

**CONTRIBUTIONS TO THE NONPARAMETRIC METHODS FOR COMPUTER  
EXPERIMENTS**

A Dissertation  
Presented to  
The Academic Faculty

by

Wenjia Wang

In Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy in the  
School of Industrial and Systems Engineering

Georgia Institute of Technology

August 2018

Copyright © 2018 by Wenjia Wang

**CONTRIBUTIONS TO THE NONPARAMETRIC METHODS FOR COMPUTER  
EXPERIMENTS**

Approved by:

Dr. C. F. Jeff Wu, Advisor  
School of Industrial and Systems  
Engineering  
*Georgia Institute of Technology*

Dr. Benjamin Haaland, Advisor  
School of Medicine  
*University of Utah*

Dr. Rui Tuo, Advisor  
School of Industrial and Systems  
Engineering  
*Georgia Institute of Technology*

Dr. Roshan Joseph Vengazhiyil  
School of Industrial and Systems  
Engineering  
*Georgia Institute of Technology*

Dr. Jianjun Shi  
School of Industrial and Systems  
Engineering  
*Georgia Institute of Technology*

Dr. Matthew Plumlee  
Department of Industrial Engineer-  
ing and Management Sciences  
*Northwestern University*

Date Approved: May 3rd, 2018

*To my parents and my wife.*

## ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude to my advisors Prof. C. F. Jeff Wu, Prof. Benjamin Haaland and Prof. Rui Tuo for their continuous support and guidance of my Ph.D. study. They strongly influenced me by their patience, perceptive insights, solid mathematical skills, and positive attitude. I am very fortunate to be mentored by them, and really appreciate their help from different perspectives.

Besides my advisors, I would like to thank the rest of my thesis committee, Dr. Roshan Joseph Vengazhiyil, Dr. Jianjun Shi, and Dr. Matthew Plumlee, for taking their precious time attending my defense and providing their insightful comments on my thesis, my doctoral study, and my future career. I am also thankful to all the faculty and staff members in ISyE for their help and efforts. Special thanks to faculty members from operations research for teaching me to widen my vision in research.

I am indebted to my fellow students and friends accompanying me in the past years. Many thanks to Chih-Li Sung, Simon Mak, Yuanshuo Zhao, and Lixiang Lin for being an excellent research group to work together. Thank my officemate Shuang Li for discussion on various topics. I would also like to thank my academic brothers Qianyi Wang, Qiushi Chen, Zhihao Ding, Can Zhang, Fang Cao, Helin Zhu, Fan Ye, Chengliang Zhang, Xiaowei Yue, Shan Ba, Xinyu Min, and Linwei Xin, for their valuable suggestions on discussions in academics, life, and career. I am also fortunate to have friends Di Wu, Yu Cao, Yilun Chen, Junzhuo Chen, and Tianyi Liu who were entering Georgia Tech later than me, for the time we have spent together. I would like to extend my thanks to my friends outside Georgia Tech, including but not limit to Di Wang, Zhuokang Jia, Xueqian Lu, Yanjun Zhu, Zhuoqiang Jia, and Haohao Liao, for carrying me, playing together and helping me getting through the hard times during my Ph.D. study.

Last but not the least, I would like to express my sincerest gratitude to my family for their unconditional help and care through my entire life. I especially appreciate the supports

from my parents Yihe Wang and Yanhua Wang, and my wife Yilin Li. This dissertation is dedicated to them.

## TABLE OF CONTENTS

<b>Acknowledgments</b> . . . . .	v
<b>List of Tables</b> . . . . .	xi
<b>List of Figures</b> . . . . .	xii
<b>Chapter 1: Controlling Sources of Inaccuracy in Stochastic Kriging</b> . . . . .	1
1.1 Introduction . . . . .	1
1.2 Preliminaries . . . . .	3
1.2.1 Stochastic Kriging Model . . . . .	3
1.2.2 Sources of Inaccuracy . . . . .	6
1.3 Nominal Error . . . . .	8
1.3.1 Stationary Model . . . . .	10
1.3.2 Non-Stationary Model . . . . .	11
1.3.3 Regression Functions . . . . .	13
1.3.4 Example Designs . . . . .	13
1.4 Numeric Error . . . . .	15
1.5 Parameter Estimation Error . . . . .	18
1.5.1 Example Design . . . . .	22
1.6 Parameter Estimation Numeric Error . . . . .	22

1.7	Numeric Examples . . . . .	24
1.7.1	Constant ratio of noise and process variance . . . . .	24
1.7.2	Input varying ratio of noise and process variance . . . . .	26
1.8	Discussion . . . . .	28
<b>Chapter 2: On Prediction Properties of Kriging: Uniform Error Bounds and Robustness . . . . .</b>		<b>31</b>
2.1	Introduction . . . . .	31
2.2	Review on the simple kriging method . . . . .	32
2.2.1	Goal of this work . . . . .	34
2.3	Kriging interpolant . . . . .	36
2.4	Power function and its upper bounds . . . . .	37
2.5	Uniform error bounds for kriging . . . . .	39
2.6	Simulation studies . . . . .	42
2.7	Conclusions and Discussion . . . . .	45
<b>Chapter 3: Smoothness Estimation and Adaptive Kernel Ridge Regression . . . . .</b>		<b>48</b>
3.1	Introduction . . . . .	48
3.2	Methodology . . . . .	49
3.2.1	Problem Setting . . . . .	50
3.2.2	Proposed Method . . . . .	50
3.2.3	Some Intuition and Related Methods . . . . .	52
3.3	Theoretical Results . . . . .	54
3.3.1	Mathematical Formulation of Smoothness . . . . .	54

3.3.2	Main Theorems . . . . .	56
3.3.3	Comparison with Existing Results . . . . .	60
3.4	Proofs . . . . .	61
3.4.1	Proof of Theorem 3.3.1 . . . . .	62
3.4.2	Proof of Theorem 3.3.2 . . . . .	62
<b>Appendix A: Appendix of Chapter 1 . . . . .</b>		<b>66</b>
A.1	Proof of Theorem 1.3.1 . . . . .	66
A.2	Assumptions for Theorem 1.4.1 . . . . .	68
A.3	Proof of Theorem 1.5.1 . . . . .	69
A.4	Proof of Proposition 1.5.1 . . . . .	74
A.5	Proof of Theorem 1.6.1 . . . . .	80
A.6	Proof of Lemma A.5.2 . . . . .	83
<b>Appendix B: Appendix of Chapter 2 . . . . .</b>		<b>87</b>
B.1	Auxiliary tools . . . . .	87
B.1.1	Reproducing kernel Hilbert spaces . . . . .	87
B.1.2	A Maximum inequality for Gaussian processes . . . . .	88
B.2	Proof of Theorem 2.5.1 . . . . .	89
<b>Appendix C: Appendix of Chapter 3 . . . . .</b>		<b>96</b>
C.1	Upper and Lower Bounds of the modified likelihood function . . . . .	96
C.2	Proof of Lemma C.1.1 . . . . .	100
C.3	Proof of Lemma C.1.2 . . . . .	106

C.4	Proof of Lemma C.1.3 . . . . .	107
C.5	Proof of Lemma C.2.2 . . . . .	109
C.6	Proof of Lemma C.3.1 . . . . .	119
C.7	Proof of Lemma 3.3.1 . . . . .	122
C.8	Proof of Lemma C.5.3 . . . . .	124
C.9	Proof of Lemma C.5.4 . . . . .	128
C.10	Proof of Lemma C.5.2 . . . . .	128
C.11	Proof of Lemma C.5.4 . . . . .	133
C.12	Proof of Lemma C.6.1 . . . . .	134
C.13	Asymptotic bounds of the determinant term . . . . .	135
	C.13.1 Properties of eigenvalues . . . . .	135
	C.13.2 Lower bound of the determinant term . . . . .	136
	C.13.3 Upper bound of the determinant term . . . . .	138
<b>References</b>	. . . . .	<b>149</b>

## LIST OF TABLES

1.1	Average maximum squared prediction error for a spectrum of experimental designs across numbers of replications. . . . .	25
1.2	The average maximum prediction error under the best choice of replications to the average maximum prediction error under the number of replications suggested by the nominal bound. . . . .	26
1.3	Average maximum squared prediction error comparisons across number of distinct input locations and input varying replication vs. constant replication.	28
2.1	Numerical studies on the convergence rates of kriging prediction. . . . .	43

## LIST OF FIGURES

1.1	Nominal error designs under different settings. . . . .	15
1.2	Contour of $\frac{\partial \sigma_{\tau}^2(x)}{\partial \tau}$ and parameter estimation error design. . . . .	22
1.3	Comparisons of the nominal error bound to the average maximum squared prediction error. . . . .	27
2.1	The regression line of $\log \sup_{\mathbf{x} \in \Omega} \epsilon(\mathbf{x})$ on $\log h_{\mathbf{X}}$ . . . . .	44
3.1	The plot of function $g(m)$ , where $d = 0.8m_0$ . . . . .	59

## SUMMARY

Kriging, or Gaussian process modeling, is widely used in estimating unknown functions based on the (noisy) evaluations. Originally, kriging was introduced in geostatistics by [1] and has seen revived interest (and many new results) in the areas of spatial statistics [2, 3], computer experiments [4, 5] and machine learning [6, 7].

The main idea of kriging is to assume the underlying function is a realization of a Gaussian random field. The accuracy of kriging, or more generally, nonparametric regression, depends very strongly on the manner in which data is collected [8, 9, 10] and the properties of the underlying function, especially the smoothness of the underlying function. This dissertation addresses three important problems related to: (i) What type of data collection might be expected to enable one to build an accurate model; (ii) Based on a high-quality design, what is the accuracy of the model; and (iii) Can we construct estimators that achieve the optimal convergence rate without knowing the true smoothness in advance.

In Chapter 1 we consider the first problem: What type of data collection might be expected to enable one to build an accurate model. This problem is known as *computer experimental design* in the field of computer experiments. In many situations actual physical experimentation is difficult or impossible, so scientists and engineers use simulations, or *computer experiments*, to study a system of interest. Many simulations are stochastic in the sense that repeated runs with the same input configuration will result in different outputs. For expensive or time-consuming simulations, stochastic kriging [11] is commonly used to generate predictions for simulation model outputs subject to uncertainty due to both function approximation and stochastic variation. In this chapter, we develop and justify a few guidelines for experimental design, which ensure accuracy of stochastic kriging emulators. We decompose error in stochastic kriging predictions into nominal, numeric, parameter estimation and parameter estimation numeric components and provide means to control each in terms of properties of the underlying experimental design. The design

properties implied for each source of error are weakly conflicting and broad principles are proposed. In brief, the space-filling properties “small fill distance” and “large separation distance” should balance with replication at distinct input configurations, with number of replications depending on the relative magnitudes of stochastic and process variability. Non-stationarity implies higher input density in more active regions, while regression functions imply a balance with traditional design properties. A few examples are presented to illustrate the results.

