, we can see that the regression line fits the data relatively well. The regression coefficient is 1.548, which is larger than one. This gives us an empirical confirmation that our results in Theorem 3.1 are valid, and there exists a function such that the confidence interval is not L_p-reliable. As indicated by Figure 1, we believe the results in Theorem 3.1 can be improved. In Panel 3, although the ratio increases, the largest value of the ratio is only about 10^{-4}. Panels 3 and 4 indicate that the two approaches in Section 3.2 can provide reliable confidence intervals. It can be seen that the confidence interval derived by setting $\hat{\sigma}^2 = C$ may be conservative (the ratio is very small). Therefore, other uncertainty quantification methods may be considered if the underlying function is known to be in some reproducing kernel Hilbert space.

6. Conclusions and discussion

In this work, we consider the prediction and uncertainty quantification of the Gaussian process model applied to a fixed function in the corresponding reproducing kernel Hilbert space from a frequentist perspective. The model is misspecified under Assumption* 2.1. We consider two cases, the deterministic case, in which the observations are noiseless, and the stochastic case, where the observations are corrupted by noise. In both cases, we assume that the variance is estimated by maximum likelihood estimation, according to Assumption* 2.1. In the deterministic case, we show that the constructed confidence interval in the Gaussian process model is not L_p-weakly-reliable for $p > 2$, and is not L_2-strongly-reliable. We also present two reliable confidence intervals under some scenarios. In the stochastic case, the regularization parameter value is assumed to be at a certain rate. We show that the predictor derived by the Gaussian process modeling is not optimal and/or the constructed confidence interval is not L_2-strongly-reliable. These results indicate that the optimal predictor and L_2-strong-reliability cannot be achieved at the same time if the Gaussian process model is misspecified. As by-products, we obtain several lower bounds on the mean squared prediction error under different choices of the regularization parameter value.

In the Gaussian process model, it is often assumed that there are some unknown parameters, and maximum likelihood estimation or Bayesian methods, are used to estimate these parameters, even if the Gaussian process model is misspecified. Prediction performance of the Gaussian process model with maximum likelihood estimation under misspecification has been studied by [49, 7]. In [49], they show for periodic functions, under some situations, the misspecified Gaussian process model with maximum likelihood estimation can still work well in terms of prediction. If the underlying truth is indeed a Gaussian process, using a misspecified correlation function and maximum likelihood estimation may not have desired prediction performance, as suggested in [7].

In addition to prediction, uncertainty quantification is another important problem in computer experiments. Because the Gaussian process model has a

probabilistic structure, models based on Gaussian process modeling are usually validated via confidence intervals. In order to test the performance of these models, typically, several simulations are conducted. In some literature, the test function is selected to be a deterministic function with a closed form. In these cases, the Gaussian process model is usually misspecified, i.e., the function may not be a sample path of the corresponding Gaussian process. However, an imposed pointwise confidence interval is still constructed and used to quantify the uncertainty. It has been observed that Gaussian process models often have poor coverage of their confidence intervals [23, 30, 69]. There is no theoretical result explaining this phenomenon from a frequentist perspective to the best of our knowledge. Our results provide some insights on explaining the poor coverage of the confidence interval, and a better understanding of the model misspecification in Gaussian process modeling.

Several statistical inference methods have been studied if the confidence interval is not derived directly from the Gaussian process modeling. Most of them are from a Bayesian perspective. For example, in [48, 74], credible intervals are constructed and analyzed for Gaussian process models (or particularly Brownian motion). Additionally, [70, 71] derive finite sample bounds on frequentist coverage errors of Bayesian credible intervals for Gaussian process models. Therefore, one can also use other inference methods besides the confidence interval in the Gaussian process modeling.

Several problems are not considered in this work. First, we only consider uniformly distributed measurement locations in the stochastic case, where fixed designs are not considered. Second, we only consider the case that the regularization parameter value is predetermined with a certain rate. We could not confirm similar results if we use maximum likelihood estimation to estimate the regularization parameter value, or select parameter values using other criteria as in [33]. Also, we do not consider using maximum likelihood estimation to estimate the parameters of the correlation function Ψ. Therefore, a thorough investigation of applying maximum likelihood estimation under misspecification is needed. Third, as discussed in Section 2.3, Definition 2.1 is a necessary condition. One possible way to improve this definition is to restrict the L_p-norm of the ratio such that it is bounded away from zero.

Appendix A: Notation

We use $\langle \cdot, \cdot \rangle_n$ to denote the empirical inner product, which is defined by

$$\langle f, g \rangle_n = \frac{1}{n} \sum_{k=1}^{n} f(x_k) g(x_k)$$

for two functions f and g, and let $\|g\|_n^2 = \langle g, g \rangle_n$ be the empirical norm of function g. In particular, let

$$\langle \epsilon, f \rangle_n = \frac{1}{n} \sum_{k=1}^{n} \epsilon_k f(x_k)$$

for a function f, where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$. Let $a \vee b = \max(a, b)$ for two real numbers a, b. We use $H(\cdot, \mathcal{F}, \|\cdot\|)$ and $H_B(\cdot, \mathcal{F}, \|\cdot\|)$ to denote the entropy number and the bracket entropy number of class $\mathcal{F}$ with the (empirical) norm $\|\cdot\|$, respectively. Without loss of generality, we assume $\mathrm{Vol}(\Omega) = 1$ in the Appendix.

Appendix B: Proof of Proposition 2.1

By (2.7) and (2.5), it can be seen that $|CI_{n,\beta}(x)| = 2q_{1-\beta/2}\sqrt{\mathrm{Var}[Z(x)|\mathcal{Z}]}$ for a point x. By Fubini's theorem,

$$
\begin{aligned}
\mathbb{E}\|(Z - I_X^{(1)}Z)/|CI_{n,\beta}|\|_{L_p(\Omega)}^p &= \int_{x \in \Omega} \frac{\mathbb{E}(Z(x) - I_X^{(1)}Z(x))^p}{C_1(\sqrt{\mathrm{Var}[Z(x)|\mathcal{Z}]})^p} dx \\
&= \int_{x \in \Omega} \frac{2^{p/2}\Gamma(\frac{p+1}{2})}{\sqrt{\pi}} \frac{(\sqrt{\mathrm{Var}[Z(x)|\mathcal{Z}]})^p}{C_1(\sqrt{\mathrm{Var}[Z(x)|\mathcal{Z}]})^p} dx \\
&= C_2,
\end{aligned}
$$

where the second equality can be derived by the calculation of pth moment of normal distribution. See [60].

Appendix C: Proof of Theorem 3.1

The proof of Theorem 3.1 relies on approximation numbers. The nth approximation number of the embedding $id : H^\nu(\Omega) \to L_p(\Omega)$, denoted by b_n, is defined by

$$
b_n = \inf\{\|id - L\|, L \in \mathcal{L}(H^\nu(\Omega), L_p(\Omega)), \mathrm{rank}(L) < n\}, \tag{C.1}
$$

where $\mathcal{L}(H^\nu(\Omega), L_p(\Omega))$ is the family of all bounded linear mappings $H^\nu(\Omega) \to L_p(\Omega)$, $\|\cdot\|$ is the operator norm, and $\mathrm{rank}(L)$ is the dimension of the range of L. In (C.1), we define $\mathrm{rank}(L) = 0$ if $L(f) = 0$ for all $f \in H^\nu(\Omega)$. Therefore, it can be seen that $\|id\| = b_1 \geq b_2 \geq \ldots > 0$. Lemma C.1 states a property of approximation numbers [18].

Lemma C.1. *Suppose $p \geq 2$. The approximation number b_n defined in (C.1) satisfies that for all $n \in \mathbb{N}$,*

$$
c_1 n^{-\frac{\nu}{d}+\frac{1}{2}-\frac{1}{p}} \leq b_n \leq c_2 n^{-\frac{\nu}{d}+\frac{1}{2}-\frac{1}{p}}, \tag{C.2}
$$

where c_1 and c_2 are two positive constants depending on Ω, ν and p.

Lemma C.2 is a direct result of Theorem 5.14 of [66], which provides an upper bound on $\mathcal{P}_{\Psi_M, X}$ defined in (2.21).

Lemma C.2. *Let Ω be compact and convex with a positive Lebesgue measure; Ψ_M be a Matérn correlation function given by (3.1). Suppose Condition 3.1 holds for a sampling scheme $\mathcal{X}$. Then there exist constants c, c_1 depending only on Ω, $\mathcal{X}$, and ν in (3.1), such that $\mathcal{P}_{\Psi_M, X} \leq cn^{-\frac{\nu}{d}+\frac{1}{2}}$ provided that $n^{-\frac{\nu}{d}+\frac{1}{2}} \leq c_1$ and $X \in \mathcal{X}$.*

Lemma C.3 is a direct result of Corollary 10.13 in [64] and the extension theorem [15]. Lemma C.3 states that the reproducing kernel Hilbert space $\mathcal{N}_\Psi(\Omega)$ coincides with the Sobolev space with smoothness ν $H^\nu(\Omega)$, for correlation functions satisfying Condition 2.1.

Lemma C.3. *Suppose Condition 2.1 is satisfied. We have the following.*

1. *The reproducing kernel Hilbert space $\mathcal{N}_\Psi(\mathbb{R}^d)$ coincides with the Sobolev space with smoothness ν $H^\nu(\mathbb{R}^d)$, and the norms $\|\cdot\|_{\mathcal{N}_\Psi(\mathbb{R}^d)}$ and $\|\cdot\|_{H^\nu(\mathbb{R}^d)}$ are equivalent.*
2. *Suppose Ω is compact and convex. Then the reproducing kernel Hilbert space $\mathcal{N}_\Psi(\Omega)$ coincides with the Sobolev space with smoothness ν $H^\nu(\Omega)$, and the norms $\|\cdot\|_{\mathcal{N}_\Psi(\Omega)}$ and $\|\cdot\|_{H^\nu(\Omega)}$ are equivalent.*

Now we are ready to prove Theorem 3.1.

By Lemma C.1, there exists a function $\phi_n \in H^\nu(\Omega)$ satisfying $\|\phi_n\|_{H^\nu(\Omega)} = 1$ such that

$$c_1 n^{-\frac{2\nu}{d}+1-\frac{2}{p}} \leq \|\phi_n - \mathcal{I}_{\Psi,X}\phi_n\|^2_{L_p(\Omega)},$$

since $\mathcal{I}_{\Psi,X}$ is a rank n linear operator. By Lemma C.2, (3.5), and Lemma C.3, we have for sufficiently large N such that $N^{-\frac{\nu}{d}+\frac{1}{2}} \leq c_1$ and for all $n \geq N$,

$$\begin{aligned}
\hat{c}_{n,\beta}(x)^2 &= \frac{C_1}{n} Y^T R^{-1} Y (1 - r(x)^T R^{-1} r(x)) \leq \frac{C_1}{n} Y^T R^{-1} Y \mathcal{P}^2_{\Psi_M,X} \\
&\leq \frac{C_2}{n} \|\phi_n\|^2_{\mathcal{N}_\Psi(\Omega)} n^{-\frac{2\nu}{d}+1} \leq \frac{C_3}{n} \|\phi_n\|^2_{H^\nu(\Omega)} n^{-\frac{2\nu}{d}+1} = C_3 n^{-\frac{2\nu}{d}}
\end{aligned}$$

for any $x \in \Omega$, where $Y = (\phi_n(x_1), ..., \phi_n(x_n))^T$. Let f in (3.6) be equal to ϕ_n. Therefore, we have

$$\|(f - \hat{f}_n)/\hat{c}_{n,\beta}\|^2_{L_p(\Omega)} \geq C_4 \|f - \hat{f}_n\|^2_{L_p(\Omega)} n^{\frac{2\nu}{d}} \geq C_5 n^{1-\frac{2}{p}},$$

which finishes the proof of (3.6).

The case $p = 2$ can be proved similarly. The only difference is that we let $f = a_n \phi_n$ such that $\|a_n \phi_n\|_{H^\nu(\Omega)} = a_n$.

Appendix D: Proof of Theorem 3.2

We first show that (3.8) holds. Plugging $\hat{c}_{n,\beta}(x) = q_{1-\beta/2} P_{\Psi,X}(x)$, it suffices to show

$$\left| \frac{f(x) - \hat{f}_n(x)}{P_{\Psi,X}(x)} \right| \leq 1$$

holds for all f with $\|f\|_{\mathcal{N}_\Psi(\Omega)} \leq 1$ and $x \in \Omega$. This is a direct result of (2.20).