In Chapter 2 we derive error bounds of the (simple) kriging predictor under a uniform metric. The kriging method has pointwise predictive distributions which are computationally simple. However, in many applications one would like to predict for a range of untried points simultaneously. In this chapter we introduce some error bounds for the (simple) kriging predictor under the uniform metric. The predictive error is bounded in terms of the maximum pointwise predictive variance of kriging, which can be further bounded with the fill distance of the design set. It works for a scattered set of input points in an arbitrary dimension, and also covers the case where the covariance function of the Gaussian process is misspecified. These results lead to a better understanding of the rate of convergence of kriging under the Gaussian or the Matérn correlation functions, the relationship between space-filling designs and the accuracy of kriging models, and the robustness of the Matérn correlation functions.

In Chapter 3 we consider identifying the smoothness of an underlying function, by employing maximum likelihood estimation for the Gaussian process model. The function estimator based on the smoothness estimator is also constructed in this chapter. This maximum likelihood approach is widely used in estimating the smoothness parameter in practice, but theoretical studies are lacking. We propose a modified maximum likelihood method to estimate the underlying function as well as its smoothness based on noisy evaluations. We prove the consistency of the proposed smoothness estimator and that the function estimator achieves a nearly optimal rate of convergence for all degrees of smoothness.

## CHAPTER 1

### CONTROLLING SOURCES OF INACCURACY IN STOCHASTIC KRIGING

#### 1.1 Introduction

In many situations actual physical experimentation is difficult or impossible, so scientists and engineers use simulations, or *computer experiments*, to study a system of interest. For example, [12] study a complex simulation model for turbulent flows in swirl injectors, which are used in a spectrum of propulsion and power-generation applications, under a range of geometric conditions, [13] estimate sexual transmissibility of human papillomavirus infection via a stochastic simulation model, and [14] use the Cardiovascular Disease Policy Model to project cost-effectiveness of treating hypertension in the U.S. according to 2014 guidelines. Commonly, these simulations require a cascade of complex calculations and simulator runs are *expensive* relative to their information content. To enable exploration of the relationship between inputs and outputs in the system of interest, a typical and apparently high-quality solution is to collect data at several input configurations, then build an inexpensive approximation, or *emulator*, for the simulation.

In many cases, the data collected from the computer simulation is *stochastic* in the sense that repeated runs with the same input configuration will have different outputs, driven primarily by elements of the simulation model which are inherently stochastic. Consider for example, the Coronary Heart Disease Policy Model, which is the simulation backbone underlying the cost-effectiveness study in [14]. For each subject in a large cohort (the U.S. adult population), this model generates a simulated Markov trajectory through risk and event categories. These trajectories involve, for each subject and time-increment, randomly assigning a new state according to a specified distribution. Even if all the simulation settings, what we are calling inputs here, are unchanged, a new run of the simulation model

will have slightly different random trajectories, and in turn slightly different outputs. For emulation of stochastic computer experiments, the stochastic kriging model proposed in [11] has gained considerable traction as a quality approximation in a broad spectrum of real applications. In the stochastic kriging model, output associated with each input is decomposed as the sum of a mean (Gaussian process) output and random (Gaussian) noise.

The accuracy of the stochastic kriging emulator depends strongly on how the data is collected [8, 9, 10]. Notably, [11] provides a few useful results relating to mean squared prediction error (MSPE) *integrated over the design space* indicating that the distinct data sites should be relatively space-filling, while the number of replications is driven by the relative magnitudes of process and stochastic variability. Unfortunately, these results are limited to stationary process covariance with no non-trivial regression functions in the process mean. Further, no explicit consideration is given to very important experimental design impacts on numeric stability and parameter estimation (or numeric stability in parameter estimation). A spectrum of practical sequential design heuristics for stochastic kriging are explored in [15].

[10] examine the qualitative features of high-quality experimental designs for building accurate Gaussian process emulators of *deterministic* computer experiments. For deterministic emulators, it is shown that the weakly conflicting space-filling properties “small fill distance” and “large separation distance” ensure well-controlled error. Non-stationarity in the process’s correlation decay indicates a higher density of input locations in regions with more quickly decaying correlation, while non-trivial regression functions indicate a balance between the space-filling properties and traditional design properties targeting small variances of least squares coefficient estimates. In the common situation where correlation parameters are estimated within the Gaussian process framework, space-filling designs are slightly shifted to emphasis particular sizes and orientations of pairwise differences between input locations.

Here, we seek to develop and justify overarching principles of data collection for *stochas-*

*tic* kriging. Importantly, the primary target here is a *qualitative* indication of what type of designs might be expected to enable one to build an accurate model, not optimal design. Broadly, the development here follows the framework and many of the results laid out in [10]. Throughout, results which extend in a relatively straightforward manner from the deterministic case to the stochastic case will be described in brief, at a high level with differences highlighted, while completely unique results and those for which extension is more complex will be described in more depth.

Inaccuracy in stochastic kriging will be decomposed into four components, nominal, numeric, parameter estimation, and parameter estimation numeric error. The overall approach is to bound these four types of error in terms of experimental design properties. It will be shown that the implied design characteristics for these four sources of error are *weakly conflicting*. In Section 1.2, the problem is formally stated, some notation provided, and several important well-known results stated. Then, in respective Sections 1.3, 1.4, 1.5, and 1.6, the nominal, numeric, parameter estimation, and parameter estimation numeric error are bounded. Designs which are high-quality with respect to the provided bounds are discussed and a few examples are given, with consideration to stationary and non-stationary cases as well as non-trivial regression functions. In Section 1.7, a few numeric examples comparing the accuracy of stochastic kriging emulators based on a spectrum of designs, numbers of replications, and process/noise variances are presented to illustrate the proposed principles. Conclusions and implications are discussed briefly in Section 1.8.

## **1.2 Preliminaries**

### 1.2.1 Stochastic Kriging Model

We consider the situation where a *noisy* output  $y(x)$  can be observed at an input configuration  $x$  in a compact set  $\Omega \subset \mathbb{R}^d$ . The output is noisy, or stochastic, in the sense that another run, or observation, at  $x$  will give a different output value. The noisy outputs are modeled

as the sum of a deterministic function plus mean zero Gaussian noise. That is,

$$y(x) = f(x) + \epsilon(x), \quad (1.1)$$

where  $\epsilon(x) \sim N(0, \sigma_{\tau_*}^2(x))$  and  $\tau_* \in \mathbb{R}^{p_1}$  is a vector of parameters. Throughout, we will annotate *true* parameter values, which are not subject to estimation or any type of numeric error, with an  $*$  whenever this distinction between the true parameter values and their estimated or noisy counterparts is useful. Notably, the noise components are taken as independent across both input locations and replications at the same input location, and we have suppressed the dependence of  $y(x)$  and  $\epsilon(x)$  on a random element, say  $v$ . Following [11], the deterministic component  $f : \Omega \rightarrow \mathbb{R}$  is modeled as a Gaussian process (GP) (see, for example, [4] and [5]),  $f \sim \text{GP}(h(\cdot)' \beta_*, \Psi_{\theta_*}(\cdot, \cdot))$  for some fixed, known *regression* functions  $h : \Omega \rightarrow \mathbb{R}^q$  and a positive definite covariance function  $\Psi_{\theta_*}(\cdot, \cdot)$ . Here, the process mean and covariance depend on respective unknown parameters  $\beta_* \in \mathbb{R}^q$  and  $\theta_* \in \mathbb{R}^{p_2}$ . Let  $\vartheta = (\beta^T, \theta^T, \tau^T)^T$  denote the vector consisting of all the parameters. Throughout, the underlying mean function in the stochastic kriging model will be considered as the primary estimation target.

As shown in [11], the best linear unbiased predictor, as well as its MSPE, can be expressed in terms of the *distinct* data locations and the average output at each. The likelihood of the unknown parameters given the data, on the other hand, depends on all the individual outputs, not just the average at each distinct location, as shown in [16]. Throughout, we will use notation following [16]. Let  $\bar{Y}$  denote the vector of average responses at each of the  $n$  distinct locations and  $\bar{X}$  to denote the *corresponding* distinct design locations. On the other hand, we will use  $Y$  to denote the full vector of  $m$  outputs (not averaged) and  $X$  to denote the corresponding (potentially non-distinct) design locations. For the  $i^{\text{th}}$  *distinct* design location  $x_i$ , let  $k_i$  denote the number of replications observed at  $x_i$ . Then, the experimental design corresponding to the  $i^{\text{th}}$  component of  $\bar{Y}$  can be described in terms of the

pair  $(x_i, k_i)$  for  $i = 1, \dots, n$ , where  $x_i \in \Omega$  denotes a distinct design point, and  $k_i$  denotes the number of replicates at  $x_i$ . Let

$$\bar{y}(x_i) = \frac{1}{k_i} \sum_{j=1}^{k_i} y_j(x_i).$$

denote the sample mean at point  $x_i$ , where  $y_j(x_i)$  denotes the  $j$ th experiment at  $x_i$ . Similarly, let  $\bar{\Sigma}_\epsilon = \text{diag}\{\sigma_\tau^2(x_1)/k_1, \dots, \sigma_\tau^2(x_n)/k_n\}$  denote the diagonal matrix of marginal *noise* variances of the components  $\bar{Y}$ , and let  $\Sigma_\epsilon$  denote the diagonal matrix of marginal *noise* variances of the components  $Y$ .

If  $\beta_*$  is unknown, but both  $\theta_*$  and  $\tau_*$  are known, then the BLUP for  $f$  at an arbitrary location of interest  $x \in \Omega$  is [8]

$$\hat{f}_\theta(x) = h(x)^T \hat{\beta} + \Psi_\theta(x, \bar{X}) [\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1} (\bar{Y} - H(\bar{X}) \hat{\beta}), \quad (1.2)$$

where  $\hat{\beta} = (H(\bar{X})^T [\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1} H(\bar{X}))^{-1} H(\bar{X})^T [\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1} \bar{Y}$ ,  $H(\bar{X})$  has  $i^{\text{th}}$  row  $h(x_i)'$  for distinct data location  $x_i$ , and  $\Psi_\theta(A, B)$  has elements  $\Psi_\theta(a_i, b_j)$ . Similarly, the BLUP (1.2) has expected squared prediction error (conditional on the observed data), or mean squared prediction error, (MSPE)

$$\Psi_\theta(x, x) - (h(x)^T, \Psi_\theta(x, \bar{X})) \begin{pmatrix} 0 & H(\bar{X})^T \\ H(\bar{X}) & \Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon \end{pmatrix}^{-1} \begin{pmatrix} h(x) \\ \Psi_\theta(\bar{X}, x) \end{pmatrix}. \quad (1.3)$$

Applying block matrix inverse results [17], the MSPE (1.3) can be written as

$$\begin{aligned} & \Psi_\theta(x, x) - \Psi_\theta(x, \bar{X}) [\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1} \Psi_\theta(\bar{X}, x) \\ & + (h(x) - H(\bar{X})^T [\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1} \Psi_\theta(\bar{X}, x))^T \\ & \quad \times (H(\bar{X})^T [\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1} H(\bar{X}))^{-1} \\ & \quad \times (h(x) - H(\bar{X})^T [\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1} \Psi_\theta(\bar{X}, x)). \end{aligned} \quad (1.4)$$

At first glance, the stochastic kriging model, which assumes a Gaussian process mean with Gaussian noise, appears quite narrow and restrictive. In fact, the model is not as restrictive as it appears. In particular, if one believes that the target function  $f$  lies in a reproducing kernel Hilbert space (say for example,  $f$  has a fixed number of continuous partial derivatives), then a representer theorem [18] ensures that the solution to a very broad range of loss or likelihood-based penalized regression problems has the form given in (1.2), although  $\hat{\beta}$  would be estimated differently and the regularizing matrix  $\bar{\Sigma}_\epsilon$  constructed differently, depending on the loss or likelihood. In practice, the stochastic kriging model is typically a high-accuracy non-parametric estimate of the underlying function  $f$ , and would represent a high-quality starting approximation for each of the three examples mentioned in the first paragraph of this article (turbulent flows, sexual transmissibility, and cardiovascular policy).

The stochastic kriging model, which is adapted to simulations with noisy outputs, differs from a kriging model for simulations with deterministic outputs only by the inclusion of  $\bar{\Sigma}_\epsilon = \text{diag}\{\sigma_\tau^2(x_1)/k_1, \dots, \sigma_\tau^2(x_n)/k_n\}$  in the BLUP and MSPE formulas above. In a sense, the kriging model for deterministic simulations is a special case of the stochastic kriging model for which  $\sigma_\tau^2(\cdot) \equiv 0$ . Many of results developed below extend immediately to the kriging model for deterministic simulations by taking  $\sigma_\tau^2(\cdot) \equiv 0$  or the number of replications at the  $i^{\text{th}}$  distinct input location  $k_i \rightarrow \infty$  across  $i$ . Similarly, many of the results developed in [10] for deterministic kriging translate directly to the stochastic kriging context with only a cosmetic rework. The aspects of parameter estimation error that relate to estimation of  $\sigma_\tau^2(\cdot)$  are an exception.

### 1.2.2 Sources of Inaccuracy

As stated in the final paragraph of Section 1.1, inaccuracy in stochastic kriging will be decomposed into four components, nominal, numeric, parameter estimation, and parameter estimation numeric error. The numeric emulator is, in a sense, the actual, tangible emula-

tor, which is subject to parameter estimation error as well as numeric error in both emulator calculation and parameter estimation. Let  $\vartheta_*$ ,  $\hat{\vartheta}$ , and  $\tilde{\vartheta}$  respectively denote the true parameters, estimated parameters *not* subject to floating point errors, and estimated parameters subject to floating point error in both computation and optimization. As noted previously,  $*$  will be used throughout to annotate true parameter values. Similarly, we will use  $\hat{\cdot}$  and  $\tilde{\cdot}$  to identify quantities subject to estimation and numeric error, respectively. Similar to decompositions in [9] and [10], the norm of the difference between the estimator of the unknown function and real function can be decomposed into nominal, numeric, parameter estimation, and parameter estimation numeric components using the triangle inequality as follows,

$$\begin{aligned} \|f - \tilde{f}_{\tilde{\vartheta}}\| &= \|f - \hat{f}_{\vartheta_*} + \hat{f}_{\vartheta_*} - \hat{f}_{\hat{\vartheta}} + \hat{f}_{\hat{\vartheta}} - \hat{f}_{\tilde{\vartheta}} + \hat{f}_{\tilde{\vartheta}} - \tilde{f}_{\tilde{\vartheta}}\| \\ &\leq \underbrace{\|f - \hat{f}_{\vartheta_*}\|}_{\text{nominal}} + \underbrace{\|\hat{f}_{\vartheta_*} - \hat{f}_{\hat{\vartheta}}\|}_{\text{parameter estimation}} + \underbrace{\|\hat{f}_{\hat{\vartheta}} - \hat{f}_{\tilde{\vartheta}}\|}_{\text{parameter numeric}} + \underbrace{\|\hat{f}_{\tilde{\vartheta}} - \tilde{f}_{\tilde{\vartheta}}\|}_{\text{numeric}}. \end{aligned} \quad (1.5)$$

Here  $\tilde{f}_{\tilde{\vartheta}}$  denotes the nominal emulator subject to floating point errors. Nominal error refers to the difference between the target function  $f$  and its *idealized* approximation  $\hat{f}_{\vartheta_*}$ , which is not subject to floating point or parameter estimation error. Numeric error refers to the difference between the computed emulator  $\tilde{f}_{\tilde{\vartheta}}$ , which is subject to floating point arithmetic, and an idealized version of the emulator which is not subject to floating point error in emulator computation  $\hat{f}_{\tilde{\vartheta}}$ . Parameter estimation error represents the difference between emulators with the true and estimated parameters,  $\hat{f}_{\vartheta_*}$  and  $\hat{f}_{\hat{\vartheta}}$ , respectively. Parameter estimation numeric error refers to the difference between the emulator with numerically estimated parameters under floating point arithmetic  $\hat{f}_{\tilde{\vartheta}}$  and the emulator under an *exactly* estimated parameter  $\hat{f}_{\hat{\vartheta}}$ . While decomposition (1.5) holds for any norm, here the  $L_2(\Omega)$  norm will be the primary focus. Taking the expectation (conditional on the data) of (1.5)

and applying Jensen's inequality and Fubini's theorem [19] gives

$$\begin{aligned} \mathbb{E}\|f - \tilde{f}_{\hat{\vartheta}}\| &\leq \sqrt{\int_{\Omega} \mathbb{E}(f(x) - \hat{f}_{\vartheta_*}(x))^2 dx} + \sqrt{\int_{\Omega} \mathbb{E}(\hat{f}_{\vartheta_*}(x) - \hat{f}_{\hat{\vartheta}}(x))^2 dx} \\ &\quad + \sqrt{\int_{\Omega} \mathbb{E}(\hat{f}_{\hat{\vartheta}}(x) - \tilde{f}_{\hat{\vartheta}}(x))^2 dx} + \sqrt{\int_{\Omega} \mathbb{E}(\tilde{f}_{\hat{\vartheta}}(x) - f(x))^2 dx}. \end{aligned} \quad (1.6)$$

Notice that the BLUP with parameter  $\vartheta_*$  is the nominal emulator  $\hat{f}_{\vartheta_*}$  in the first term in (1.6) above, while the portion of the first term, bounding the nominal error above, under the square root and inside the integral,  $\mathbb{E}(f(x) - \hat{f}_{\vartheta_*}(x))^2$ , equals the MSPE (1.3).

### 1.3 Nominal Error

For a particular design problem, we have two approaches to reduce MSPE. The first approach is to add more distinct input locations to reduce the distance between potential inputs and design points, the other is to take more experimental runs at a particular location to reduce the predictive variance at that location. Intuitively, if there is a cluster of design points, then the MSPE of the experimental design including the cluster is almost the same as the MSPE of the experimental design with multiple experiments at one of the points in this cluster. Our intuition is correct, as a consequence of the continuity of matrix summation, inverses, and quadratic forms, as summarized in Proposition 1.3.1 below.

**Proposition 1.3.1.** *Suppose  $f \sim \text{GP}(h(\cdot)'\beta, \Psi_{\theta}(\cdot, \cdot))$ , for some fixed, known functions  $h(\cdot)$  and a positive definite function  $\Psi_{\theta}(\cdot, \cdot)$ , with stochastic observations generated by the stochastic kriging model described in Section 1.2.1. Let  $X = (X_1, X_2)$ , where  $X_1 = (x_1, x_2, \dots, x_r)$  and*

$$X' = (\underbrace{x^*, \dots, x^*}_{r \text{ replications}}, X_2).$$

*If  $\sigma_{\tau}^2(\cdot) > 0$  and  $\sigma_{\tau}^2(\cdot)$ ,  $h(\cdot)$ , and  $\Psi_{\theta}(\cdot, \cdot)$  are continuous, then  $\text{MSPE}(x) \rightarrow \text{MSPE}'(x)$*

as  $x_i \rightarrow x^*$  for  $i = 1, \dots, r$ , where  $MSPE(x)$  is the MSPE of the BLUP based on  $X$  and  $MSPE'(x)$  is the MSPE of the BLUP based on  $X'$ .

A bound on the nominal error for the uppermost terms of the MSPE (1.4), which provide the MSPE for a mean model with no regression functions, is provided in Theorem 1.3.1. A proof is given in Section A.1 of Appendix. Notably, the proof follows the strategy laid out in the proof of Theorem 3.1 in [10], with a few additional complexities in handling  $\bar{\Sigma}_\epsilon$ . In fact, the deterministic kriging result in Theorem 3.1 of [10], can be obtained as a special case of the Theorem below by setting  $\lambda_{\max}(\bar{\Sigma}_\epsilon) = 0$ . Throughout, we will use the notation  $\lambda_{\max}(A)$  and  $\lambda_{\min}(A)$  to denote the maximum and minimum eigenvalues of a symmetric matrix  $A$ .

**Theorem 1.3.1.** *Suppose  $f \sim GP(0, \Psi_\theta(\cdot, \cdot))$  for a positive definite function  $\Psi_\theta(\cdot, \cdot)$  with  $\Psi_\theta(\cdot, \cdot) \geq 0$ , stochastic observations are generated by the stochastic kriging model described in Section 1.2.1,  $(n - 2) \sup_{u, v \in \Omega} \Psi_\theta(u, v) > \lambda_{\max}(\bar{\Sigma}_\epsilon)$ , then the MSPE of  $f$  has upper bound*

$$\begin{aligned} & \Psi_\theta(x, x) - 2\Psi_\theta(x_i, x) + \Psi_\theta(x_i, x_i) - \frac{(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i))^2}{n \sup_{u, v \in \Omega} \Psi_\theta(u, v) + \lambda_{\max}(\bar{\Sigma}_\epsilon)} \\ & + \frac{\lambda_{\max}(\bar{\Sigma}_\epsilon)(n \sup_{u, v \in \Omega} \Psi_\theta(u, v) + 2(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i)))}{n \sup_{u, v \in \Omega} \Psi_\theta(u, v) + \lambda_{\max}(\bar{\Sigma}_\epsilon)}. \end{aligned} \quad (1.7)$$

Two special cases are examined. These cases respectively represent broadly applicable *stationary* and *non-stationary* covariance models for the *process*  $f$ , and will be referred to as the Stationary Model and Non-Stationary Model. These models provide a concrete structure within which we can gain a qualitative understanding of the design implications of both stationarity and non-stationarity. In the upcoming development, the overall bound on the uppermost terms of (1.4) will be expressed in terms of the maximum of local bounds,

$$\sup_{x \in \Omega} \mathbb{E}\{f(x) - \hat{f}_\vartheta(x)\}^2 = \max_i \sup_{x \in A_i} \mathbb{E}\{f(x) - \hat{f}_\vartheta(x)\}^2, \quad (1.8)$$

where  $\cup_i A_i = \Omega$ . The maximum over  $i$  in (1.8) can be controlled by imposing a uniform bound over each of its components. Below,  $\varphi(\cdot)$  is a decreasing function of its non-negative argument and  $\bar{\Gamma}$  is diagonal.