The second inequality (3.9) can be shown by a similar approach as in the proof of Theorem 3.1. By Lemma C.1, there exists a function $\phi_n \in H^\nu(\Omega)$ satisfying $\|\phi_n\|_{H^\nu(\Omega)} = 1$ such that

$$c_1 n^{-\frac{2\nu}{d}+1} \leq \|\phi_n - \mathcal{I}_{\Psi,X}\phi_n\|^2_{L_\infty(\Omega)}.$$

By Lemma C.2, we have

$$\hat{c}_{n,\beta}(x)^2 \leq C_1\|\phi_n\|^2_{\mathcal{N}_\Psi(\Omega)} n^{-\frac{2\nu}{d}+1} \leq C_2\|\phi_n\|^2_{H^\nu(\Omega)} n^{-\frac{2\nu}{d}+1} = C_2 n^{-\frac{2\nu}{d}+1}$$

for any $x \in \Omega$ and sufficiently large n such that $n^{-\frac{\nu}{d}+\frac{1}{2}} \leq c_1$. Letting $f = a_n\phi_n$, we have

$$\|(f - \hat{f}_n)/\hat{c}_{n,\beta}\|^2_{L_\infty(\Omega)} \geq C_3 a_n,$$

which finishes the proof.

Appendix E: Proof of Theorem 3.3

By plugging $\hat{c}_{n,\beta}(x) = q_{1-\beta/2}\sqrt{Y^T R^{-1} Y}P_{\Psi,X}(x)$, it suffices to show that there exists $N > 0$ such that for all $n > N$,

$$\left| \frac{f(x) - \hat{f}_n(x)}{\sqrt{Y^T R^{-1} Y}P_{\Psi,X}(x)} \right| \leq C.$$

By (2.20) and Lemma 3.1, for sufficiently large n, it can be seen that

$$\left| \frac{f(x) - \hat{f}_n(x)}{\sqrt{Y^T R^{-1} Y}P_{\Psi,X}(x)} \right|^2 \leq \frac{\|f\|^2_{\mathcal{N}_\Psi(\Omega)}}{Y^T R^{-1} Y} = \frac{\|f\|^2_{\mathcal{N}_\Psi(\Omega)}}{\|f\|^2_{\mathcal{N}_\Psi(\Omega)} - \|f - \mathcal{I}_{\Psi,X}f\|^2_{\mathcal{N}_\Psi(\Omega)}}$$

$$\leq \frac{\|f\|^2_{\mathcal{N}_\Psi(\Omega)}}{\|f\|^2_{\mathcal{N}_\Psi(\Omega)} - C^2 P^2_{\Psi,X}\|T^{-1}f\|^2_{L_2(\Omega)}}$$

$$\leq \frac{\|f\|^2_{\mathcal{N}_\Psi(\Omega)}}{\|f\|^2_{\mathcal{N}_\Psi(\Omega)} - C_1 n^{-\frac{2\nu}{d}+1}\|T^{-1}f\|^2_{L_2(\Omega)}} \leq \frac{1}{2},$$

where the first inequality is by (3.5), and the last inequality follows from $n^{-\frac{2\nu}{d}+1}$ converges to 0. Then we finish the proof.

Appendix F: Proof of Theorem 4.1

Recall that in the stochastic case, we assume $x_1, ..., x_n$ are drawn uniformly from Ω. Before we show the proof of Theorem 4.1, we first present some lemmas used in this section. Note that the proof of Lemma F.1 is based on Lemma 8.4 of [58]; thus it is omitted here. Lemmas F.2 and F.3 are Theorem 10.2 of [58] and Theorem 2.1 of [59], respectively.

Lemma F.1. *Suppose $\epsilon_1, ..., \epsilon_n$ are independent and identically normally distributed variables. Then for all $t > C$, with probability at least $1 - C_1 \exp(-C_2 t^2)$,*

$$\sup_{g \in \mathcal{N}_\Psi(\Omega)} \frac{|\langle \epsilon, g \rangle_n|}{\|g\|_n^{1 - \frac{d}{2\nu}} \|g\|_{\mathcal{N}_\Psi(\Omega)}^{\frac{d}{2\nu}}} \leq t n^{-\frac{1}{2}}.$$

Lemma F.2. *Suppose $f \in H^\nu(\Omega)$ and $\hat{\mu}_n^{-1} = O_P(n^{-\frac{d}{2\nu+d}})$. Then we have*

$$\|f - \hat{f}_n\|_n = O_P(\hat{\mu}_n^{\frac{1}{2}} n^{-\frac{1}{2}} \vee n^{\frac{d-2\nu}{4\nu}} \hat{\mu}_n^{-\frac{d}{4\nu}}),$$

$$\|\hat{f}_n\|_{\mathcal{N}_\Psi(\Omega)} = O_P(1 \vee n^{\frac{d}{4\nu}} \hat{\mu}_n^{-\frac{2\nu+d}{4\nu}}),$$

where $\hat{f}_n$ is defined as in (4.1).

Lemma F.3. *Let $\mathcal{R} := \sup_{f \in \mathcal{F}} \|f\|_{L_2(\Omega)}, K := \sup_{f \in \mathcal{F}} \|f\|_{L_\infty(\Omega)}$, where $\mathcal{F}$ is a function class. Then for all $t > 0$, with probability at least $1 - \exp(-t)$,*

$$\sup_{f \in \mathcal{F}} \left| \|f\|_n^2 - \|f\|_{L_2(\Omega)}^2 \right| \leq C_1 \left(\frac{2\mathcal{R}J_\infty(K, \mathcal{F}) + \mathcal{R}K\sqrt{t}}{\sqrt{n}} + \frac{4J_\infty^2(K, \mathcal{F}) + K^2 t}{n} \right),$$

where C_1 is a constant, and

$$J_\infty^2(z, \mathcal{F}) = C_2^2 \inf_{\delta > 0} \mathbb{E} \left[z \int_\delta^1 \sqrt{H(uz/2, \mathcal{F}, \| \cdot \|_{L_\infty(\Omega)})} du + \sqrt{n} z \delta \right]^2, \quad \text{(F.1)}$$

with C_2 another constant.

The following lemma is a Bernstein-type inequality for a single function g. See, for example, [38].

Lemma F.4. *Suppose X_i, $i = 1, \ldots, n$ are uniformly distributed on Ω. Let $Z_i = (\|g\|_{L_2(\Omega)}^2 / Vol(\Omega) - g(X_i)^2)/\|g\|_{\mathcal{N}_\Psi(\Omega)}^2$ for a function $g \in \mathcal{N}_\Psi(\Omega)$. Suppose $|Z_i| \leq b$ for some constant $b > 0$. For all $t > 0$, we have*

$$P\left(\frac{1}{n} \sum_{i=1}^n Z_i \geq t \right) \leq \exp \left[-\frac{nt^2/2}{E(Z_1^2) + bt/3} \right],$$

which is the same as

$$P\left(\frac{\|g\|_{L_2(\Omega)}^2 / Vol(\Omega)}{\|g\|_{\mathcal{N}_\Psi(\Omega)}^2} - \frac{\|g\|_n^2}{\|g\|_{\mathcal{N}_\Psi(\Omega)}^2} \geq t \right) \leq \exp \left[-\frac{nt^2/2}{E(Z_1^2) + bt/3} \right].$$

Lemma F.5. *Assume for class $\mathcal{G}$, $\sup_{g \in \mathcal{G}} \|g\|_{L_\infty(\Omega)} \leq K < 1$, and the bracket entropy $H_B(\delta_n/Vol(\Omega), \mathcal{G}, \| \cdot \|_{L_\infty(\Omega)}) \leq \frac{n\delta_n^2}{1200 \, Vol(\Omega) K^2}$, and $n\delta_n^2 \to \infty$, where $Vol(\Omega)$ denotes the volume of Ω and $0 < \delta_n < 1$. Then we have*

$$P\left(\inf_{\|g\|_{L_2(\Omega)} \geq 2\delta_n, g \in \mathcal{G}} \frac{\|g\|_n^2}{\|g\|_{L_2(\Omega)}^2} < \eta_1 \right) \leq C_1 \exp(-C_2 n\delta_n^2/K^2),$$

and

$$P\left(\sup_{\|g\|_{L_2(\Omega)} \geq 2\delta_n, g \in \mathcal{G}} \frac{\|g\|_n^2}{\|g\|_{L_2(\Omega)}^2} > \eta_2\right) \leq C_3 \exp(-C_4 n \delta_n^2 / K^2),$$

for some constants $\eta_1, \eta_2 > 0$ and C_i's only depending on Ω.

Lemma F.6. *For any $\mu \asymp n^\alpha$ with $0 \leq \alpha < 1$, with probability at least $1 - C \exp(-n^\eta)$,*

$$\frac{\mu}{n} Y^T (R + \mu I_n)^{-1} Y \asymp \sigma_\epsilon^2,$$

where $Y = (y_1, ..., y_n)^T$ with y_k defined in (2.1), and R is as in (2.4). Furthermore, if $\mu^{-1} = O_P(n^{-\frac{d}{2\nu+d}})$, then

$$\frac{\mu}{n} Y^T (R + \mu I_n)^{-1} Y \to \sigma_\epsilon^2$$

in probability.

Lemma F.7. *Suppose A, B and $C \in \mathbb{R}^{n \times n}$ are positive definite matrices. We have*

$$tr((A + B)(A + B + C)^{-1}) \geq tr(A(A + C)^{-1}),$$

and

$$tr((A + B)^2 (A + B + C)^{-2}) \geq tr(A^2 (A + C)^{-2}).$$

Lemma F.8. *Suppose $f \in \mathcal{N}_\Psi(\Omega)$ and $0 \leq \alpha < 1$. With probability at least $1 - C_1 \exp(-C_2 n^{\eta_1})$, we have*

$$(f(x) - r(x)^T (R + \hat{\mu}_n I_n)^{-1} f(X))^2 \leq (1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x)) \|f\|_{\mathcal{N}_\Psi(\Omega)}^2,$$

and

$$1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x) \lesssim n^{(\alpha-1)(1-\frac{d}{2\nu})},$$

where $r(x)$ and R are as in (2.4), $\hat{\mu}_n \asymp n^\alpha$, and $f(X) = (f(x_1), ..., f(x_n))^T$.

We first state the intuition behind the proof. Direct calculation shows that

$$\mathbb{E}_\epsilon \|f - \hat{f}_n\|_{L_2(\Omega)}^2$$

$$= \mathbb{E}_\epsilon \int_\Omega (f(x) - r(x)^T (K + \hat{\mu}_n I_n)^{-1} f(X) - r(x)^T (K + \hat{\mu}_n I_n)^{-1} \epsilon)^2 dx$$

$$= \underbrace{\sigma_\epsilon^2 \int_\Omega r(x)^T (K + \hat{\mu}_n I_n)^{-2} r(x) dx}_{variance} + \underbrace{\int_\Omega (f(x) - r(x)^T (K + \hat{\mu}_n I_n)^{-1} f(X))^2 dx}_{bias}.$$

If $\hat{\mu}_n$ is large $(\alpha > \frac{d}{2\nu+d})$, the bias dominates. Therefore, to obtain a lower bound of the mean-squared prediction error for the case $\alpha > \frac{d}{2\nu+d}$, we only need to obtain a lower bound of the bias term. On the other hand, if $\hat{\mu}_n$ is small $(\alpha < \frac{d}{2\nu+d})$, we only need to obtain a lower bound for the variance term.

Now we present the proof of Theorem 4.1. We first consider the case $\alpha > \frac{d}{2\nu+d}$.

By the proof of Lemma F.6, it can be seen that

$$\hat{f}_n = \operatorname*{argmin}_{g \in \mathcal{N}_\Psi(\Omega)} \left(\frac{1}{n} \sum_{i=1}^n (y_i - g(x_i))^2 + \frac{\hat{\mu}_n}{n} \|g\|_{\mathcal{N}_\Psi(\Omega)}^2 \right). \tag{F.2}$$

In the rest of proof we will write $\hat{f}_n$ as $\hat{f}$ for simplification. Plugging (2.1) into the objective function of (F.2), we have

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \frac{\hat{\mu}_n}{n} \|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2$$

$$= \|f - \hat{f}\|_n^2 + 2\langle \epsilon, f - \hat{f} \rangle_n + \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 + \frac{\hat{\mu}_n}{n} \|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2. \tag{F.3}$$

By Lemma F.1,

$$|\langle \epsilon, f - \hat{f} \rangle_n| \le t n^{-\frac{1}{2}} \|f - \hat{f}\|_n^{1 - \frac{d}{2\nu}} \|f - \hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^{\frac{d}{2\nu}},$$

with probability at least $1 - C_1 \exp(-C_2 t^2)$. By Lemma F.2 and the triangle inequality,

$$\|f - \hat{f}\|_{\mathcal{N}_\Psi(\Omega)} \le \|f\|_{\mathcal{N}_\Psi(\Omega)} + \|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)} \le C_3.$$

Therefore, (F.3) can be lower bounded by

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \frac{\hat{\mu}_n}{n} \|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2$$

$$\ge \|f - \hat{f}\|_n^2 + \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 + \frac{\hat{\mu}_n}{n} \|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2 - 2C_3 t n^{-\frac{1}{2}} \|f - \hat{f}\|_n^{1 - \frac{d}{2\nu}}. \tag{F.4}$$

By Lemma F.2 and the interpolation inequality, we have $\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)} \le C_4$, and $\|\hat{f}\|_{L_\infty(\Omega)} \le c \|\hat{f}\|_{L_2(\Omega)}^{1 - \frac{d}{2\nu}} \|\hat{f}\|_{H^\nu(\Omega)}^{\frac{d}{2\nu}} \le c_1 \|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)} \le (c_1 \vee 1) C_4 \le C_5$.