### 1.3.1 Stationary Model

Suppose  $\Psi_\theta(u, v) = \sigma^2 \varphi(\|\Theta(u - v)\|_2)$  with  $\bar{\Sigma}_\epsilon = \sigma^2 \bar{\Gamma}$ , where  $\sigma \in \mathbb{R}_+$  is a parameter and  $\Theta \in \mathbb{R}^{d \times d}$  is a non-singular matrix, which could be a parameter in its own right or a function of a lower dimensional parameter. Consider using the bound (1.7) as a guidepost for identifying the features of a high-quality experimental design. Unlike in the deterministic case discussed in [10], in the stochastic kriging case, the denominator influences the bound (1.7) through  $\bar{\Sigma}_\epsilon$ , inducing a balance between the variance at each point and the fill distance. Notice that in (1.7), the bound is an increasing function of  $\Psi_\theta(x, x) - \Psi_\theta(x_i, x) = \Psi_\theta(x_i, x_i) - \Psi_\theta(x_i, x)$ . Let  $A_i = V_i(\Theta)$ , the Voronoi cell [20] anchored by distinct data point  $x_i$ , with respect to a Mahalanobis distance [21]

$$V_i(\Theta) = \{x \in \Omega : d_\Theta(x, x_i) \leq d_\Theta(x, x_j) \quad \forall j \neq i\},$$

where  $d_\Theta(u, v) = \sqrt{(u - v)' \Theta' \Theta (u - v)}$ . On  $A_i = V_i(\Theta)$ , the bound given by (1.7) can be bounded in terms of the smallest value of  $\Psi_\theta(x_i, x)$ , which is attained for  $x$  maximizing  $d_\Theta(x_i, x)$ . Taking the maximum over  $i$ , and letting  $\nu = \varphi(0) - \varphi(\max_i \sup_{x \in V_i(\Theta)} d_\Theta(x_i, x))$ , (1.7) can be rewritten as

$$\sigma^2 \left( 2\nu - \frac{\nu^2}{n\varphi(0) + \lambda_{\max}(\bar{\Gamma})} + \frac{\lambda_{\max}(\bar{\Gamma})(n\varphi(0) + 2\nu)}{n\varphi(0) + \lambda_{\max}(\bar{\Gamma})} \right). \quad (1.9)$$

Notice that

$$\max_i \sup_{x \in V_i(\Theta)} d_\Theta(x_i, x) = \sup_{x \in \Omega} \min_i d_\Theta(x_i, x)$$

is the fill distance with respect to the distance  $d_\Theta$ . Since (1.9) is an increasing function of  $\nu \in [0, \varphi(0)]$ , the upper bound can be controlled by demanding the fill distance is small, *balanced with small largest element of  $\bar{\Gamma}$ .*

**Interpretation.** *In the context of the stationary stochastic kriging model described above, experimental designs which balance small fill distance, with respect to the distance  $d_\Theta$ , for the distinct input locations with replication targeting uniformly small  $\bar{\Sigma}_\epsilon$  ensure well-controlled nominal error.*

### 1.3.2 Non-Stationary Model

Here, we consider a relatively simple model of non-stationarity, adapted from [22], which forms a good approximation in many practical situations. In brief, the correlation decay is taken to be composed of more rapidly and more slowly decaying components, with the emphasis on the components depending on the input locations. This model of non-stationarity is reasonably well-suited to situations where the surface of interest is varying more quickly in some input regions and more slowly in others, and the model provides a structure for examining the design implications of this type of non-stationarity.

Suppose  $\Psi_\Theta(u, v) = \sigma^2(\omega_1(u)\omega_1(v)\varphi(\|\Theta_1(u - v)\|_2) + \omega_2(u)\omega_2(v)\varphi(\|\Theta_2(u - v)\|_2))$  with  $\bar{\Sigma}_\epsilon = \sigma^2\bar{\Gamma}$ , where  $\sigma \in \mathbb{R}_+$  is a parameter, and  $\Theta_1, \Theta_2 \in \mathbb{R}^{d \times d}$  are non-singular matrices, either parameters in their own right or functions of lower dimensional parameters. For the Non-Stationary Model case, assume in addition  $\omega_1(\cdot), \omega_2(\cdot) \geq 0$  have Lipschitz continuous derivatives on  $\Omega$  with Lipschitz constants  $k_1$  and  $k_2$ , respectively,  $\omega_1^2(\cdot) + \omega_2^2(\cdot) = 1$ ,  $\Theta_1, \Theta_2$  are non-singular, and  $\lambda_{\max}(\Theta_1'\Xi_2'\Xi_2\Theta_1) < 1$ , where  $\Xi_2 = \Theta_2^{-1}$ . The final assumption can be interpreted as  $\varphi(\|\Theta_2(\cdot - \cdot)\|_2)$  is narrower than  $\varphi(\|\Theta_1(\cdot - \cdot)\|_2)$ . For the Non-Stationary Model, we will localize the bounds over *unions* of Voronoi cells  $V_i^* = V_i(\Theta_1) \cup V_i(\Theta_2)$ . Note that  $V_i^* \approx V_i(\Theta_1)$  and  $V_i^* \approx V_i(\Theta_2)$  if  $\Theta_1 \approx c\Theta_2$  for some  $c$ .

Similar to the Stationary Model, we take the maximum over  $i$ , and let

$$\nu = \varphi(0) - \inf_{x \in V_i^*} \{ \omega_1(x) \omega_1(x_i) \varphi(\|\Theta_1(x - x_i)\|_2) + \omega_2(x) \omega_2(x_i) \varphi(\|\Theta_2(x - x_i)\|_2) \}. \quad (1.10)$$

Then, (1.7) again gives upper bound (1.9). Using Lipschitz continuity of  $\omega_1(\cdot), \omega_2(\cdot)$  and Taylor's theorem [23], it can be shown that [10],

$$\begin{aligned} \nu \leq & \varphi(0) - (\omega_1^2(x_i) \varphi(\sup_{x \in V_i^*} d_{\Theta_1}(x_i, x)) + \omega_2^2(x_i) \varphi(\sup_{x \in V_i^*} d_{\Theta_2}(x_i, x))) \\ & - \varphi(0)(k_1 + k_2) \max_i \sup_{x \in V_i^*} \|x - x_i\|_2. \end{aligned} \quad (1.11)$$

By plugging the right-hand side of (1.11) into (1.9), we can obtain an upper bound, and corresponding guidepost for identifying features of a high-quality nominal error experimental design.

Following the development in [10], it can be shown that for fixed  $\sigma_\tau^2(x_i)/k_i$ ,  $i = 1, \dots, n$  (or equivalently  $\bar{\Gamma}$ ), (1.9) is bounded uniformly over the design space by an experimental design with smaller *union* of Voronoi cells, with respect to both  $d_{\Theta_1}$  and  $d_{\Theta_2}$ , in regions with more emphasis on the quickly decaying correlation, and vice versa. Similar to [10], the global and local correlation emphases are given concretely by

$$\omega_k(x_i)^2 \left( \varphi \left( \sup_{x \in V_i^*} d_{\Theta_1}(x_i, x) \right) - \varphi \left( \sup_{x \in V_i^*} d_{\Theta_2}(x_i, x) \right) \right), \quad k = 1, 2.$$

**Interpretation.** *In the context of the non-stationary stochastic kriging model described above, experimental designs which balance smaller fill distances for the distinct input locations, with respect to distances  $d_{\Theta_1}$  and  $d_{\Theta_2}$ , in regions of the input space with more rapidly decaying correlation and larger fill distances in regions with more slowly decaying correlation, with replication targeting uniformly small  $\bar{\Sigma}_\epsilon$  ensure well-controlled nominal error.*

### 1.3.3 Regression Functions

Next, we consider the lowermost terms in (1.4), expressing the contribution of the regression terms to the overall accuracy. The regression terms can be bounded as

$$\begin{aligned}
& (h(x) - H(\bar{X})^T[\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1}\Psi_\theta(\bar{X}, x))^T (H(\bar{X})^T[\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1}H(\bar{X}))^{-1} \\
& \quad \times (h(x) - H(\bar{X})^T[\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1}\Psi_\theta(\bar{X}, x)) \\
& \leq \sigma^2 \frac{n\varphi(0) + \lambda_{\max}(\bar{\Gamma})}{\lambda_{\min}(H(\bar{X})'H(\bar{X}))} \|h(x) - H(\bar{X})^T[\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1}\Psi_\theta(\bar{X}, x)\|_2^2.
\end{aligned} \tag{1.12}$$

The term  $\lambda_{\max}(\bar{\Gamma})$  encourages balanced replication in the sense that it encourages a small maximum of  $\sigma_\tau^2(x_i)/k_i$ . The term  $\lambda_{\min}(H(\bar{X})'H(\bar{X}))$  in the denominator, on the other hand, encourages some degree of *traditional* design properties. For example, linear regression functions would push input locations towards the edges or corners of the design space. On the other hand, the final term is the sum of squared errors for smoothed estimates of the regression functions and would be expected to be small in precisely the same situations when the topmost terms in (1.4) are small, under the assumption that the regression functions can be well-approximated using the kernel  $\Psi_\theta$  [10]. That is, replication and traditional design properties need to be balanced with fill distance-based criteria.

**Interpretation.** *In the context of stochastic kriging models with non-trivial regression functions, experimental designs which balance space-filling properties, of the stationary or non-stationary variety as appropriate, replication targeting uniformly small  $\bar{\Sigma}_\epsilon$ , and traditional design properties targeting low-variance regression function coefficient estimates ensure well-controlled nominal error.*

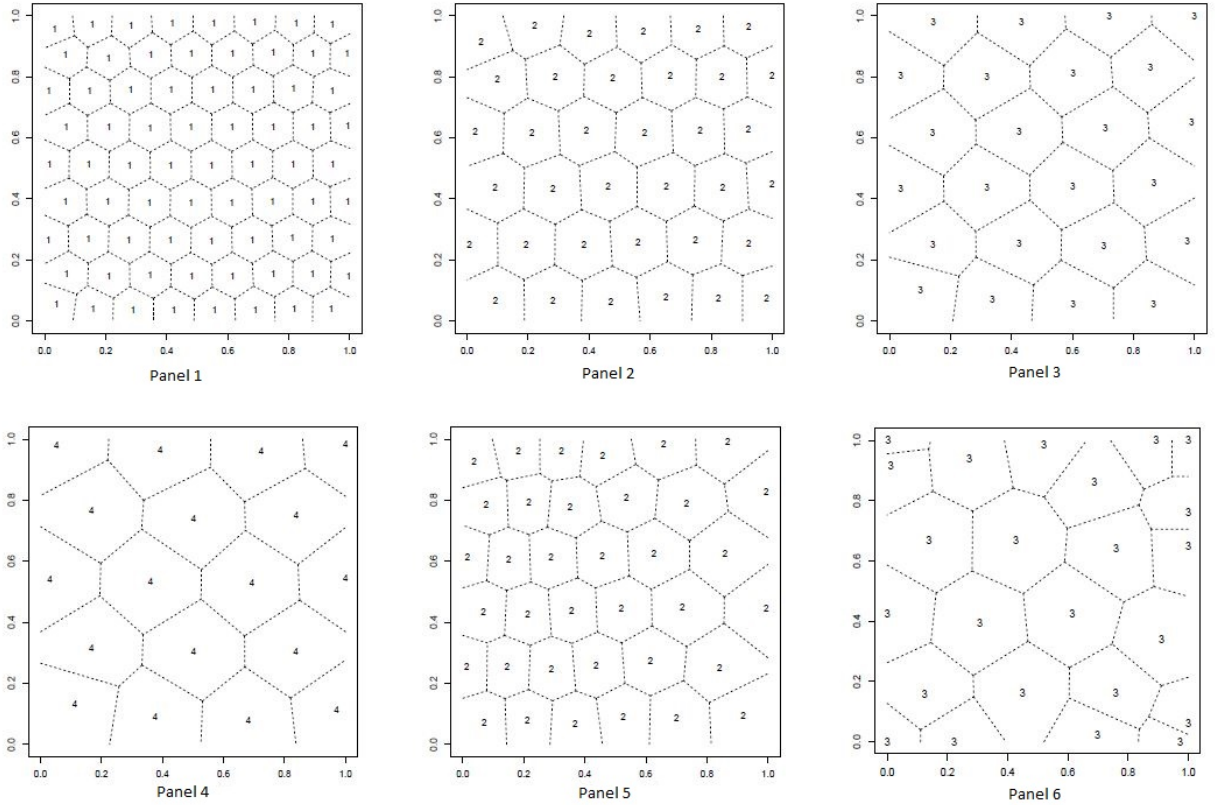
### 1.3.4 Example Designs

Here, we seek to illustrate the type of designs indicated by the nominal error bounds, and provide a measure of corroboration for the qualitative features of good experimental de-

signs that the bounds suggest. For a given practical context and hypothetical values for the covariance parameters, the actual nominal error (1.4) is computable, and could represent a component of a reasonable objective.

Example high quality designs for stochastic kriging problems in the stationary situation, across a range of ratios  $\sigma_\tau^2(x_i)/\sigma^2$ , and the non-stationary situation, as well as the stationary situation along with a constant and linear regression functions are shown in Figure 1.1. For the stationary cases shown in Panels 1-4, distinct design locations arrange themselves in a space-filling pattern, minimizing the fill distance. For the the non-stationary case shown in Panel 5, more distinct design locations are needed in portions of the input space with more emphasis on the more rapidly decaying correlation. For the situation where a constant and linear regression functions are included with a stationary stochastic process variance, distinct design locations are pushed towards the corners of the input space, balancing space-filling and traditional design properties. As the ratio of noise variance to functional variance  $\sigma_\tau^2(x_i)/\sigma^2$  increases, more replications are needed at each distinct design location, moving from no replication when  $\sigma_\tau^2(x_i)/\sigma^2 = 0.03$  to four replications when  $\sigma_\tau^2(x_i)/\sigma^2 = 0.45$ .

These designs are obtained by minimizing the nominal error bounds given by (1.9), by plugging (1.11) into (1.9), and by taking the summation of (1.12) and (1.9), for the respective stationary, non-stationary, and non-trivial regression functions situations. Since the noise variance is constant over the region, we need only consider the case where the number of replications at each distinct input location are equal. The designs which minimize the upper bounds can then be obtained by minimizing over the number of replications. In general, finding high-quality experimental designs is challenging, particularly when the value of the objective function is non-smooth and non-convex. For a given number of replications, we can adopt the homotopy continuation [24] procedure applied in [10]. In brief, we optimize the bounds over several iterations, slowly transitioning from an easier objective to the target objective.



**Figure 1.1: Panels 1-4:** Nominal error designs for *stationary* correlation with  $\varphi(d) = \exp\{-d^2\}$  and respective ratios  $\sigma_\tau^2(x_i)/\sigma^2$ , which is a constant, of 0.03, 0.10, 0.25, and 0.45. **Panel 5:** Nominal error design for *non-stationary* correlation with  $\varphi(d) = \exp\{-d^2\}$ ,  $\omega_1(x) = x_1$ ,  $\Theta_1 = I_2$ ,  $\Theta_2 = 4I_2$ , and ratio  $\sigma_\tau^2(x_i)/\sigma^2$  of 0.10. **Panel 6:** Nominal error design for *stationary* correlation with  $\varphi(d) = \exp\{-d^2\}$ , and ratio  $\sigma_\tau^2(x_i)/\sigma^2$  of 0.25, along with a constant and two linear regression functions. Design points annotated with number of replications throughout.

## 1.4 Numeric Error

Numeric error comes from at least two sources. The first source is rounding error in the computer's representation of real numbers, and the second source is numeric solution to the parameter optimization problem. In this section we develop bounds, in terms of properties of the experimental design, on the numeric error coming from the first numeric source of error, namely  $\|\hat{f}_{\tilde{y}} - \tilde{f}_{\tilde{y}}\|$ . It can be shown that, similar to the non-stochastic kriging situation [10], increasing the number of data points always decreases the nominal error. Unlike non-

stochastic kriging, increasing the number of data points in the stochastic situation has far less ability to adversely affect numeric accuracy, particularly when  $\sigma_r^2(x_i)$  is non-negligible. It will be shown that the first source of numeric error can be controlled via the minimum eigenvalue of  $\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon$ , which has

$$\lambda_{\min}(\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon) \geq \lambda_{\min}(\Psi_\theta(\bar{X}, \bar{X})) + \lambda_{\min}(\bar{\Sigma}_\epsilon).$$

Numeric accuracy depends on the accuracy of floating point matrix manipulations. Commonly, computer and software have 15 digits of accuracy meaning roughly that

$$\|\tilde{x} - x\|_2 / \|x\|_2 \leq 10^{-15},$$

where  $x$  denotes the actual value and  $\tilde{x}$  denotes the value that the computer stores. Theorem 1.4.1 provides a bound on numeric error. The proof is essentially identical to the proof of Theorem 4.1 provided in the Appendix to [10], except with the additional  $\bar{\Sigma}_\epsilon$  in the representation of the emulator (1.2), so is omitted for brevity.

**Theorem 1.4.1.** *Suppose  $f \sim \text{GP}(h(\cdot)'\beta, \Psi_\theta(\cdot, \cdot))$ , for some fixed, known functions  $h(\cdot)$  and a positive definite function  $\Psi_\theta(\cdot, \cdot)$ , with stochastic observations generated by the stochastic kriging model described in Section 1.2.1. For any fixed parameter estimate  $\tilde{\vartheta}$ , under Assumption A.2.1 (see Section A.2 of Appendix),*

$$\begin{aligned} & |\hat{f}_{\tilde{\vartheta}}(x) - \tilde{f}_{\tilde{\vartheta}}(x)| \\ & \leq \delta \|h(x)\|_2 \|\tilde{\beta}\|_2 + \frac{2\delta}{1-r} \|\Psi_{\tilde{\vartheta}}(\bar{X}, x)\|_2 (\|H(\bar{X})\|_2 \|\tilde{\beta}\|_2 + \|\hat{f}(\bar{X})\|_2) g(\Psi_\theta(\bar{X}, \bar{X}), \bar{\Sigma}_\epsilon), \end{aligned}$$

where

$$g(\Psi_\theta(\bar{X}, \bar{X}), \bar{\Sigma}_\epsilon) = \frac{1 + \kappa(\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon)}{\lambda_{\min}(\Psi_\theta(\bar{X}, \bar{X})) + \lambda_{\min}(\bar{\Sigma}_\epsilon)},$$

and  $\kappa(\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon)$  denotes the condition number of  $\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon$ .

Assumption A.2.1 requires the calculation of functions  $h$ ,  $\hat{f}$  and  $\Psi$  to be relative accurate (see Section A.2 of Appendix). Note that

$$g(\Psi_\theta(\bar{X}, \bar{X}), \bar{\Sigma}_\epsilon) \leq \frac{1}{\lambda_{\min}(\Psi_\theta(\bar{X}, \bar{X})) + \lambda_{\min}(\bar{\Sigma}_\epsilon)} \left( 1 + \frac{n \sup_{u,v \in \Omega} \Psi_\theta(u, v) + \lambda_{\max}(\bar{\Sigma}_\epsilon)}{\lambda_{\min}(\Psi_\theta(\bar{X}, \bar{X})) + \lambda_{\min}(\bar{\Sigma}_\epsilon)} \right), \quad (1.13)$$

where the inequality follows from Gershgorin's theorem [25]. See, equation (A.3).

The norm  $\|h(x)\|_2$  does not depend on the experimental design. For experimental designs which are not too small and have reasonable *parameter estimation numeric* properties, it will be shown in Section 1.6 that  $\|\tilde{\beta}\|_2$  will approximately equal  $\|\beta\|_2$ . Similarly, for experimental designs which are not too small and have reasonable *nominal* properties,  $\|\hat{f}(\bar{X})\|_2$  depends primarily on the sample size and large sample distribution of the inputs, as well as the target function  $f$ . Further, for experimental designs which are not too small, the norms  $\|\Psi_{\hat{\theta}}(\bar{X}, x)\|_2$  and  $\|H(\bar{X})\|_2$  depend primarily on the sample size and large sample distribution of the inputs. Thus, aside from  $g(\Psi_\theta(\bar{X}, \bar{X}), \bar{\Sigma}_\epsilon)$ , the other terms in the bound in the theorem influence the numeric error only weakly. The bound depends on the experimental design primarily through  $g(\Psi_\theta(\bar{X}, \bar{X}), \bar{\Sigma}_\epsilon)$ , which can be controlled via  $\lambda_{\min}(\Psi_\theta(\bar{X}, \bar{X})) + \lambda_{\min}(\bar{\Sigma}_\epsilon)$  as seen in (1.13). Unless  $\lambda_{\min}(\bar{\Sigma}_\epsilon)$  is *very* near zero, the numeric error associated with generating predictions from a stochastic kriging model may be expected to be substantially less than in the deterministic case. On the other hand, when  $\lambda_{\min}(\bar{\Sigma}_\epsilon)$  is very near zero, the numeric error in generating predictions would behave in a manner described in Section 3 of [10], favoring designs with *well separated* distinct locations and, in the presence of non-stationarity, a greater (lesser) density of distinct input locations in sub-regions of the input space with more emphasis on local (global) correlation. Within the framework described above, the relatively common practice in deterministic kriging of including a small so-called *nugget*  $\delta$ , corresponding to  $\bar{\Sigma}_\epsilon = \delta I_n$ , has the ef-

fect of greatly reducing numeric and parameter estimation numeric error, while (hopefully) only slightly increasing nominal and parameter estimation error.

**Interpretation.** *In the context of the stochastic kriging model described above, numeric error is well-controlled for experimental designs with either well-separated distinct input locations or  $\lambda_{\min}(\bar{\Sigma}_\epsilon)$  not too small.*

## 1.5 Parameter Estimation Error

Throughout this section, the variance of the noise component  $\sigma_\tau^2(x)$  is taken as a continuously *differentiable* function of the unknown parameter vector  $\tau$ . Maximum likelihood estimation of parameters is considered. As shown in [16], the likelihood of the parameters given the observed data depends on each individual output, not just their average at each distinct design location. In this section, we will work with the full observation vector  $Y$  and corresponding (potentially repeated) full design  $X$ . Up to an additive constant, the log-likelihood is

$$l = -\frac{1}{2} \log \det[\Psi_\theta(X, X) + \Sigma_\epsilon] - \frac{1}{2} (Y - H(X)\beta)^T [\Psi_\theta(X, X) + \Sigma_\epsilon]^{-1} (Y - H(X)\beta).$$

Note that this log-likelihood can be computed in an efficient manner using results in [16].