Let $\mathcal{F} = H^\nu(C_5)$, where $H^\nu(C_5)$ denotes the ball in the Sobolev space $H^\nu(\Omega)$ with radius C_5. Thus, the bracket entropy number can be bounded by [1]

$$H_B(\delta_n, \mathcal{F}, \|\cdot\|_{L_\infty(\Omega)}) \le C_6 \left(\frac{1}{\delta_n} \right)^{d/\nu},$$

and $\hat{f} \in \mathcal{F}$. Hence, $J_\infty^2(C_5, \mathcal{F})$ in (F.1) can be bounded by

$$J_\infty^2(C_5, \mathcal{F}) = C_7^2 \inf_{\delta > 0} E \left[C_5 \int_\delta^1 \sqrt{H(uC_5/2, \mathcal{F}, \|\cdot\|_{L_\infty(\Omega)})} du + \sqrt{n} z \delta \right]^2$$

$$\leq C_8^2\left[C_5\int_0^1\left(\frac{1}{uC_5}\right)^{d/(2\nu)}du\right]^2$$

$$=C_9^2C_5^2\left(\frac{1}{C_5}\right)^{d/\nu}\leq C_{10}. \tag{F.5}$$

By Lemma F.3, for all $t>0$, with probability at least $1-\exp(-t)$,

$$\sup_{f\in\mathcal{F}}\left|\|f\|_n^2-\|f\|_{L_2(\Omega)}^2\right|\leq C_{11}\left(\frac{2\mathcal{R}\sqrt{C_{10}}+\mathcal{R}C_5\sqrt{t}}{\sqrt{n}}+\frac{4C_{10}+C_5^2t}{n}\right),$$

where $\mathcal{R}=\sup_{f\in\mathcal{F}}\|f\|_{L_2(\Omega)}$. Choosing $t=n^\eta$, where $\eta=\left(\frac{(\alpha-1)(2\nu+d)}{2\nu}+1\right)/4$, for sufficient large n, we have that the right-hand side of (F.4) can be lower bounded by

$$\|f-\hat{f}\|_n^2+\frac{\hat{\mu}_n}{n}\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2-2C_3n^\eta n^{-\frac{1}{2}}\|f-\hat{f}\|_n^{1-\frac{d}{2\nu}}+\frac{1}{n}\sum_{i=1}^n\epsilon_i^2$$

$$\geq\|f-\hat{f}\|_{L_2(\Omega)}^2+\frac{\hat{\mu}_n}{n}\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2$$

$$-C_{12}n^{(\eta-1)/2}R-C_{12}n^{\eta-1}-C_{13}n^\eta n^{-\frac{1}{2}}\|f-\hat{f}\|_{L_2(\Omega)}^{1-\frac{d}{2\nu}}$$

$$-C_{13}n^\eta n^{-\frac{1}{2}}n^{(\eta-1)\left(1-\frac{d}{2\nu}\right)/4}R^{\left(1-\frac{d}{2\nu}\right)}-C_{13}n^\eta n^{-\frac{1}{2}}n^{(\eta-1)\left(1-\frac{d}{2\nu}\right)/2}+\frac{1}{n}\sum_{i=1}^n\epsilon_i^2, \tag{F.6}$$

where we also apply Jensen's inequality.

Case 1: If $2C_3n^\eta n^{-\frac{1}{2}}\|f-\hat{f}\|_{L_2(\Omega)}^{1-\frac{d}{2\nu}}\leq\|f-\hat{f}\|_{L_2(\Omega)}^2$, then we have $\|f-\hat{f}\|_{L_2(\Omega)}\gtrsim n^{(2\eta-1)\frac{\nu}{2\nu+d}}$.

Case 2: If $2C_3n^\eta n^{-\frac{1}{2}}\|f-\hat{f}\|_{L_2(\Omega)}^{1-\frac{d}{2\nu}}>\|f-\hat{f}\|_{L_2(\Omega)}^2$, we have $\|f-\hat{f}\|_{L_2(\Omega)}<C_{14}n^{(2\eta-1)\frac{\nu}{2\nu+d}}$. Consider function class $\mathcal{G}=\{g:\|g\|_{L_2(\Omega)}\leq C_{14}n^{(2\eta-1)\frac{\nu}{2\nu+d}}\}\cap\mathcal{F}$, we have $f-\hat{f}\in\mathcal{G}$ and $R_G=\sup_{g\in\mathcal{G}}\|g\|_{L_2(\Omega)}\leq C_{14}n^{(2\eta-1)\frac{\nu}{2\nu+d}}$. By (F.6), we have for sufficient large n,

$$\|f-\hat{f}\|_n^2+\frac{\hat{\mu}_n}{n}\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2-2C_3n^\eta n^{-\frac{1}{2}}\|f-\hat{f}\|_n^{1-\frac{d}{2\nu}}$$

$$\geq\|f-\hat{f}\|_{L_2(\Omega)}^2+\frac{\hat{\mu}_n}{n}\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2-C_{12}n^{(\eta-1)/2}R_G-C_{12}n^{\eta-1}-C_{13}n^\eta n^{-\frac{1}{2}}R_G^{1-\frac{d}{2\nu}}$$

$$-C_{13}n^\eta n^{-\frac{1}{2}}n^{(\eta-1)\left(1-\frac{d}{2\nu}\right)/4}R_G^{\left(1-\frac{d}{2\nu}\right)}-C_{13}n^\eta n^{-\frac{1}{2}}n^{(\eta-1)\left(1-\frac{d}{2\nu}\right)/2}$$

$$\geq\|f-\hat{f}\|_{L_2(\Omega)}^2+\frac{\hat{\mu}_n}{n}\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2-C_{15}n^\eta n^{-\frac{1}{2}}n^{(2\eta-1)\frac{\nu}{2\nu+d}(1-\frac{d}{2\nu})}. \tag{F.7}$$

Let

$$f_1=\underset{g\in\mathcal{N}_\Psi(\Omega)}{\arg\min}\|f-g\|_{L_2(\Omega)}^2+\frac{\hat{\mu}_n}{2n}\|g\|_{\mathcal{N}_\Psi(\Omega)}^2. \tag{F.8}$$

Therefore, by (F.2), we have

$$\|f - \hat{f}\|_n^2 + 2\langle \epsilon, f - \hat{f}\rangle_n + \frac{\hat{\mu}_n}{n}\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2$$
$$\leq \|f - f_1\|_n^2 + 2\langle \epsilon, f - f_1\rangle_n + \frac{\hat{\mu}_n}{n}\|f_1\|_{\mathcal{N}_\Psi(\Omega)}^2,$$

which, together with (F.4) and Lemma F.1, implies

$$\|f - \hat{f}\|_n^2 + \frac{\hat{\mu}_n}{n}\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2 - 2C_3 n^\eta n^{-\frac{1}{2}}\|f - \hat{f}\|_n^{1-\frac{d}{2\nu}}$$
$$\leq \|f - f_1\|_n^2 + 2C_{16}n^\eta n^{-\frac{1}{2}}\|f - f_1\|_n^{1-\frac{d}{2\nu}} + \frac{\hat{\mu}_n}{n}\|f_1\|_{\mathcal{N}_\Psi(\Omega)}^2. \tag{F.9}$$

By (F.7) and (F.9), we have

$$\|f - \hat{f}\|_{L_2(\Omega)}^2 + \frac{\hat{\mu}_n}{n}\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2 - C_{15}n^\eta n^{-\frac{1}{2}}n^{(2\eta-1)\frac{\nu}{2\nu+d}(1-\frac{d}{2\nu})}$$
$$\leq \|f - f_1\|_n^2 + 2C_{16}n^\eta n^{-\frac{1}{2}}\|f - f_1\|_n^{1-\frac{d}{2\nu}} + \frac{\hat{\mu}_n}{n}\|f_1\|_{\mathcal{N}_\Psi(\Omega)}^2, \tag{F.10}$$

By (F.8), we have

$$\|f - f_1\|_{L_2(\Omega)}^2 + \frac{\hat{\mu}_n}{2n}\|f_1\|_{\mathcal{N}_\Psi(\Omega)}^2 \leq \frac{\hat{\mu}_n}{2n}\|f\|_{\mathcal{N}_\Psi(\Omega)}^2 \leq C_{17}\frac{\hat{\mu}_n}{n}. \tag{F.11}$$

By Lemma F.4, it can be shown that with probability at least $1 - \exp(-C_{18}n^\eta)$,

$$\|f - f_1\|_n^2 + 2C_{16}n^{\eta_1}n^{-\frac{1}{2}}\|f - f_1\|_n^{1-\frac{d}{2\nu}} + \frac{\hat{\mu}_n}{n}\|f_1\|_{\mathcal{N}_\Psi(\Omega)}^2$$
$$\leq \|f - f_1\|_{L_2(\Omega)}^2 + \frac{\hat{\mu}_n}{n}\|f_1\|_{\mathcal{N}_\Psi(\Omega)}^2$$
$$+ C_{17}n^\eta n^{-\frac{1}{2}}\|f - f_1\|_{L_2(\Omega)}^{1-\frac{d}{2\nu}} + C_{17}n^\eta n^{-\frac{1}{2}}n^{(\eta-1)\left(1-\frac{d}{2\nu}\right)/2} + C_{17}n^{\eta-1}$$
$$\leq \|f - f_1\|_{L_2(\Omega)}^2 + \frac{\hat{\mu}_n}{n}\|f_1\|_{\mathcal{N}_\Psi(\Omega)}^2$$
$$+ C_{17}n^\eta n^{-\frac{1}{2}}\|f - f_1\|_{L_2(\Omega)}^{1-\frac{d}{2\nu}} + 2C_{17}n^\eta n^{-\frac{1}{2}}n^{(\eta-1)\left(1-\frac{d}{2\nu}\right)/2}. \tag{F.12}$$

Combining (F.10) and (F.12) yields

$$\|f - \hat{f}\|_{L_2(\Omega)}^2 + \frac{\hat{\mu}_n}{n}\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2 - C_{15}n^\eta n^{-\frac{1}{2}}n^{(2\eta-1)\frac{\nu}{2\nu+d}(1-\frac{d}{2\nu})}$$
$$\leq \|f - f_1\|_{L_2(\Omega)}^2 + \frac{\hat{\mu}_n}{n}\|f_1\|_{\mathcal{N}_\Psi(\Omega)}^2$$
$$+ C_{17}n^\eta n^{-\frac{1}{2}}\|f - f_1\|_{L_2(\Omega)}^{1-\frac{d}{2\nu}} + 2C_{17}n^\eta n^{-\frac{1}{2}}n^{(\eta-1)\left(1-\frac{d}{2\nu}\right)/2}. \tag{F.13}$$

By (F.8), we have

$$\|f - f_1\|_{L_2(\Omega)}^2 + \frac{\hat{\mu}_n}{2n}\|f_1\|_{\mathcal{N}_\Psi(\Omega)}^2 \leq \|f - \hat{f}\|_{L_2(\Omega)}^2 + \frac{\hat{\mu}_n}{2n}\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2,$$

which implies

$$2\|f - f_1\|_{L_2(\Omega)}^2 + \frac{\hat{\mu}_n}{n}\|f_1\|_{\mathcal{N}_\Psi(\Omega)}^2 \leq 2\|f - \hat{f}\|_{L_2(\Omega)}^2 + \frac{\hat{\mu}_n}{n}\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2. \qquad (\text{F.14})$$