Let  $\mathbb{E}$  denote the expectation conditional on  $X$  and  $Y$ . Then, for  $n$  and  $k_i$  not too small,

$$\mathbb{E}\{f_{\hat{\vartheta}_*}(x) - \hat{f}_{\hat{\vartheta}}(x)\}^2 \approx \frac{\partial \hat{f}(x)}{\partial \vartheta'_*} \mathcal{I}(\vartheta_*)^{-1} \frac{\partial \hat{f}(x)}{\partial \vartheta_*}, \quad (1.14)$$

where  $\mathcal{I}(\vartheta_*)$  denotes the information matrix. For approximation (1.14) to hold, we need the sequence of likelihood functions to become increasingly peaked. For more details, see [26].

In general, parameter estimates might be expected to affect the accuracy of Gaussian process regression models relatively weakly. In fact, the order of approximation error will

be the same across a wide range of parameter estimates, as long as the target function is in the reproducing kernel Hilbert space associated with the basic kernel [9]. The Fisher's information based error approximation in (1.14), while highly accurate only for large (and informative) samples, provides guidance for ensuring that the data we collect will enable construction of parameter estimates within this wide acceptable range.

For parameter estimation error, we have the following theorem, whose proof is provided in Section A.3 of Appendix. Similar to Theorem 1.3.1, the proof of Theorem 1.5.1 follows the strategy laid out in the proof of Theorem 5.1 in [10], with a few additional complexities in handling the noise variance parameters  $\tau$ . Once again, the deterministic kriging result in Theorem 5.1 of [10], can be obtained as a special case of the Theorem below by setting  $\sigma_\tau^2(\cdot) = 0$  and omitting the  $c_4$  terms. In the theorem, the Gaussian process covariance's parameters are separated as  $\Psi_\theta(\cdot, \cdot) = \sigma^2 \Phi_\rho(\cdot, \cdot)$ .

**Theorem 1.5.1.** *Let  $f \sim \text{GP}(h(\cdot)' \beta, \sigma^2 \Phi_\rho(\cdot, \cdot))$  for some fixed, known functions  $h(\cdot)$  and positive definite function  $\Phi_\rho(\cdot, \cdot)$ , with stochastic observations generated by the stochastic kriging model described in Section 1.2.1. Suppose  $\hat{\vartheta}$  is the maximum likelihood estimator of the full set of unknown parameters  $\vartheta = (\beta, \sigma^2, \rho, \tau)$ . Then, an approximate upper bound for  $\mathbb{E}\{\hat{f}_{\vartheta_*}(x) - \hat{f}_{\hat{\vartheta}}(x)\}^2$  is given by*

$$\frac{\sigma^2 \|c_1\|_2^2 (m \sup_{u,v \in \Omega} \Phi_\rho(u, v) + \lambda_{\max}(\Sigma_\gamma))}{ms_2} + \frac{\sigma^4 \|c\|_2^2 (m \sup_{u,v \in \Omega} \Phi_\rho(u, v) + \lambda_{\max}(\Sigma_\gamma))^2}{m^2 s_1}, \quad (1.15)$$

where

$$\begin{aligned}
c_1 &= \frac{\partial \hat{f}(x)}{\partial \beta} = h(x) - H(X)^T (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} \Phi_\rho(X, x), \\
(c_3)_j &= \frac{\partial \hat{f}(x)}{\partial \rho_j} = \left( \frac{\partial \Phi_\rho(x, X)}{\partial \rho_j} - \Phi_\rho(x, X) (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} \frac{\partial \Phi_\rho(X, X)}{\partial \rho_j} \right) \\
&\quad \times (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} (Y - H(X)\beta), \quad j = 1, \dots, p_2, \\
(c_4)_t &= \frac{\partial \hat{f}(x)}{\partial \tau_t} = \Phi_\rho(x, X) (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} \text{diag} \left( \frac{\partial \gamma_1}{\partial \tau_t} I_{k_1}, \dots, \frac{\partial \gamma_i}{\partial \tau_t} I_{k_i}, \dots, \frac{\partial \gamma_n}{\partial \tau_t} I_{k_n} \right) \\
&\quad \times (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} (Y - H(X)\beta), \quad t = 1, \dots, p_1, \\
c &= (c_3^T, c_4^T),
\end{aligned}$$

where  $(c_j)_i$  denotes the  $i^{\text{th}}$  element in vector  $c_j$ ,  $\gamma_i = \sigma_\tau^2(x_i)/\sigma^2$ ,  $\Sigma_\gamma = \text{diag}(\gamma_1 I_{k_1}, \dots, \gamma_n I_{k_n})$ ,  $k_i$  is the number of replicates on  $i^{\text{th}}$  point,  $m = \sum_{i=1}^n k_i$ , and  $s_1$  and  $s_2$  are respectively defined in (A.7) and (A.10) in Appendix.

The upper bound is approximate in the sense that for a sequence of experimental designs with convergent large sample distribution and maximum likelihood parameter estimates, the probability that the upper bound is violated by more than  $\varepsilon > 0$  goes to zero.

Following the development in [10], both  $\|c_1\|_2^2$  and  $\|c_3\|_2^2$  involve *interpolation* errors, for the regression functions and the derivatives of the Gaussian process covariance, respectively, and these components would be expected to be small for high quality nominal designs. The remaining terms in  $c_3$  are either well-controlled for high quality numeric designs, in the case of  $(\Psi_\theta(X, X) + \Sigma_\epsilon)^{-1}$ , or depend only weakly on aspects of the experimental design beyond its size and large sample distribution, in the case of  $Y - H(X)\beta$ . For  $c_4$ , we have the following proposition, whose proof is given in Section A.4 of Appendix.

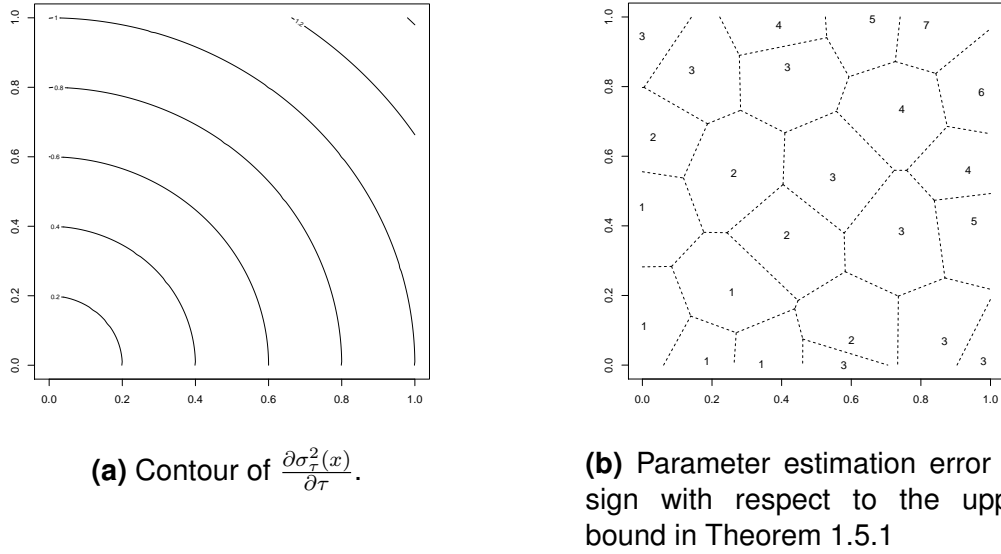
**Proposition 1.5.1.** *Under the conditions of Theorem 1.5.1,*

$$|(c_4)_t| \leq \frac{\|\Phi_\rho(x, \bar{X})\|_2 \|\bar{Y} - H(\bar{X})\beta\|_2}{(\lambda_{\min}(\Phi_\rho(\bar{X}, \bar{X}) + \bar{\Sigma}_\gamma))^2} \max_{i: x_i \in \bar{X}} \left| \frac{1}{k_i} \frac{\partial \gamma_i}{\partial \tau_t} \right|. \quad (1.16)$$

The initial terms in (1.16) are either well-controlled for high quality numeric designs, for  $\lambda_{\min}(\Phi_\rho(\bar{X}, \bar{X}) + \bar{\Sigma}_\gamma)$ , or depend only weakly on aspects of the experimental design beyond its number of distinct locations and their large sample distribution, for  $\|\Phi_\rho(x, \bar{X})\|_2$  and  $\|\bar{Y} - H(\bar{X})\beta\|_2$ . The last term in (1.16),  $\max \left| \frac{1}{k_i} \frac{\partial \gamma_i}{\partial \tau_t} \right|$ , encourages replication, since it is a decreasing function of  $k_i$ . Moreover, replication is more strongly encouraged near locations  $x_i$  where  $\gamma_i = \sigma_\tau^2(x_i)/\sigma^2$  is changing more rapidly with respect to one of the parameters  $\tau_t$ . The term  $s_2$  introduces a *push* towards experimental design properties targeting reduction in variance of the regression function coefficients.

The term  $s_1$  is somewhat more complex. Let  $W_1(x, y) = \Phi_\rho(x, y) + \sigma_\tau^2(x)/\sigma^2 \mathbb{I}_{\{x=y\}}$  and  $\xi = (\rho, \tau)'$ . By (A.7),  $s_1 \geq 0$  and  $s_1 > 0$  unless  $\frac{\partial W_1(x, y)}{\partial \xi} a = W_1(x, y) b$  with probability 1 for some  $(a, b) \neq 0$ . There are two parts to  $\frac{\partial W_1(x, y)}{\partial \xi}$ ,  $\frac{\partial \Phi_\rho(x, y)}{\partial \rho}$  and  $\frac{\partial \sigma_\tau^2(x)}{\partial \tau} \mathbb{I}_{\{x=y\}}$ . Consider the *distinct* and *replicated* locations,  $x \neq y$  and  $x = y$ , separately. The term  $s_1$  will be large if two conditions are met. First, the differences between distinct locations  $\{x_i - x_j\}$  make  $\frac{\partial \Phi_\rho(x_i, x_j)}{\partial \rho}$  far from zero, balanced with respect to a basis of  $\mathbb{R}^{\dim \rho}$ , and not collinear with  $\Phi_\rho(x_i, x_j)$ , similar to [10]. Second, the locations of replications make  $\frac{\partial \sigma_\tau^2(x_i)}{\partial \tau}$  far from zero, not collinear with  $\Phi_\rho(x, x) + \sigma_\tau^2(x)/\sigma^2$ , and *balanced* in the sense that locations for which the derivative  $\frac{\partial \sigma_\tau^2(x_i)}{\partial \tau}$  is *small* in magnitude require *more* replicates and *vice versa*. Notice that this encouragement of more replications where the derivative is smaller runs contrary to the influence of the term  $\max \left| \frac{1}{k_i} \frac{\partial \gamma_i}{\partial \tau_t} \right|$  in  $c_4$ , which encourages more replications where the derivative is large in magnitude. Taken together, numeric studies suggest that the bound (1.15) is small for experimental designs whose distinct locations have good nominal and numeric properties, balanced with sufficient replications at each distinct data site.