Combining (F.11), (F.13) and (F.14), we have for sufficient large n,

$$
\begin{aligned}
\|f - \hat{f}\|_{L_2(\Omega)}^2 \geq &\|f - f_1\|_{L_2(\Omega)}^2 - C_{15}n^\eta n^{-\frac{1}{2}}n^{(2\eta-1)\frac{\nu}{2\nu+d}(1-\frac{d}{2\nu})} \\
&- C_{17}n^\eta n^{-\frac{1}{2}}\|f - f_1\|_{L_2(\Omega)}^{1-\frac{d}{2\nu}} - 2C_{17}n^\eta n^{-\frac{1}{2}}n^{(\eta-1)\left(1-\frac{d}{2\nu}\right)/2} \\
\geq &\|f - f_1\|_{L_2(\Omega)}^2 - C_{15}n^\eta n^{-\frac{1}{2}}n^{(2\eta-1)\frac{\nu}{2\nu+d}(1-\frac{d}{2\nu})} \\
&- C_{17}n^\eta n^{-\frac{1}{2}}n^{\left(1-\frac{d}{2\nu}\right)(\alpha-1)/2} - 2C_{17}n^\eta n^{-\frac{1}{2}}n^{(\eta-1)\left(1-\frac{d}{2\nu}\right)/2} \\
\geq &\|f - f_1\|_{L_2(\Omega)}^2 - 4C_{15}n^\eta n^{-\frac{1}{2}}n^{(2\eta-1)\frac{\nu}{2\nu+d}(1-\frac{d}{2\nu})}, \qquad (\text{F.15})
\end{aligned}
$$

where the last inequality is because $(2\eta-1)\frac{2\nu}{2\nu+d} > \alpha - 1$. Let

$$\tilde{f} = \operatorname*{argmin}_{g \in \mathcal{N}_\Psi(\mathbb{R}^d)} \|f - g\|_{L_2(\mathbb{R}^d)}^2 + \frac{\hat{\mu}_n}{C_{18}n}\|g\|_{\mathcal{N}_\Psi(\mathbb{R}^d)}^2, \qquad (\text{F.16})$$

where C_{18} is a constant determined later. By (F.8) and the definition of $\|\cdot\|_{\mathcal{N}_\Psi(\Omega)}^2$, we have

$$
\begin{aligned}
\|f - f_1\|_{L_2(\Omega)}^2 + \frac{\hat{\mu}_n}{2n}\|f_1\|_{\mathcal{N}_\Psi(\Omega)}^2 &\leq \|f - \tilde{f}\|_{L_2(\Omega)}^2 + \frac{\hat{\mu}_n}{2n}\|\tilde{f}\|_{\mathcal{N}_\Psi(\Omega)}^2 \\
&\leq \|f - \tilde{f}\|_{L_2(\mathbb{R}^d)}^2 + \frac{\hat{\mu}_n}{2n}\|\tilde{f}\|_{\mathcal{N}_\Psi(\mathbb{R}^d)}^2. \qquad (\text{F.17})
\end{aligned}
$$

By (F.16) and the extension theorem (we still use f_1 to denote the extension of f_1 for notational simplicity),

$$
\begin{aligned}
&\|f - \tilde{f}\|_{L_2(\mathbb{R}^d)}^2 + \frac{\hat{\mu}_n}{C_{18}n}\|\tilde{f}\|_{\mathcal{N}_\Psi(\mathbb{R}^d)}^2 \\
\leq &\|f - f_1\|_{L_2(\mathbb{R}^d)}^2 + \frac{\hat{\mu}_n}{C_{18}n}\|f_1\|_{\mathcal{N}_\Psi(\mathbb{R}^d)}^2 \\
\leq &\frac{C_{19}}{C_{18}}\left(C_{18}\|f - f_1\|_{L_2(\Omega)}^2 + \frac{\hat{\mu}_n}{n}\|f_1\|_{\mathcal{N}_\Psi(\Omega)}^2\right) \\
\leq &\frac{C_{19}}{C_{18}}\left((C_{18}-1)\|f - f_1\|_{L_2(\Omega)}^2 + \|f - \tilde{f}\|_{L_2(\Omega)}^2 + \frac{\hat{\mu}_n}{n}\|\tilde{f}\|_{\mathcal{N}_\Psi(\mathbb{R}^d)}^2\right), \qquad (\text{F.18})
\end{aligned}
$$

where $C_{19} > 1$ is a constant. Therefore, combining (F.17) and (F.18) yields

$$\|f - f_1\|_{L_2(\Omega)}^2 \geq \frac{1}{C_{18}-1}\left((1 - \frac{C_{19}}{C_{18}})\|f - \tilde{f}\|_{L_2(\mathbb{R}^d)}^2 - \frac{\hat{\mu}_n(C_{19}-1)}{C_{18}n}\|\tilde{f}\|_{\mathcal{N}_\Psi(\mathbb{R}^d)}^2\right) \qquad (\text{F.19})$$

Next, we calculate $\|f - \tilde{f}\|_{L_2(\mathbb{R}^d)}^2$ and $\|\tilde{f}\|_{\mathcal{N}_\Psi(\mathbb{R}^d)}^2$ with respect to f.

By Fourier transform and (F.16),

$$
\begin{aligned}
&\|f - \tilde{f}\|^2_{L_2(\mathbb{R}^d)} + \frac{\hat{\mu}_n}{C_{18}n}\|\tilde{f}\|^2_{\mathcal{N}_\Psi(\mathbb{R}^d)} \\
&= \int_{\mathbb{R}^d} |\mathcal{F}(f)(\omega) - \mathcal{F}(\tilde{f})(\omega)|^2 d\omega + \frac{\hat{\mu}_n}{C_{18}n}\int_{\mathbb{R}^d}|\mathcal{F}(\tilde{f})(\omega)|^2(1+|\omega|^2)^\nu d\omega \\
&= \int_{\mathbb{R}^d} |\mathcal{F}(f)(\omega) - \mathcal{F}(\tilde{f})(\omega)|^2 + \frac{\hat{\mu}_n}{C_{18}n}|\mathcal{F}(\tilde{f})(\omega)|^2(1+|\omega|^2)^\nu d\omega \\
&= \int_{\mathbb{R}^d} \frac{\frac{\hat{\mu}_n}{C_{18}n}(1+|\omega|^2)^\nu}{1+\frac{\hat{\mu}_n}{C_{18}n}(1+|\omega|^2)^\nu}|\mathcal{F}(f)(\omega)|^2 d\omega,
\end{aligned}
$$

where

$$
\mathcal{F}(\tilde{f})(\omega) = \frac{\mathcal{F}(f)(\omega)}{1+\frac{\hat{\mu}_n}{C_{18}n}(1+|\omega|^2)^\nu}.
$$

Therefore,

$$
\int_{\mathbb{R}^d}|\mathcal{F}(f)(\omega) - \mathcal{F}(\tilde{f})(\omega)|^2 d\omega = \int_{\mathbb{R}^d}|\mathcal{F}(f)(\omega)|^2 \frac{(\frac{\hat{\mu}_n}{C_{18}n}(1+|\omega|^2)^\nu)^2}{(1+\frac{\hat{\mu}_n}{C_{18}n}(1+|\omega|^2)^\nu)^2}d\omega, \tag{F.20}
$$

and

$$
\frac{\hat{\mu}_n}{C_{18}n}\int_{\mathbb{R}^d}|\mathcal{F}(\tilde{f})(\omega)|^2(1+|\omega|^2)^\nu d\omega = \int_{\mathbb{R}^d}|\mathcal{F}(f)(\omega)|^2\frac{\frac{\hat{\mu}_n}{C_{18}n}(1+|\omega|^2)^\nu}{(1+\frac{\hat{\mu}_n}{C_{18}n}(1+|\omega|^2)^\nu)^2}d\omega. \tag{F.21}
$$

Let $h(|\omega|) = \frac{\hat{\mu}_n}{C_{18}n}(1+|\omega|^2)^\nu$ and $C_{18} = C_{19}^2$. Plugging (F.20) and (F.21) into (F.19), we have

$$
\begin{aligned}
&\|f - f_1\|^2_{L_2(\Omega)} \\
&\geq \frac{1}{C_{18}-1}\int_{\mathbb{R}^d}\frac{|\mathcal{F}(f)(\omega)|^2h(|\omega|)}{(1+h(|\omega|))^2}\left(\frac{C_{18}-C_{19}}{C_{18}}h(|\omega|) - (C_{19}-1)\right)d\omega \\
&\geq \frac{1}{C_{19}+1}\int_{\mathbb{R}^d}\frac{|\mathcal{F}(f)(\omega)|^2h(|\omega|)}{(1+h(|\omega|))^2}(h(|\omega|)/C_{19}-1)d\omega \\
&= \frac{1}{C_{19}+1}\int_{\{\omega:h(|\omega|)\leq 2C_{19}\}}\frac{|\mathcal{F}(f)(\omega)|^2h(|\omega|)}{(1+h(|\omega|))^2}(h(|\omega|)/C_{19}-1)d\omega \\
&\quad + \frac{1}{C_{19}+1}\int_{\{\omega:h(|\omega|)>2C_{19}\}}\frac{|\mathcal{F}(f)(\omega)|^2h(|\omega|)}{(1+h(|\omega|))^2}(h(|\omega|)/C_{19}-1)d\omega \\
&\geq \frac{1}{C_{19}+1}\int_{\{\omega:h(|\omega|)\leq 2C_{19}\}}\frac{|\mathcal{F}(f)(\omega)|^2h(|\omega|)}{(1+h(|\omega|))^2}(h(|\omega|)/C_{19}-1)d\omega \\
&\quad + \frac{1}{4(C_{19}+1)}\int_{\{\omega:h(|\omega|)>2C_{19}\}}\frac{|\mathcal{F}(f)(\omega)|^2}{2C_{19}}d\omega
\end{aligned}
$$

$$\geq -\frac{1}{C_{19}+1} \int_{\{\omega:h(|\omega|)\leq 2C_{19}\}} |\mathcal{F}(f)(\omega)|^2 d\omega$$

$$+ \frac{1}{4(C_{19}+1)} \int_{\{\omega:h(|\omega|)>2C_{19}\}} \frac{|\mathcal{F}(f)(\omega)|^2}{2C_{19}} d\omega. \tag{F.22}$$

Now we can build our function f. Let

$$\mathcal{F}(f)(\omega) = \begin{cases} 0 & h(|\omega|) \leq 2C_{19} - \delta_n, \\ g_1(|\omega|) & 2C_{19} - \delta_n \leq h(|\omega|) \leq 2C_{19}, \\ (1+|\omega|^2)^{-\nu/2-d/2-1} & h(|\omega|) > 2C_{19}, \end{cases}$$

where g_1 and δ_n are chosen such that f is continuous and

$$\frac{1}{C_{19}+1} \int_{\{\omega:2C_{19}-\delta_n \leq h(|\omega|)\leq 2C_{19}\}} |\mathcal{F}(f)(|\omega|)|^2 d\omega$$

$$< \frac{1}{8(C_{19}+1)} \int_{\{\omega:h(|\omega|)>2C_{19}\}} \frac{|\mathcal{F}(f)(|\omega|)|^2}{2C_{19}}.$$

Then we normalize f such that $\|f\|_{\mathcal{N}_\Psi(\mathbb{R}^d)} = 1$. Therefore, by (F.22), direct calculation shows that

$$\|f - f_1\|_{L_2(\Omega)}^2 \geq C_{20} \frac{\hat{\mu}_n}{n}.$$

By (F.15) and $(2\eta-1)\frac{\nu}{2\nu+d}(1-\frac{d}{2\nu})+\eta-1/2 < \alpha-1$, we have

$$\|f - \hat{f}\|_{L_2(\Omega)}^2 \geq C_{21} \frac{\hat{\mu}_n}{n}.$$

Note that $(2\eta-1)\frac{2\nu}{2\nu+d} > \alpha-1$, which leads to a contradiction of Case 2 as n increases.

Using this constructed f, it can be seen that

$$\mathbb{E}_\epsilon \|f - \hat{f}\|_{L_2(\Omega)}^2 \geq C_6 n^{-(1-2\eta)\frac{2\nu}{2\nu+d}}.$$

Therefore, we finish the proof of the case $\alpha > \frac{d}{2\nu+d}$.

Next, we prove the case $\alpha < \frac{d}{2\nu+d}$. We consider the case $0 \leq \alpha < \frac{d}{2\nu+d}$.

By Fubini's theorem,

$$\mathbb{E}\|f - \hat{f}_n\|_{L_2(\Omega)}^2$$

$$= \mathbb{E} \int_\Omega (f(x) - r(x)^T(R+\hat{\mu}_n I_n)^{-1}f(X) - r(x)^T(R+\hat{\mu}_n I_n)^{-1}\epsilon)^2 dx$$

$$= \sigma_\epsilon^2 \int_\Omega r(x)^T(R+\hat{\mu}_n I_n)^{-2}r(x)dx + \int_\Omega (f(x) - r(x)^T(R+\hat{\mu}_n I_n)^{-1}f(X))^2 dx$$

$$\geq \sigma_\epsilon^2 \int_\Omega r(x)^T(R+\hat{\mu}_n I_n)^{-2}r(x)dx := \sigma_\epsilon^2 I. \tag{F.23}$$

We consider a discrete version of I. Let $I_1 = \operatorname{tr}(R^2(R + \hat{\mu}_n I_n)^{-2})$. Let $p = \lfloor (n/\hat{\mu}_n)^{d/(2\nu)} \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function, and $p_1 = \min\{p, C_1 n^{1/2}\}$. Let $\Psi_1 = \frac{1}{\sqrt{n}}(\varphi_1(X), ..., \varphi_{p_1}(X))$, and $\Psi_2 = \frac{1}{\sqrt{n}}(\varphi_{p_1+1}(X), \varphi_{p_1+2}(X), ...)$, where $\varphi_k(X) = (\varphi_k(x_1), ..., \varphi_k(x_n))^T$ for $k = 1, 2, ...$, and φ_k's are as in (L.1). Let $\Lambda_1 = \operatorname{diag}(n\lambda_1, ..., n\lambda_p)$ and $\Lambda_2 = \operatorname{diag}(n\lambda_{p+1}, ...)$, where λ_k's are as in (L.1). Therefore, $R = \sum_{k=1}^{\infty} \lambda_i \varphi_k(X) \varphi_k(X)^T = \Psi_1 \Lambda_1 \Psi_1^T + \Psi_2 \Lambda_2 \Psi_2^T$.

By Lemma F.7,

$$
\begin{aligned}
I_1 \geq & \operatorname{tr}((\Psi_1 \Lambda_1 \Psi_1^T)^2 (\Psi_1 \Lambda_1 \Psi_1^T + \hat{\mu}_n I_n)^{-2}) \\
= & \sum_{i=1}^{p_1} \left(\frac{\lambda_i(\Psi_1 \Lambda_1 \Psi_1^T)}{\lambda_i(\Psi_1 \Lambda_1 \Psi_1^T) + \hat{\mu}_n} \right)^2,
\end{aligned}
\tag{F.24}
$$

where $\lambda_i(\Psi_1 \Lambda_1 \Psi_1^T)$ denote the i-th eigenvalue of $\Psi_1 \Lambda_1 \Psi_1^T$. Note that the i-th eigenvalue $\lambda_i(\Psi_1 \Lambda_1 \Psi_1^T) = \lambda_i(\Psi_1^T \Psi_1 \Lambda_1)$ for $i = 1, ..., p$, because if u_i is eigenvector corresponding to i-th eigenvalue of $\Psi_1 \Lambda_1 \Psi_1^T$, then

$$
\Psi_1 \Lambda_1 \Psi_1^T u_i = \lambda_i u_i \Rightarrow \Psi_1^T \Psi_1 \Lambda_1 \Psi_1^T u_i = \lambda_i \Psi_1^T u_i.
$$

Therefore, (F.24) implies

$$
\begin{aligned}
I_1 \geq & \operatorname{tr}((\Lambda_1 \Psi_1^T \Psi_1)^2 (\Lambda_1 \Psi_1^T \Psi_1 + \hat{\mu}_n I_n)^{-2}) \\
= & \operatorname{tr}(\Lambda_1^2 (\Lambda_1 + \hat{\mu}_n (\Psi_1^T \Psi_1)^{-1})^{-2}).
\end{aligned}
$$

By Lemma L.2, it can be shown that $\lambda_{\min}(\Psi_1^T \Psi_1) \geq \eta_1$ with probability at least $1 - C_{22} \exp(-C_{23} n^{\eta_2})$, where the constant in the expression of p_1 is chosen such that the condition of Lemma F.5 is satisfied. Combining this with Lemma L.1, we have $I_1 \geq C_{24} p_1$.

Notice that for any $u = (u_1, ..., u_n)^T \in \mathbb{R}^n$,

$$
1 - 2 \sum_{i=1}^{n} u_i \Psi(x - x_i) + \sum_{i=1}^{n} \sum_{j=1}^{n} u_i u_j \Psi(x_i - x_j) + \hat{\mu}_n \|u\|_2^2 \geq \hat{\mu}_n \|u\|_2^2.
$$

Plugging $u = (R + \hat{\mu}_n I_n)^{-1} r(x)$, we have

$$
\hat{\mu}_n r(x)^T (R + \hat{\mu}_n I_n)^{-2} r(x) \leq 1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x).
\tag{F.25}
$$

By Lemma F.8, with probability at least $1 - C_{25} \exp(-C_{26} n^{\eta_2})$, $1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x) \leq C_{27} \left(\frac{\hat{\mu}_n}{n} \right)^{\frac{2\nu - d}{2\nu}}$.

Let $\mathcal{H}_1 = \{h | h(x)^2 = r(x)^T (R + \hat{\mu}_n I_n)^{-2} r(x), \text{with } x_i \in \Omega, i = 1, ..., n\}$. Let $\mathcal{H} = \mathcal{H}_1 \bigcap \{\|h\|_{L_\infty(\Omega)}^2 \leq C_{27} \left(\frac{\hat{\mu}_n}{n} \right)^{\frac{2\nu - d}{2\nu}} / \hat{\mu}_n\}$. It can be seen from (F.25) that with probability at least $1 - C_{25} \exp(-C_{26} n^{\eta_3})$, $\mathcal{H}$ is true. It can be also seen that $\|h_1\|_n^2 = \frac{1}{n} \operatorname{tr}(R^2 (R + \hat{\mu}_n I_n)^{-2})$ for some $h_1(x) = r(x)^T (R + \hat{\mu}_n I_n)^{-2} r(x) \in \mathcal{H}_1$.

By Lemma F.3, we have with another probability at least $1 - \exp(-n^{\eta_3})$ with $\eta_3 = (1 - \alpha) \frac{2\nu - d}{2d}$,

$$
\sup_{h \in \mathcal{H}_1} \left| \|h\|_n^2 - \|h\|_{L_2(\Omega)}^2 \right| \leq C_7 n^{-\frac{1}{2} + \left((\alpha - 1) \frac{2\nu - d}{2\nu}\right)\left(2 - \frac{d}{2\nu}\right)},
$$

where $J_\infty^2(\sqrt{C_{27}}\left(\frac{\hat\mu_n}{n}\right)^{\frac{2\nu-d}{4\nu}}/\hat\mu_n^{1/2}, \mathcal{F})$ can be calculate similarly. Therefore, we have

$$\|h\|_{L_2(\Omega)}^2 \geq \|h\|_n^2 - C_{27}n^{-\frac{1}{2}+\left((\alpha-1)\frac{2\nu-d}{2\nu}\right)\left(2-\frac{d}{2\nu}\right)}. \tag{F.26}$$

Note

$$\|h\|_n^2 = I_1/n \geq C_4 p_1/n \geq C_8 \min\{n^{-1+(1-\alpha)\frac{d}{2\nu}}, n^{-1/2}\}$$
$$> n^{-\frac{1}{2}+\left((\alpha-1)\frac{2\nu-d}{2\nu}\right)\left(2-\frac{d}{2\nu}\right)}$$

with probability at least $1 - C_{22}\exp(-C_{23}n^{\eta_2})$. Therefore, combining all probabilities together and by (F.26), with probability at least $1 - C_{28}\exp(-C_{29}n^{\eta_4})$, we have $\|h\|_{L_2(\Omega)}^2 \geq C_{30}p_1/n$, which finishes the proof of the case $\alpha \in [0, \frac{d}{2\nu+d})$. If $\alpha < 0$, then from (F.23) it can be seen that the error is larger than choosing $\alpha = 0$. Thus, we complete the proof.

Appendix G: Proof of Theorem 4.2

In the proof of Theorem 4.2, we hide all the probabilities with the form $1 - C\exp(-C'n^{-\eta})$ for the conciseness of the proof.

Case 1: $\alpha > \frac{d}{2\nu+d}$.

By (F.23) and Lemma F.6, we have

$$\mathbb{E}_\epsilon\|(f-\hat f_n)/\tilde c_{n,\beta}(\cdot;\hat\mu_n)\|_{L_2(\Omega)}^2 \geq C_1\frac{\mathbb{E}_\epsilon\|f-\hat f_n\|_{L_2(\Omega)}^2}{\mathbb{E}_\epsilon\|\tilde c_{n,\beta}(\cdot;\hat\mu_n)\|_{L_2(\Omega)}^2}$$

$$\geq C_2\frac{\mathbb{E}_\epsilon\int_\Omega \hat\mu_n(f(x)-r(x)^T(R+\hat\mu_n I_n)^{-1}f(X)-r(x)^T(R+\hat\mu_n I_n)^{-1}\epsilon)^2 dx}{\int_\Omega(1-r(x)^T(R+\hat\mu_n I_n)^{-1}r(x))dx}$$

$$= C_2\frac{\int_\Omega \hat\mu_n(r(x)^T(R+\hat\mu_n I_n)^{-2}r(x)+(f(x)-r(x)^T(R+\hat\mu_n I_n)^{-1}f(X))^2)dx}{\int_\Omega 1-r(x)^T(R+\hat\mu_n I_n)^{-1}r(x)dx}$$

$$\geq C_3 n^{(1-\alpha)(1-\frac{d}{2\nu})}\hat\mu_n\int_\Omega(f(x)-r(x)^T(R+\hat\mu_n I_n)^{-1}f(X))^2 dx,$$

where $r(x)$ and R are as in (2.4), and $f(X) = (f(x_1),...,f(x_n))^T$. The first inequality is true because of the Cauchy-Schwarz inequality and that $f-\hat f_n$ is normal. The last inequality follows Lemma F.8. If $\alpha > \frac{d}{2\nu+d}$, then by the proof of Theorem 4.1, we have $n^{(1-\alpha)(1-\frac{d}{2\nu})}\hat\mu_n\int_\Omega(f(x)-r(x)^T(R+\hat\mu_n I_n)^{-1}f(X))^2 \gtrsim n^\beta$ with $\beta > 0$ for $f \in \mathcal{N}_\Psi(\Omega)$ constructed in the proof of Theorem 4.1, which finishes the proof of Case 1.

Case 2: $0 \leq \alpha < \frac{d}{2\nu+d}$.

First, we prove (4.7). By the proof of Theorem 4.1 and Lemma F.6, it can be shown that

$$\hat\mu_n\mathbb{E}_\epsilon\|\tilde c_{n,\beta}(\cdot;\hat\mu_n)\|_{L_2(\Omega)}^2 \gtrsim n^{(\alpha-1)(1-\frac{d}{2\nu})}.$$

Noting that $f - \hat{f}_n$ is normal, we have

$$\hat{\mu}_n \mathbb{E}_\epsilon \| f - \hat{f}_n \|^2_{L_2(\Omega)}$$

$$\leq C_1 \hat{\mu}_n \int_\Omega r(x)^T (R + \hat{\mu}_n I_n)^{-2} r(x) + (f(x) - r(x)^T (R + \hat{\mu}_n I_n)^{-1} f(X))^2 dx$$

$$\leq 2 C_1 \hat{\mu}_n \int_\Omega r(x)^T (R + \hat{\mu}_n I_n)^{-2} r(x) + (f(x) - r(x)^T R^{-1} f(X))^2$$
$$+ (r(x)^T R^{-1} f(X) - r(x)^T (R + \hat{\mu}_n I_n)^{-1} f(X))^2 dx$$

$$\leq 2 C_1 \hat{\mu}_n \int_\Omega r(x)^T (R + \hat{\mu}_n I_n)^{-2} r(x) + (1 - r(x)^T R^{-1} r(X)) \log n$$
$$+ (r(x)^T R^{-1} f(X) - r(x)^T (R + \hat{\mu}_n I_n)^{-1} f(X))^2 dx$$

$$\leq 2 C_1 \hat{\mu}_n \int_\Omega r(x)^T (R + \hat{\mu}_n I_n)^{-2} r(x) + (1 - r(x)^T R^{-1} r(x)) \log n$$
$$+ (r(x)^T R^{-1} f(X) - r(x)^T (R + C_2 n^{\frac{d}{2\nu+d}} I_n)^{-1} f(X))^2 dx$$

$$\leq 2 C_1 \int_\Omega (1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x)) + \hat{\mu}_n (1 - r(x)^T R^{-1} r(x)) \log n$$
$$+ \hat{\mu}_n (r(x)^T R^{-1} f(X) - r(x)^T (R + C_2 n^{\frac{d}{2\nu+d}} I_n)^{-1} f(X))^2 dx, \qquad \text{(G.1)}$$

where the first inequality is by Fubini's theorem, the second inequality is by the Cauchy-Schwarz inequality, the third inequality is by (2.20) and the fact $\| f \|^2_{\mathcal{N}_\Psi(\Omega)} \leq \log n$, the fourth inequality is by $\hat{\mu}_n \leq C_2 n^{\frac{d}{2\nu+d}}$, and the fifth inequality is by (F.25). By Lemma C.2 and Lemma F.8, (G.1) can be further bounded by

$$\hat{\mu}_n \mathbb{E}_\epsilon \| f - \hat{f}_n \|^2_{L_2(\Omega)}$$
$$\leq C_2 n^{(\alpha-1)(1-\frac{d}{2\nu})} + n^{\alpha - \frac{\nu}{d} + 1/2} \log n$$
$$+ \hat{\mu}_n \int_\Omega (r(x)^T R^{-1} f(X) - r(x)^T (R + C_2 n^{\frac{d}{2\nu+d}} I_n)^{-1} f(X))^2 dx.$$

Let $f_1(x) = (r(x)^T R^{-1} f(X) - r(x)^T (R + C_2 n^{\frac{d}{2\nu+d}} I_n)^{-1} f(X)) / \sqrt{\log n}$. Then $\| f_1 \|^2_{\mathcal{N}_\Psi(\Omega)} \leq 1$, which implies either $\| f_1 \|^2_{L_2(\Omega)} \leq C_3 n^{-\frac{2\nu}{2\nu+d}}$ or the conditions of Lemma F.5 are satisfied. If the later happens, by Lemma F.5, it can be shown that

$$\| f_1 \|^2_{L_2(\Omega)} \leq \eta_2 \| f_1 \|^2_n$$
$$\leq \eta_2 (\| r(x)^T R^{-1} f(X) - r(x)^T (R + C_2 n^{\frac{d}{2\nu+d}} I_n)^{-1} f(X) \|^2_n$$
$$+ C_2 n^{-\frac{2\nu}{2\nu+d}} \| r(x)^T (R + C_2 n^{\frac{d}{2\nu+d}} I_n)^{-1} f(X) \|^2_{\mathcal{N}_\Psi(\Omega)})$$
$$\leq C_4 n^{-\frac{2\nu}{2\nu+d}}.$$

By noticing that $n^{\alpha - \frac{\nu}{d} + 1/2} \log n + n^{-\frac{2\nu}{2\nu+d}} n^\alpha \log n \leq C_5 n^{(\alpha-1)(1-\frac{d}{2\nu})}$, we finish the proof of the first part.