**Interpretation.** *In the context of the stochastic kriging model described above with parameters estimated via maximum likelihood, experimental designs with good nominal and numeric properties ensure well-controlled parameter estimation error.*



**Figure 1.2:** Contour of  $\frac{\partial \sigma_\tau^2(x)}{\partial \tau}$  and parameter estimation error design.

### 1.5.1 Example Design

Consider an example with  $\Psi(d) = \exp(-d^T d)$ ,  $\Psi_\rho(\cdot) = \Psi(\text{diag}\{\rho\}(\cdot))$ , and  $\rho = (1, 1)^T$ . In addition, suppose the stochastic error is given by  $\sigma_\tau^2(x) = \tau \|x\|_2 + 0.04$ , where  $\tau$  is a parameter with true value 1. Suppose we want design points on  $\Omega = [0, 1]^2$ . Since  $\frac{\partial \sigma_\tau^2(x)}{\partial \tau} = \|x\|_2$ , by (1.16), a high quality experimental design should put more replicates on the locations that are far from zero. The total number of design points (may not be distinct locations) is 72, and the number of unique location is 24. The corresponding design that minimizes the parameter estimation error bound (1.15) is shown in Figure 1.2. Notice that by balancing the stochastic error and  $\frac{\partial \sigma_\tau^2(x)}{\partial \tau}$ , the number of replicates are consistent with the contours of  $\frac{\partial \sigma_\tau^2(x)}{\partial \tau}$ , subject to edge effects.

## 1.6 Parameter Estimation Numeric Error

In this section, the numeric error coming from numeric optimization of parameter estimates, the second source of numeric error,  $\|\hat{f}_{\hat{\vartheta}} - \hat{f}_{\vartheta}\|_2$ , is discussed. Recall that  $\hat{f}_{\vartheta}(x) = h(x)^T \beta + \Psi_\theta(x, \bar{X})[\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\tau]^{-1}(\bar{y} - H(\bar{X})\beta)$ , where each element of  $\bar{X}$  denotes

a distinct data location, and  $\bar{\Sigma}_\tau = \text{diag}(\sigma_\tau^2(x_1)/k_1, \dots, \sigma_\tau^2(x_i)/k_i, \dots, \sigma_\tau^2(x_n)/k_n)$ . Let  $\tilde{A} = \Psi_{\hat{\theta}}(\bar{X}, \bar{X}) + \bar{\Sigma}_\tau$  and  $\hat{A} = \Psi_{\hat{\theta}}(\bar{X}, \bar{X}) + \bar{\Sigma}_{\hat{\tau}}$  denote the corresponding quantities subject to parameter estimation numeric error from numeric optimization and theoretical parameter estimates. The below result links experimental design properties to parameter estimation numeric error. A proof is provided in Section A.5 of Appendix.

**Theorem 1.6.1.** *Suppose  $f \sim \text{GP}(h(\cdot)'\beta, \Psi_\theta(\cdot, \cdot))$  for some fixed, known function  $h(\cdot)$  and positive definite function  $\Psi_\theta(\cdot, \cdot)$ , with stochastic observations generated by model (1.1). Let  $\tilde{\vartheta}$  denote the parameter we derive from numeric optimization and let  $\hat{\vartheta}$  denote the true solution to the parameter optimization problem. Let  $\hat{f}_{\tilde{\vartheta}}$  and  $\hat{f}_{\hat{\vartheta}}$  denote the BLUPs for  $f$  with respective parameters  $\hat{\vartheta}$  and  $\tilde{\vartheta}$ . Then, under Assumptions A.2.1, A.5.1, and A.5.2,*

$$\begin{aligned} |\hat{f}_{\tilde{\vartheta}} - \hat{f}_{\hat{\vartheta}}| &\leq \frac{2\delta\kappa(\hat{A})}{(1-r)\lambda_{\min}(\hat{A})} \|\Psi_{\hat{\theta}}(\bar{X}, x)\|_2 (\|f(x)\|_2 + \|H(\bar{X})\|_2 \|\hat{\beta}\|_2) \\ &\quad + 2\delta \left( \kappa(\hat{A})\kappa(H(\bar{X})^T H(\bar{X})) \left( 1 + (1+\delta)^2 + \frac{(1+\delta)^2}{1-r} \kappa(\hat{A}) \right) + 1 \right) \\ &\quad \times (\|h(x)\|_2 + \frac{1+r}{(1-r)\lambda_{\min}(\hat{A})} \|H(\bar{X})\|_2 \|\Psi_{\hat{\theta}}(\bar{X}, x)\|_2) \|\hat{\beta}\|_2. \end{aligned} \quad (1.17)$$

**Remark 1.6.2.** *If Assumption A.5.2 does not hold, we can still use Lemma A.5.1 to derive an upper bound of  $|\hat{f}_{\tilde{\vartheta}} - \hat{f}_{\hat{\vartheta}}|$ , which is of order  $\delta\kappa(\hat{A})^3$ .*

Most of the terms above also appeared in Theorem 1.4.1. The parameter estimation numeric error can also be controlled via  $\lambda_{\min}(\Psi_\theta(\bar{X}, \bar{X})) + \lambda_{\min}(\bar{\Sigma}_\epsilon)$  as seen in (1.13). See Section 1.4 for a detailed discussion. The term  $\kappa(H(\bar{X})^T H(\bar{X}))$  requires some degree of *traditional* design properties, as discussed in Section 1.3. However, the parameter estimation numeric error has a higher order of influence on the error as a whole on the left-hand side of (1.5) than the numeric error since there is a  $\kappa(\hat{A})^2$  on the right-hand side of (1.17).

**Interpretation.** *In the context of the stochastic kriging model described above with variance-covariance parameters estimated via numerically maximizing the likelihood, experimental designs with good numeric properties, slightly shifted towards good traditional design*

*properties if non-trivial regression functions are included, ensure well-controlled parameter estimation numeric error.*

## 1.7 Numeric Examples

In this section, we report simulation studies comparing designs with different numbers of replications. Notably, we focus on the relationship between the number of replications at each distinct input location and the relative sizes of process and noise variation, potentially varying over the input space. The relationship between the space-filling properties of the distinct input locations and emulator accuracy is examined empirically in [10].

### 1.7.1 Constant ratio of noise and process variance

Take  $\Psi(u, v) = \exp(-\|u - v\|_2^2)$ ,  $\sigma^2 = 1$ , and space of interest  $\Omega = [0, 1]^2$ . The total number of design points (potentially non-distinct) is set at 72, and the number of replicates varied across 1, 2, 3, and 4, for 72, 36, 24, and 18 distinct locations. Take  $\epsilon(x) \sim N(0, \sigma_\epsilon^2)$  for all  $x \in \Omega$ .

For the initial study, set  $\sigma_\epsilon^2$  to be 0.5, 0.1, and 0.01. Designs examined for the distinct input locations include the nominal designs shown in the first four panels of Figure 1.1,  $S$ -optimal Latin hypercubes [27], random Latin hypercubes [28], random uniform designs, and MaxPro designs [29]. First, 300 draws from the Gaussian process with mean zero and the correlation function  $\Psi(\cdot, \cdot)$  are generated. For each generating, the observations based on the design and 100 point random uniform testing set are made. Random errors draw from  $N(0, \sigma_\epsilon^2)$  are added to each of the observation on the design points. Based on the observations with random noise on the design points, predictions generated on the testing set are calculated, and the maximum squared prediction error are computed. The R packages `lhs` [30] and `MaxPro` [31] were used for generating Latin hypercube and MaxPro designs. The average maximum squared prediction error over 300 draws is calculated, and the results are reported in Table 1.1.

**Table 1.1:** Average maximum squared prediction error for a spectrum of experimental designs across numbers of replications.

$\sigma_\epsilon^2 = 0.5$				
Design	rep = 4	rep = 3	rep = 2	rep = 1
nominal	0.206	0.202	0.221	0.212
optLHS	0.236	0.229	0.221	0.244
randLHS	0.261	0.240	0.246	0.217
random	0.295	0.278	0.249	0.237
MaxPro	0.192	0.214	0.214	0.203
$\sigma_\epsilon^2 = 0.1$				
Design	rep = 4	rep = 3	rep = 2	rep = 1
nominal	0.071	0.071	0.071	0.065
optLHS	0.088	0.083	0.079	0.073
randLHS	0.109	0.092	0.091	0.081
random	0.137	0.117	0.095	0.084
MaxPro	0.059	0.063	0.067	0.067
$\sigma_\epsilon^2 = 0.01$				
Design	rep = 4	rep = 3	rep = 2	rep = 1
nominal	0.012	0.012	0.013	0.012
optLHS	0.020	0.018	0.017	0.014
randLHS	0.033	0.024	0.020	0.017
random	0.047	0.036	0.025	0.017
MaxPro	0.012	0.011	0.011	0.010

For a particular choice of experimental design strategy for the distinct input locations, we see an overall trend favoring replication as noise increases and space-fillingness as noise decreases. Similar to [10], we see good performance for MaxPro designs [29] for the distinct input locations, as well as designs selected via the nominal error bound (1.9).

Next, we examine the quality of the nominal error bound (1.9), as well as any potential losses in accuracy due to following the guidance of the nominal error bounds in terms of the number of replications. Here, the Gaussian process draws follow the same settings as the previous study. The designs examined here are optimal Latin hypercube and MaxPro. The total number of (potentially non-distinct) design points is set at 72, with numbers of replicates in  $\{24, 18, 12, 9, 8, 6, 4, 3, 2, 1\}$ , for numbers of distinct locations in  $\{3, 4, 6, 8, 9, 12, 18, 24, 36, 72\}$ . Noise standard deviations  $\sigma_\epsilon$  are taken in  $\{0.05, 0.35, 0.5\}$ . Comparisons of the nominal error bound (1.9) to the average maximum squared prediction

error over 300 draws of the Gaussian process are presented in Figure 1.3. As the number of replicates increases, the bound decreases, then increases. The simulated average maximum squared prediction error, on the other hand, varies somewhat more slowly than the upper bound.

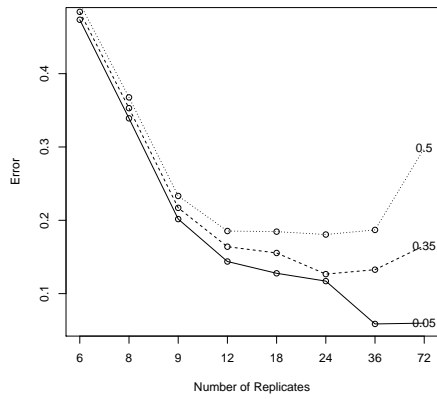
Consider using the nominal error bound (1.9) as guidance for choosing the number of replicates. Here, we compare the average maximum prediction error under the best choice of replications to the average maximum prediction error under the number of replications suggested by the nominal bound. The noise standard deviations are taken to be  $\sigma_\epsilon = 0.05k$  for  $k = 1, \dots, 10$ . Relative and absolute differences in error are shown in Table 1.2. Results suggest that the bound provides useful guidance describing the qualities of a high-quality experimental design.

**Table 1.2:** The average maximum prediction error under the best choice of replications to the average maximum prediction error under the number of replications suggested by the nominal bound.