For the second part, by the standard minimax theory in nonparametric regression, there exists a function f such that $\mathbb{E}_\epsilon \|f - \hat{f}_n\|^2_{L_2(\Omega)} \geq C_6 n^{-\frac{2\nu}{2\nu+d}} \|f\|^2_{\mathcal{N}_\Psi(\Omega)}$, which implies

$$\mathbb{E}_\epsilon \|(f - \hat{f}_n)/\tilde{c}_{n,\beta}(\cdot; \hat{\mu}_n)\|^2_{L_2(\Omega)} \geq C_7 \frac{\mathbb{E}_\epsilon \|f - \hat{f}_n\|^2_{L_2(\Omega)}}{\mathbb{E}_\epsilon \|\tilde{c}_{n,\beta}(\cdot; \hat{\mu}_n)\|^2_{L_2(\Omega)}}$$

$$\geq C_8 n^{(1-\alpha)(1-\frac{d}{2\nu})} \hat{\mu}_n \mathbb{E}_\epsilon \|f - \hat{f}_n\|^2_{L_2(\Omega)}$$

$$\geq C_9 a_n.$$

Case 3: $\alpha < 0$.

By Lemma F.6, we have

$$\mathbb{E}_\epsilon \|(f - \hat{f}_n)/\tilde{c}_n(\cdot, \beta; \hat{\mu}_n)\|^2_{L_2(\Omega)}$$

$$\leq C_{10} \hat{\mu}_n \int_\Omega \frac{r(x)^T (R + \hat{\mu}_n I_n)^{-2} r(x) + (f(x) - r(x)^T (R + \hat{\mu}_n I_n)^{-1} f(X))^2}{(1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x))} dx$$

$$\leq C_{10} + C_{10} \hat{\mu}_n \int_\Omega \frac{(f(x) - r(x)^T (R + \hat{\mu}_n I_n)^{-1} f(X))^2}{(1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x))} dx$$

$$\leq C_{10} + C_{10} \hat{\mu}_n \log n.$$

The second inequality is true because of (F.25), and the third inequality is true because of Lemma F.8. Note $\hat{\mu}_n \log n \to 0$, which finishes the proof of the case $\alpha < 0$.

Appendix H: Proof of Lemma F.5

The idea of the proof is to use the bracket entropy number. By the definition of the bracket entropy number, we can find finite functions g_s's such that the ball with small radius centered on g_s's can cover the function class $\mathcal{G}$. By showing the results hold for these g_s's, we can show the results hold for all function $g \in \mathcal{G}$.

Take $g \in \mathcal{G}$, and suppose that $s\delta_n \leq \|g\|_{L_2(\Omega)} \leq (s+1)\delta_n$, where $s \in \{2, 3, ...\}$. Let $-K \leq g_L \leq g \leq g_U \leq K$, and $\|g_U - g_L\|_{L_\infty(\Omega)} \leq \delta_n/\text{Vol}(\Omega)$, for functions g_L and g_U. For $0 < C \leq \frac{1}{4\text{Vol}(\Omega)}$, by Cauchy-Schwarz inequality, we have

$$g_L^2 \leq 2g^2/C + 2C(g - g_L)^2 \leq 2g^2/C + 2C\delta_n^2/\text{Vol}(\Omega)^2,$$

which implies

$$2\|g\|_n^2 \geq C\|g_L\|_n^2 - 2C^2 \delta_n^2/\text{Vol}(\Omega)^2.$$

The inequality $\|g\|_n^2/\|g\|^2_{L_2(\Omega)} < \eta_1$ implies

$$\|g_L\|_n^2 - \|g_L\|^2_{L_2(\Omega)}/\text{Vol}(\Omega)$$

$$\leq 2\eta_1 \|g\|^2_{L_2(\Omega)}/C - \|g_L\|_2^2/\text{Vol}(\Omega) + 2C\delta_n^2/\text{Vol}(\Omega)^2$$

$$\leq 2\eta_1(s+1)^2\delta_n^2/C - (s-1)^2\delta_n^2/\text{Vol}(\Omega) + 2C\delta_n^2/\text{Vol}(\Omega)^2$$
$$\leq 2\eta_1(s+1)^2\delta_n^2/C - (s-1)^2\delta_n^2/\text{Vol}(\Omega) + 2C\delta_n^2/\text{Vol}(\Omega)^2.$$

By choosing appropriate C and η_1 (the choice only depends on $\text{Vol}(\Omega)$), we obtain

$$\|g_L\|_n^2 - \|g_L\|_{L_2(\Omega)}^2/\text{Vol}(\Omega) \leq -\frac{1}{2}(s-1)^2\delta_n^2/\text{Vol}(\Omega). \qquad \text{(H.1)}$$

Note that

$$\left|\|g_L\|_n^2 - \|g_L\|_{L_2(\Omega)}^2/\text{Vol}(\Omega)\right| \leq K^2 \qquad \text{(H.2)}$$

and

$$\mathbb{E}(g_L^2 - \|g_L\|_{L_2(\Omega)}^2/\text{Vol}(\Omega))^2 \leq 4K^2\|g_L\|_{L_2(\Omega)}^2/\text{Vol}(\Omega) \leq 4K^2(s+2)^2\delta_n^2/\text{Vol}(\Omega). \qquad \text{(H.3)}$$

Combining (H.1), (H.2) and (H.3) and Lemma F.4, we have

$$\mathbb{P}\left(\|g_L\|_{L_2(\Omega)}^2/\text{Vol}(\Omega) - \|g_L\|_n^2 \geq \frac{1}{2\text{Vol}(\Omega)}(s-1)^2\delta_n^2\right)$$

$$\leq \exp\left[-\frac{n\frac{1}{8\text{Vol}(\Omega)^2}(s-1)^4\delta_n^4}{4K^2(s+2)^2\delta_n^2/\text{Vol}(\Omega) + K^2\frac{1}{6\text{Vol}(\Omega)}(s-1)^2\delta_n^2}\right]$$

$$\leq \exp\left[-\frac{n\frac{1}{8\text{Vol}(\Omega)}(s-1)^4\delta_n^4}{4K^2(s+2)^2\delta_n^2 + K^2\frac{1}{6}(s-1)^2\delta_n^2}\right]$$

$$\leq \exp\left[-\frac{1}{8\text{Vol}(\Omega)}\frac{n(s-1)^2\delta_n^2}{36K^2 + K^2\frac{1}{6}}\right]$$

$$\leq \exp\left[-\frac{1}{8\text{Vol}(\Omega)}\frac{n(s-1)^2\delta_n^2}{37K^2}\right]$$

$$\leq \exp\left[-\frac{n(s-1)^2\delta_n^2}{296\text{Vol}(\Omega)K^2}\right]. \qquad \text{(H.4)}$$

Therefore, taking all $g \in \mathcal{G}$ yields

$$\mathbb{P}\left(\inf_{\|g\|_{L_2(\Omega)}\geq 2\delta_n}\frac{\|g\|_n^2}{\|g\|_{L_2(\Omega)}^2} < \eta_1\right)$$

$$\leq \sum_{s=2}^{\infty}\exp\left[H_B(\delta_n/\text{Vol}(\Omega), \mathcal{G}', \|\cdot\|_{L_\infty(\Omega)}) - \frac{n(s-1)^2\delta_n^2}{300\text{Vol}(\Omega)K^2}\right].$$

Since $H_B(\delta_n/\text{Vol}(\Omega), \mathcal{G}', \|\cdot\|_{L_\infty(\Omega)}) \leq \frac{n\delta_n^2}{1200\text{Vol}(\Omega)K^2}$, it can be seen that

$$\mathbb{P}\left(\inf_{\|g\|_{L_2(\Omega)}\geq 2\delta_n}\frac{\|g\|_n^2}{\|g\|_{L_2(\Omega)}^2} < \eta_1\right) \leq \sum_{s=2}^{\infty}\exp\left[-\frac{n(s-1)^2\delta_n^2}{1200\text{Vol}(\Omega)K^2}\right]$$

$$\leq C_1 \exp(-C_2 n \delta_n^2 / K^2)$$

for some constants C_1 and C_2 only related to $\mathrm{Vol}(\Omega)$, which finishes the proof of the first part.

For $C_0 \leq \frac{1}{4\mathrm{Vol}(\Omega)}$, it can be verified that

$$g^2 \leq 2g_R^2/C_0 + 2C_0(g - g_R)^2 \leq 2g_R^2/C_0 + 2C_0\delta_n^2/\mathrm{Vol}(\Omega)^2,$$

which yields

$$\|g\|_n^2 \leq 2\|g_R\|_n^2/C_0 + 2C_0\delta_n^2/\mathrm{Vol}(\Omega)^2.$$

The inequality $\|g\|_n^2/\|g\|_{L_2(\Omega)}^2 > \eta_2$ implies

$$\|g_R\|_n^2 - \|g_R\|_{L_2(\Omega)}^2/\mathrm{Vol}(\Omega) \geq \frac{1}{2}\eta_2 C_0 s^2 \delta_n^2 - \|g_R\|_{L_2(\Omega)}^2/\mathrm{Vol}(\Omega) - C_4^2 \delta_n^2/\mathrm{Vol}(\Omega)^2$$

$$\geq \frac{1}{2}\eta_2 C_0 s^2 \delta_n^2 - (s-1)^2\delta_n^2/\mathrm{Vol}(\Omega) - C_0^2 \delta_n^2/\mathrm{Vol}(\Omega)^2.$$

By choosing appropriate C_0 and η_2, we have

$$\|g_R\|_n^2 - \|g_R\|_{L_2(\Omega)}^2/\mathrm{Vol}(\Omega) \geq \frac{1}{4}(s-1)^2\delta_n^2/\mathrm{Vol}(\Omega). \tag{H.5}$$

Note that

$$\left| \|g_R\|_n^2 - \|g_R\|_{L_2(\Omega)}^2/\mathrm{Vol}(\Omega) \right| \leq K^2 \tag{H.6}$$

and

$$\mathbb{E}(g_R^2 - \|g_R\|_{L_2(\Omega)}^2/\mathrm{Vol}(\Omega))^2 \leq 4K^2\|g_R\|_{L_2(\Omega)}^2/\mathrm{Vol}(\Omega) \leq 4K^2(s+2)^2\delta_n^2/\mathrm{Vol}(\Omega). \tag{H.7}$$

By combining (H.5), (H.6) and (H.7) and Lemma F.4, similar to (H.4), we obtain

$$\mathbb{P}\left(\frac{\|g_R\|_{L_2(\Omega)}^2}{\mathrm{Vol}(\Omega)} - \|g_R\|_n^2 \geq \frac{1}{4\mathrm{Vol}(\Omega)}(s-1)^2\delta_n^2 \right) \leq \exp\left[-\frac{n(s-1)^2\delta_n^2}{600\mathrm{Vol}(\Omega)K^2} \right].$$

Taking all $g \in \mathcal{G}$ leads to

$$\mathbb{P}\left(\inf_{\|g\|_{L_2(\Omega)} \geq 2\delta_n} \frac{\|g\|_n^2}{\|g\|_{L_2(\Omega)}^2} < \eta_2 \right)$$

$$\leq \sum_{s=2}^{\infty} \exp\left[H_B(\delta_n/\mathrm{Vol}(\Omega), \mathcal{G}', \|\cdot\|_{L_\infty(\Omega)}) - \frac{n(s-1)^2\delta_n^2}{600\mathrm{Vol}(\Omega)K^2} \right].$$

Since $H_B(\delta_n/\mathrm{Vol}(\Omega), \mathcal{G}', \|\cdot\|_{L_\infty(\Omega)}) \leq \frac{n\delta_n^2}{1200\mathrm{Vol}(\Omega)K^2}$, we have

$$\mathbb{P}\left(\inf_{\|g\|_{L_2(\Omega)} \geq 2\delta_n} \frac{\|g\|_n^2}{\|g\|_{L_2(\Omega)}^2} < \eta_2 \right) \leq \sum_{s=2}^{\infty} \exp\left[-\frac{n(s-1)^2\delta_n^2}{1200\mathrm{Vol}(\Omega)K^2} \right]$$

$$\leq C_3 \exp(-C_4\delta_n^2/K^2)$$

for some constants C_3 and C_4 related to $\mathrm{Vol}(\Omega)$, which finishes the proof of the second part.

Appendix I: Proof of Lemma F.6

Notice that

$$\frac{\mu}{n} Y^T (R + \mu I)^{-1} Y = \min_{\hat{Y}} \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \frac{\mu}{n} \hat{Y}^T R^{-1} \hat{Y}, \qquad \text{(I.1)}$$

which can be verified by taking minimization of the objective function inside the right-hand side of (I.1). Let $\hat{u} = R^{-1} \hat{Y}$. By plugging $\hat{u}$ into the right-hand side of (I.1), we have

$$\min_{\hat{Y}} \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \frac{\mu}{n} \hat{Y}^T R^{-1} \hat{Y} = \min_{\hat{u}} \frac{1}{n} (Y - R\hat{u})^T (Y - R\hat{u}) + \frac{\mu}{n} \hat{u}^T R \hat{u}. \tag{I.2}$$

Therefore, by the representer theorem and (2.1), the right-hand side of (I.2) is the same as

$$\min_{\hat{f} \in \mathcal{N}_\Psi(\Omega)} \left(\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \frac{\mu}{n} \|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)} \right). \tag{I.3}$$

Notice the objective function in (I.3) can be written as

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \frac{\mu}{n} \|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)}$$

$$= \frac{1}{n} \sum_{i=1}^n (f(x_i) - \hat{f}(x_i))^2 + \frac{\mu}{n} \|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)} + \frac{2}{n} \sum_{i=1}^n \epsilon_i (f(x_i) - \hat{f}(x_i)) + \frac{1}{n} \sum_{i=1}^n \epsilon_i^2. \tag{I.4}$$

If $\mu^{-1} = O_P(n^{-\frac{d}{2\nu+d}})$, by Lemma F.2 and the proof of Lemma F.2, it can be shown that (I.4) converges to σ_ϵ^2.

If $\mu = O_P(n^{\frac{d}{2\nu+d}})$, then for any $\hat{f}$, we have

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \frac{\mu}{n} \|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)} \leq \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + C n^{-\frac{2\nu}{2\nu+d}} \|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)}.$$

Applying the results in the case of $\mu^{-1} = O_P(n^{-\frac{d}{2\nu+d}})$, we obtain

$$\frac{\mu}{n} Y^T (R + \mu I)^{-1} Y \leq 2\sigma_\epsilon^2,$$

with probability at least $1 - C_1 \exp(-C_2 n^\eta)$. The lower bound can be obtained by

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \frac{\mu}{n} \|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \hat{f}(x_i))^2 + \frac{\mu}{n}\|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)} + \frac{2}{n}\sum_{i=1}^{n}\epsilon_i(f(x_i) - \hat{f}(x_i)) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2$$

$$\geq -\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \hat{f}(x_i))^2 + \frac{\mu}{n}\|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)} + \frac{1}{2n}\sum_{i=1}^{n}\epsilon_i^2$$

$$\geq -\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \hat{f}(x_i))^2 - \frac{\mu}{n}\|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)} + \frac{1}{2n}\sum_{i=1}^{n}\epsilon_i^2 \geq \sigma_\epsilon^2/4,$$

with probability tending to one, where the first inequality is because of the Cauchy-Schwarz inequality. The last inequality is true because $\|f - \hat{f}\|_n^2$ and $\frac{\mu}{n}\|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)}$ converge to zero [58, 24]. This completes the proof.

Appendix J: Proof of Lemma F.7

We only present the proof of the first inequality. The second inequality can be proved similarly. By direct calculation, it can be shown that

$$\text{tr}((A + B)(A + B + C)^{-1}) \geq \text{tr}(A(A + C)^{-1})$$
$$\Leftrightarrow \text{tr}(C(A + B + C)^{-1}) \leq \text{tr}(C(A + C)^{-1})$$
$$\Leftrightarrow \text{tr}(C(A + B + C)^{-1}B(A + C)^{-1}) \geq 0,$$

which is true since A, B and C are positive definite.

Appendix K: Proof of Lemma F.8

Notice that at point x, for any $u = (u_1, ..., u_n)^T \in \mathbb{R}^n$,

$$(f(x) - \sum_{j=1}^{n}u_j f(x_j))^2$$

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

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

$$\leq \left(\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left|\sum_{j=1}^{n}u_j e^{-ix_j^T\omega} - e^{-ix^T\omega}\right|^2\mathcal{F}(\Psi)(\omega)d\omega + \hat{\mu}_n\|u\|_2^2\right)\|f\|^2_{\mathcal{N}(\Omega)}$$

$$= \left(1 - 2\sum_{i=1}^{n}u_i\Psi(x - x_i) + \sum_{i=1}^{n}\sum_{j=1}^{n}u_i u_j\Psi(x_i - x_j) + \hat{\mu}_n\|u\|_2^2\right)\|f\|^2_{\mathcal{N}(\Omega)},$$

where the first inequality is because of the Cauchy-Schwarz inequality. Plugging $u = (R + \hat{\mu}_n I_n)^{-1}r(x)$ finishes the proof of the first part.

Now we prove the second part of this lemma.

Consider function $g(t) = \Psi(x - t)$. By the interpolation inequality, we have

$$1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x) \le \|g(t) - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(t)\|_{L_\infty(\Omega)}$$

$$\le C_1 \|g(t) - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(t)\|_{L_2(\Omega)}^{1 - \frac{d}{2\nu}}$$

$$\times \|g(t) - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(t)\|_{\mathcal{N}_\Psi(\Omega)}^{\frac{d}{2\nu}}. \tag{K.1}$$

By direct calculation, it can be seen that

$$\|g(t) - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(t)\|_{\mathcal{N}_\Psi(\Omega)}^2 \le 1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x). \tag{K.2}$$

Combining (K.1) and (K.2) leads to

$$1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x) \le C_2 \|g(t) - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(t)\|_{L_2(\Omega)}^{\frac{4\nu - 2d}{4\nu - d}}. \tag{K.3}$$

Let $f_1(t) = r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(t)$. It can be seen by the representer theorem that

$$f_1 = \operatorname*{argmin}_{h \in \mathcal{N}_\Psi(\Omega)} \|g - h\|_n^2 + \frac{\hat{\mu}_n}{n} \|h\|_{\mathcal{N}_\Psi(\Omega)}^2.$$

Take $\delta_n = n^{-\frac{4\nu - d}{8\nu}}$. Direct calculation shows that if $\alpha < \frac{1}{2}$, either $1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x) \lesssim n^{(\alpha-1)(1-\frac{d}{2\nu})}$, or the conditions of Lemma F.5 hold. If $\alpha \ge \frac{1}{2}$, then either $\|g - f_1\|_2^2 \ge C_3 n^{2(\alpha-1)}$ for some constant C_3, which implies the conditions of Lemma F.5 hold, or

$$1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x) \le C_2 \|g(t) - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(t)\|_2^{\frac{4\nu - 2d}{4\nu - d}}$$

$$\le C_4 n^{(\alpha-1)\frac{4\nu - 2d}{4\nu - d}} \lesssim n^{(\alpha-1)(1-\frac{d}{2\nu})}.$$

Thus, combining these two cases, either the conditions of Lemma F.5 hold, or $1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x) \lesssim n^{(\alpha-1)(1-\frac{d}{2\nu})}$. If the conditions of Lemma F.5 holds, then with probability at least $1 - C_7 \exp(-C_8 n^{\eta_1})$,

$$\|g - f_1\|_{L_2(\Omega)}^2$$

$$\le \eta_2 \|g - f_1\|_n^2$$

$$= \eta_2 (\|g - f_1\|_n^2 + \frac{\hat{\mu}_n}{n} \|f_1\|_{\mathcal{N}_\Psi(\Omega)}^2 - \frac{\hat{\mu}_n}{n} \|f_1\|_{\mathcal{N}_\Psi(\Omega)}^2)$$

$$\le \eta_2 \frac{\hat{\mu}_n}{n} (1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} R (R + \hat{\mu}_n I_n)^{-1} r(x))$$

$$= \eta_2 \frac{\hat{\mu}_n}{n} (1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x)$$

$$\quad + r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x) - r(x)^T (R + \hat{\mu}_n I_n)^{-1} R (R + \hat{\mu}_n I_n)^{-1} r(x))$$

$$= \eta_2 \frac{\hat{\mu}_n}{n} (1 - r(x)^T (R + \hat{\mu}_n I_n)^{-1} r(x) + \hat{\mu}_n r(x)^T (R + \hat{\mu}_n I_n)^{-2} r(x))$$

$$\leq 2\eta_2 \frac{\hat{\mu}_n}{n}(1 - r(x)^T(R + \hat{\mu}_n I_n)^{-1}r(x)), \tag{K.4}$$

where the last inequality is by (F.25). Plugging (K.4) into (K.3) yields

$$1 - r(x)^T(R + \hat{\mu}_n I_n)^{-1}r(x) \lesssim n^{(\alpha-1)(1-\frac{d}{2\nu})},$$

which finishes the proof.

Appendix L: Properties of eigenvalues

Lemma L.1 states the asymptotic rate of the eigenvalues of $\Psi(\cdot - \cdot)$. Lemma L.2 states the minimum eigenvalue of $\Psi_1^T \Psi_1$, where Ψ_1 will be defined later.

Since $\Psi(\cdot - \cdot)$ is a positive definite function, by Mercer's theorem, there exists a countable set of positive eigenvalues $\lambda_1 \geq \lambda_2 \geq ... > 0$ and an orthonormal basis for $L_2(\Omega)$ $\{\varphi_k\}_{k \in \mathbb{N}}$ such that

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

where the summation is uniformly and absolutely convergent.

Lemma L.1. *Let λ_k be as in (L.1). Then, $\lambda_k \asymp k^{-2\nu/d}$.*

Proof. 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 [64],

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

By Lemma C.3, $H^\nu(\Omega)$ coincide with $\mathcal{N}_\Psi(\Omega)$. By Theorem 5.7 in [17], T and T^* have the same singular values. By Theorem 5.10 in [17], 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 Theorem in Section 3.3.4 in [18], the embedding operator T has approximation numbers satisfying

$$C_3 k^{-\nu/d} \leq a_k \leq C_4 k^{-\nu/d}, \forall k \in \mathbb{N}, \tag{L.2}$$

where C_3 and C_4 are two positive numbers. By Theorem 5.7 in [17], $T^*T\varphi_k = \mu_k^2 \varphi_k$, and $T^*T\varphi_k = T^*\varphi_k = \lambda_k \varphi_k$, we have $\lambda_k = \mu_k^2$. By (L.2), $\lambda_k \asymp k^{-2\nu/d}$ holds. $\qquad\square$

Lemma L.2. *Define two matrices $\Psi_1 = \frac{1}{\sqrt{n}}(\varphi_1(X), ..., \varphi_{p_1}(X))$, and $\Psi_2 = \frac{1}{\sqrt{n}}(\varphi_{p_1+1}(X), \varphi_{p_1+2}(X), ...)$, where φ_k's are eigenfunctions as in (L.1) and $\varphi_k(X) = (\varphi_k(x_1), ..., \varphi_k(x_n))^T$ for $k = 1, 2, ...$. With probability at least $1 - C_1 e^{-C_2 n^{\eta_1}}$,*

$$\lambda_{\min}(\Psi_1^T \Psi_1) \geq \eta,$$

for any $\mu = O(\sqrt{n})$, where $p = \lfloor (n/\mu)^{d/(2\nu)} \rfloor$, $p_1 = \min\{p, C_3 n^{1/2}\}$, η, η_1, C_1, C_2 are positive constants.

Proof. Consider $u^T \Psi_1^T \Psi_1 u$, where $u = (u_1, ..., u_p)^T \in \mathbb{R}^p$ with $\|u\|_2 = 1$. Let $\mathcal{Q} = \{g : g = \sum_{i=1}^p u_i \varphi_i\}$. Since φ_i's are orthonormal, $\|g\|_{L_2(\Omega)} = 1$. For any $g \in \mathcal{Q}$, by Lemma L.1, $\|g\|_{H^\nu(\Omega)}^2 \le \frac{C_1}{\lambda_{p_1}} \le C_2 p_1^{2\nu/d}$. By the interpolation inequality,

$$\|g\|_{L_\infty(\Omega)} \le C_3 \|g\|_{H^\nu(\Omega)}^{\frac{d}{2\nu}} = C_4 p_1^{1/2}.$$

We shall use Lemma F.5 to link $\|g\|_n$ to $\|g\|_{L_2(\Omega)}$. First we need to check the conditions of Lemma F.5 hold. Since $\|g\|_{L_2(\Omega)} = 1$, it suffices to check the entropy condition. Let $\rho = C_2^{1/2} p_1^{\nu/d}$. Consider class $\mathcal{Q}' = \{g : g = \frac{f}{\rho}, f \in \mathcal{Q}\}$. Since $\mathcal{Q}' \subset H^\nu(\Omega)$, there exists a constant C_5 such that

$$H_B(\delta_n/\text{Vol}(\Omega), \mathcal{F}', \| \cdot \|_{L_\infty(\Omega)}) \le C_5 \left(\frac{1}{\delta_n}\right)^{d/\nu}.$$

The entropy condition is satisfied if

$$n\delta_n^{2+d/\nu}/K^2 > C_6, \tag{L.3}$$

where $\delta_n = 1/\rho$, $K \le C_4 p_1^{1/2}/\rho$, and C_6 is some constant depending on C_1–C_5 and Ω. By direct calculations, if $p_1 \le C_7\sqrt{n}$ for some constant C_7, (L.3) is satisfied.

By Lemma F.5,

$$u^T \Psi_1^T \Psi_1 u = \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p u_j \varphi_j(X_i)\right)^2 = \|g\|_n^2 \ge \eta, \tag{L.4}$$

with probability at least $1 - C_8 \exp(-C_9 n^{\eta_1})$ for some constant η and η_1. This finishes the proof. □

Acknowledgments

References

[1] ADAMS, R. A. and FOURNIER, J. J. (2003). *Sobolev Spaces*. Academic press. MR2424078

[2] ANDERES, E. et al. (2010). On the consistent separation of scale and variance for Gaussian random fields. *The Annals of Statistics* **38** 870–893. MR2604700

[3] ANKENMAN, B., NELSON, B. L. and STAUM, J. (2010). Stochastic kriging for simulation metamodeling. *Operations Research* **58** 371–382. MR2674803

[4] Ba, S., Joseph, V. R. et al. (2012). Composite Gaussian process models for emulating expensive functions. *The Annals of Applied Statistics* **6** 1838–1860. MR3058685

[5] Bachoc, F. (2013). Parametric estimation of covariance function in Gaussian-process based Kriging models. Application to uncertainty quantification for computer experiments, PhD thesis.

[6] Bachoc, F. (2013). Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics & Data Analysis* **66** 55–69. MR3064023

[7] Bachoc, F. et al. (2018). Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case. *Bernoulli* **24** 1531–1575. MR3706801

[8] Baker, E., Barbillon, P., Fadikar, A., Gramacy, R. B., Herbei, R., Higdon, D., Huang, J., Johnson, L. R., Mondal, A., Pires, B. et al. (2020). Stochastic Simulators: An Overview with Opportunities. *arXiv preprint arXiv:2002.01321*.

[9] Bevilacqua, M., Faouzi, T., Furrer, R., Porcu, E. et al. (2019). Estimation and prediction using generalized Wendland covariance functions under fixed domain asymptotics. *The Annals of Statistics* **47** 828–856. MR3909952

[10] Bostanabad, R., Liang, B., Gao, J., Liu, W. K., Cao, J., Zeng, D., Su, X., Xu, H., Li, Y. and Chen, W. (2018). Uncertainty quantification in multiscale simulation of woven fiber composites. *Computer Methods in Applied Mechanics and Engineering* **338** 506–532. MR3811636

[11] Bower, R. G., Vernon, I., Goldstein, M., Benson, A., Lacey, C. G., Baugh, C. M., Cole, S. and Frenk, C. (2010). The parameter space of galaxy formation. *Monthly Notices of the Royal Astronomical Society* **407** 2017–2045.

[12] Burchell, A. N., Richardson, H., Mahmud, S. M., Trottier, H., Tellier, P. P., Hanley, J., Coutlée, F. and Franco, E. L. (2006). Modeling the sexual transmissibility of human papillomavirus infection using stochastic computer simulation and empirical data from a cohort study of young women in Montreal, Canada. *American journal of epidemiology* **163** 534–543.

[13] Chernih, A. and Hubbert, S. (2014). Closed form representations and properties of the generalised Wendland functions. *Journal of Approximation Theory* **177** 17–33. MR3128401

[14] Dancik, G. (2007). mlegp: An R package for Gaussian process modeling and sensitivity analysis. *CRAN. https://cran.r-project.org/web/packages/mlegp/mlegp.pdf.*

[15] DeVore, R. A. and Sharpley, R. C. (1993). Besov spaces on domains in R^d. *Transactions of the American Mathematical Society* **335** 843–864. MR1152321

[16] Driscoll, M. F. (1973). The reproducing kernel Hilbert space structure of the sample paths of a Gaussian process. *Probability Theory and Related Fields* **26** 309–316. MR0370723

[17] EDMUNDS, D. E. and EVANS, W. D. (2018). *Spectral Theory and Differential Operators.* Oxford University Press. MR3823299

[18] EDMUNDS, D. E. and TRIEBEL, H. (2008). *Function Spaces, Entropy Numbers, Differential Operators.* Cambridge University Press. MR1410258

[19] GIHMAN, I. I. and SKOROKHOD, A. V. (1974). *The Theory of Stochastic Processes I.* Springer. MR2304859

[20] GIROSI, F., JONES, M. and POGGIO, T. (1995). Regularization theory and neural networks architectures. *Neural Computation* **7** 219–269.

[21] GNEITING, T. (2002). Stationary covariance functions for space-time data. *Journal of the American Statistical Association* **97** 590–600. MR1941475

[22] GRAMACY, R. B. and APLEY, D. W. (2015). Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics* **24** 561–578. MR3357395

[23] GRAMACY, R. B. and LEE, H. K. (2012). Cases for the nugget in modeling computer experiments. *Statistics and Computing* **22** 713–722. MR2909617

[24] GU, C. (2013). *Smoothing Spline ANOVA Models.* Springer Science & Business Media. MR3025869

[25] HIGDON, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* 37–56. Springer. MR2059819

[26] HOFMANN, S., MITREA, M. and TAYLOR, M. (2007). Geometric and transformational properties of Lipschitz domains, Semmes-Kenig-Toro domains, and other classes of finite perimeter domains. *The Journal of Geometric Analysis* **17** 593–647. MR2365661

[27] HUNG, Y. (2011). Penalized blind kriging in computer experiments. *Statistica Sinica* 1171–1190. MR2827519

[28] HUNG, Y., JOSEPH, V. R. and MELKOTE, S. N. (2015). Analysis of computer experiments with functional response. *Technometrics* **57** 35–44. MR3318347

[29] JOSEPH, V. R. (2006). Limit kriging. *Technometrics* **48** 458–466. MR2328615

[30] JOSEPH, V. R. and KANG, L. (2011). Regression-based inverse distance weighting with applications to computer experiments. *Technometrics* **53** 254–265. MR2857703

[31] KARVONEN, T., WYNNE, G., TRONARP, F., OATES, C. J. and SÄRKKÄ, S. (2020). Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions. *arXiv preprint arXiv:2001.10965.* MR4130422

[32] KAUFMAN, C. and SHABY, B. A. (2013). The role of the range parameter for estimation and prediction in geostatistics. *Biometrika* **100** 473–484. MR3068447

[33] KOU, S. (2003). On the efficiency of selection criteria in spline regression. *Probability Theory and Related Fields* **127** 153–176. MR2013979

[34] LEE, M. R. and OWEN, A. B. (2018). Single nugget kriging. *Statistica Sinica* 649–669. MR3791082

[35] LI, R. and SUDJIANTO, A. (2005). Analysis of computer experiments using

penalized likelihood in Gaussian Kriging models. *Technometrics* **47** 111–120. MR2188073

[36] Mak, S., Sung, C.-L., Wang, X., Yeh, S.-T., Chang, Y.-H., Joseph, V. R., Yang, V. and Wu, C. J. (2018). An efficient surrogate model for emulation and physics extraction of large eddy simulations. *Journal of the American Statistical Association* **113** 1443–1456. MR3902221

[37] Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71** 135–146. MR0738334

[38] Massart, P. (2007). *Concentration Inequalities and Model Selection* **6**. Springer. MR2319879

[39] Matheron, G. (2012). *Estimating and Choosing: An Essay on Probability in Practice.* Springer Science & Business Media. MR0979761

[40] Moran, A. E., Odden, M. C., Thanataveerat, A., Tzong, K. Y., Rasmussen, P. W., Guzman, D., Williams, L., Bibbins-Domingo, K., Coxson, P. G. and Goldman, L. (2015). Cost-effectiveness of hypertension therapy according to 2014 guidelines. *New England Journal of Medicine* **372** 447–455.

[41] Niculescu, C. and Persson, L.-E. (2006). *Convex Functions and Their Applications.* Springer. MR2178902

[42] Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods* **63**. SIAM. MR1172997

[43] Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association* **83** 1134–1143. MR0997592

[44] Peng, C.-Y. and Wu, C. J. (2014). On the choice of nugget in kriging modeling for deterministic computer experiments. *Journal of Computational and Graphical Statistics* **23** 151–168. MR3173765

[45] Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning.* MIT Press. MR2514435

[46] Santner, T. J., Williams, B. J. and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments.* Springer Science & Business Media. MR2160708

[47] Scheuerer, M., Schaback, R. and Schlather, M. (2013). Interpolation of spatial data–A stochastic or a deterministic problem? *European Journal of Applied Mathematics* **24** 601–629. MR3082868

[48] Sniekers, S. and van der Vaart, A. (2015). Credible sets in the fixed design model with Brownian motion prior. *Journal of Statistical Planning and Inference* **166** 78–86. MR3390135

[49] Stein, M. L. (1993). Spline smoothing with an estimated order parameter. *The Annals of Statistics* **21** 1522–1544. MR1241277

[50] Stein, M. L. (1995). Fixed-domain asymptotics for spatial periodograms. *Journal of the American Statistical Association* **90** 1277–1288. MR1379470

[51] Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* Springer Science & Business Media. MR1697409

[52] Steinwart, I. (2019). Convergence types and rates in generic Karhunen-Loève expansions with applications to sample path properties. *Potential*

Analysis **51** 361–395. MR4023468

[53] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics* 1040–1053. MR0673642

[54] Sung, C.-L., Wang, W., Plumlee, M. and Haaland, B. (2020). Multiresolution Functional ANOVA for Large-Scale, Many-Input Computer Experiments. *Journal of the American Statistical Association* **115** 908-919. MR4107688

[55] Tan, M. H. (2018). Gaussian process modeling of a functional output with information from boundary and initial conditions and analytical approximations. *Technometrics* **60** 209–221. MR3804249

[56] Tuo, R. and Wang, W. (2020). Kriging prediction with isotropic Matérn correlations: robustness and experimental designs. *Journal of Machine Learning Research* **21** 1–38. MR4209473

[57] Tuo, R. and Wu, C. F. J. (2016). A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification* **4** 767–795. MR3523087

[58] van de Geer, S. (2000). *Empirical Processes in M-estimation*. Cambridge University Press.

[59] van de Geer, S. (2014). On the uniform convergence of empirical norms and inner products, with application to causal inference. *Electronic Journal of Statistics* **8** 543–574. MR3211024

[60] Walck, C. (1996). Hand-book on statistical distributions for experimentalists Technical Report.

[61] Wang, D., Loh, W.-L. et al. (2011). On fixed-domain asymptotics and covariance tapering in Gaussian random field models. *Electronic Journal of Statistics* **5** 238–269. MR2792553

[62] Wang, W. and Haaland, B. (2019). Controlling sources of inaccuracy in stochastic kriging. *Technometrics* **61** 309–321. MR3987622

[63] Wang, W., Tuo, R. and Wu, C. F. J. (2020). On Prediction Properties of Kriging: Uniform Error Bounds and Robustness. *Journal of the American Statistical Association* **115** 920–930. MR4107689

[64] Wendland, H. (2004). *Scattered Data Approximation* **17**. Cambridge University Press. MR2131724

[65] Williams, B., Higdon, D., Gattiker, J., Moore, L., McKay, M., Keller-McNulty, S. et al. (2006). Combining experimental data and computer simulations, with an application to flyer plate experiments. *Bayesian Analysis* **1** 765–792. MR2282206

[66] Wu, Z. and Schaback, R. (1993). Local error estimates for radial basis function interpolation of scattered data. *IMA Journal of Numerical Analysis* **13** 13–27. MR1199027

[67] Xu, W. and Stein, M. L. (2017). Maximum likelihood estimation for a smooth Gaussian random field model. *SIAM/ASA Journal on Uncertainty Quantification* **5** 138–175. MR3601677

[68] Xu, W., Stein, M. L. and Wisher, I. (2019). Modeling and Predicting Chaotic Circuit Data. *SIAM/ASA Journal on Uncertainty Quantification*

7 31–52. MR3895127

[69] YAMAMOTO, J. K. (2000). An alternative measure of the reliability of ordinary kriging estimates. *Mathematical Geology* **32** 489–509. MR1751576

[70] YANG, Y., BHATTACHARYA, A. and PATI, D. (2017). Frequentist coverage and sup-norm convergence rate in Gaussian process regression. *arXiv preprint arXiv:1708.04753*.

[71] YANO, K. and KATO, K. (2018). On frequentist coverage errors of Bayesian credible sets in high dimensions. *arXiv preprint arXiv:1803.03450*. MR4036046

[72] YING, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis* **36** 280–296. MR1096671

[73] YING, Z. (1993). Maximum likelihood estimation of parameters under a spatial sampling scheme. *The Annals of Statistics* **21** 1567–1590. MR1241279

[74] YOO, W. W., GHOSAL, S. et al. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *The Annals of Statistics* **44** 1069–1102. MR3485954

[75] ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association* **99** 250–261. MR2054303