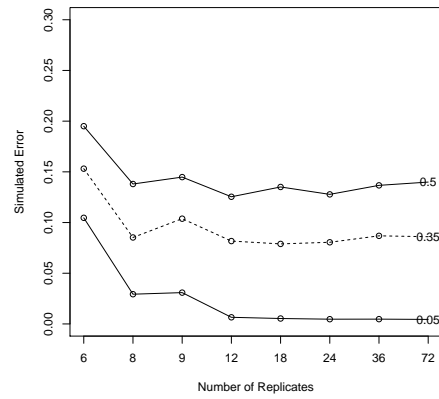
$\sigma_\epsilon$	MaxPro		optLHS	
	relative error	absolute error	relative error	absolute error
0.05	0.132	0.00052	0	0
0.10	0	0	0	0
0.15	0.194	0.00401	0.093	0.00285
0.20	0.049	0.00173	0.040	0.00176
0.25	0.065	0.00331	0	0
0.30	0.036	0.00209	0.266	0.02271
0.35	0	0	0.058	0.00577
0.40	0.003	0.00027	0	0
0.45	0.065	0.00700	0	0
0.50	0.076	0.00968	0.030	0.00459

### 1.7.2 Input varying ratio of noise and process variance

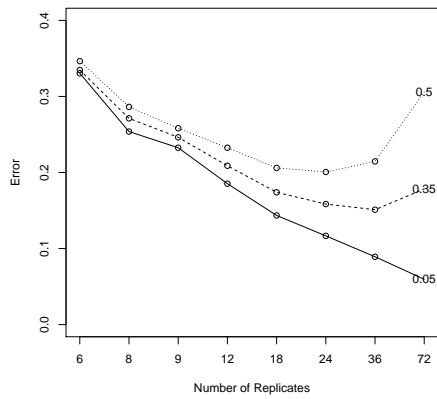
Next, we examine the model discussed in Section 1.5.1, with noise level varying over the input space. In particular,  $\Psi(d) = \exp(-d^T d)$ ,  $\Psi_\rho(\cdot) = \Psi(\text{diag}\{\rho\}(\cdot))$ , and  $\rho = (1, 1)^T$ , with stochastic error given by  $\sigma_\tau^2(x) = \tau\|x\|_2 + 0.04$ , where  $\tau$  is a parameter with true value 1. Again, the input space is  $\Omega = [0, 1]^2$ , and the total number of design points



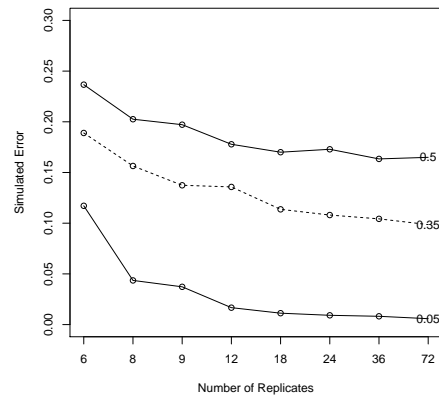
**(a)** Nominal error bound values for MaxPro designs.



**(b)** Average maximum squared prediction errors for MaxPro.



**(c)** Nominal error bound values for optimal Latin hypercube designs.



**(d)** Average maximum squared prediction errors for optimal Latin hypercube.

**Figure 1.3:** Comparisons of the nominal error bound (1.9) to the average maximum squared prediction error.

(potentially not distinct) is 72. The parameter estimation design is provided by minimizing the upper bound provided in Theorem 1.5.1. Several designs for the distinct input locations including the nominal design provided in this paper, numeric designs obtained along the lines described in [10], optimal Latin hypercubes [27], random Latin hypercubes [28], random uniform designs, and MaxPro designs [29], are considered. We compare designs with equal replication at all distinct input locations and designs with unequal replications at the input locations as guided by (1.7), in which we require that the diagonal elements in  $\bar{\Sigma}_\epsilon$  are nearly equal. The number of unique locations used in the comparison are 18, 24, and 36. Then, we run 300 independent Gaussian processes (with noise) and compare the average maximum squared prediction error of these processes. Results are shown in Table 1.3. In brief, accuracy is dramatically improved by varying the number of replications across the distinct input locations in the situation where the noise level varies across the input space.

**Table 1.3:** Average maximum squared prediction error comparisons across number of distinct input locations and input varying replication vs. constant replication.

Design	18 points		24 points		36 points	
	Varying	Const.	Varying	Const.	Varying	Const.
Nominal	0.139	0.182	0.146	0.192	0.164	0.212
Numeric	0.108	0.155	0.128	0.185	0.159	0.193
Parameter Est.	0.113	0.141	0.125	0.170	0.125	0.190
optLHS	0.157	0.209	0.144	0.209	0.129	0.198
randLHS	0.176	0.210	0.162	0.228	0.148	0.225
rand	0.229	0.267	0.184	0.239	0.160	0.227
MaxPro	0.111	0.159	0.113	0.173	0.116	0.197

## 1.8 Discussion

We have developed and justified guidelines for ensuring accuracy of stochastic kriging predictors based on experimental design. By controlling nominal, numeric, parameter estimation and parameter estimation numeric sources of error, we can control overall error in stochastic kriging. As in [10], the space-filling properties, “small fill-distance” and

“large separation-distance”, are also largely non-conflicting with each of the sources of error. Unlike [10], there is a trade-off between the number of replicates at each distinct design location and the space-filling properties of the distinct design locations. This trade-off is reflected in the upper bounds for each of the four sources of errors. The numeric error and parameter estimation numeric error are closely related to the condition number of  $\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon$ , which always becomes larger as more replicates or data locations are added. Nominal and parameter estimation error, on the other hand, tend to encourage small fill-distance.

This work has several limitations. Only upper bounds on the sources of error are considered. There may be two designs with the same upper bound, where one is much better than the other with respect to the expected error. We do not consider error from incorrectly using Gaussian process regression with maximum likelihood estimation to estimate the target function (model mis-specification). From another perspective, the order of approximation error will be the same across a huge range of parameter estimates, as long as the target function is in the reproducing kernel Hilbert space associated with the basic kernel [9]. Projection design properties have not been explicitly discussed. On the other hand, the results presented here indicate that if inert inputs are expected, then the distinct design locations should be space-filling in lower-dimensional projections of the design. Lastly, there are situations where a *stochastic* emulator is need. If the Gaussian *noise* model fits the data well, then a stochastic emulator could be constructed by adding Gaussian noise with the estimated variance. If the *noise* model fits poorly, then perhaps a localized resampling of residuals could be useful.

In brief, these results provide further motivation and rationale for using one of several apparently high-quality, space-filling experimental designs for the distinct input locations, including but not limited to [29], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], or [42], particularly when there is no reason to expect non-stationarity in the process or noise. While evidence of non-stationarity in process or noise variance, from an initial design per-

haps, would indicate a varying density of distinct input locations or number of replications at distinct locations, respectively, precise characterization of this variation across the input space is challenging. More generally, optimization of experimental designs is very challenging under many criteria, due to the high-dimensional and multi-modal nature of many of these problems. On the other hand, a fixed number of replications across the design space, paired with one (or even a few) high-quality and computationally attractive space-filling designs, as might be appropriate in a situation with stationarity in both process and noise, could conceivably be chosen in a computationally efficient manner for a moderately sized belief-set of noise to process variance ratios.

## CHAPTER 2

### ON PREDICTION PROPERTIES OF KRIGING: UNIFORM ERROR BOUNDS AND ROBUSTNESS

#### 2.1 Introduction

Kriging is a widely used methodology to reconstruct functions based on their scattered evaluations. Originally, kriging was introduced to geostatistics by [1]. Later, it has been applied to computer experiments [43], machine learning [6] and related areas. With kriging, one can obtain an interpolant of the observed data, that is, the predictive curve or surface goes through all data points. Conventional regression methods, like the linear regression, the local polynomial regression [44] and the smoothing splines [45], do not have this property. It is suitable to use interpolation in spatial statistics and machine learning when the random noise of the data is negligible. The interpolation property is particularly helpful in computer experiments, in which the aim is to construct a surrogate model for a deterministic computer code, such as a finite element solver.

A key element of kriging prediction is the use of the conditional inference of Gaussian processes. At each untried point, the conditional distribution of a Gaussian process is normal with explicit mean and variance. The confidence interval of the kriging predictor is then constructed using this conditional distribution. However, there is a gap between the theory of kriging and its practical usage. In practice, kriging is mostly used to recover a function, not just to predict at one particular point. In this situation, the pointwise predictive distributions do not contain the desired information. For example, combining the 95% confidence intervals at each point of the input space does not yield a 95% confidence limit for predicting the whole function, although this inaccurate approach is commonly used.

In this work, we derive error bounds of the (simple) kriging predictor under a uniform

metric. The predictive error is bounded in terms of the maximum pointwise predictive variance of kriging, which can be further bounded with the fill distance of the design set. This work shows that the overall predictive performance of a Gaussian process model is tied to the smoothness of the correlation function as well as the space-filling property of the design. We also show that a less smooth correlation function is more robust in prediction, in the sense that prediction consistency can be achieved for a broader range of true correlation functions. Since our work shows that the kriging predictor can achieve both the uniform convergence and robustness, we refer to this property as *universal* convergence of kriging. The theory of radial basis function approximation [46] and a maximum inequality for Gaussian processes [47, 48] are employed as axillary tools in our technical development.

This paper is organized as follows. In Section 2.2, we review the mathematical foundation of simple kriging and state the objective of this paper. In Section 2.3, we discuss kriging interpolation under a misspecified correlation function. In Section 2.4, we review some concepts and results from the theory of radial basis approximation. In Section 2.5, we present our main results on the uniform error bounds for kriging predictors. Some simulation studies are conducted in Section 2.6, which confirm our theoretical analysis. Concluding remarks and discussion are given in Section 2.7. Appendix B.1 includes some necessary mathematical tools. Appendix B.2 contains the proof of Theorem 2.5.1, the main theorem of this work.

## 2.2 Review on the simple kriging method

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

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

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

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

with some  $\phi > 0$ , where  $\|\cdot\|$  denotes the Euclidean norm. The isotropic Matérn correlation function [5, 26] is defined as

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

where  $\phi, \nu > 0$  and  $K_{\nu}$  is the modified Bessel function of the second kind. The parameter  $\nu$  is often called the smoothness parameter, because it determines the smoothness of the Gaussian process [49].

Suppose that we have observed  $Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_n)$ , in which  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are distinct points. We shall use the terminology in design of experiments [50] and call  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  the *design points*, although in some situations (e.g., in spatial statistics and machine learning) these points are observed without the use of design. In this paper, we do not assume any (algebraic or geometric) structure for the design points  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ . In other words, they are scattered points.

The aim of (simple) kriging is to predict  $Z(\mathbf{x})$  at an untried  $\mathbf{x}$  based on the observed data  $Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_n)$ , which is done by calculating the conditional distribution. It follows from standard arguments [5, 51] that, conditional on  $Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_n)$ ,  $Z(\mathbf{x})$  is normally distributed, with

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

$$\text{Var}[Z(\mathbf{x})|Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_n)] = \sigma^2(1 - \mathbf{r}^T(\mathbf{x})\mathbf{K}^{-1}\mathbf{r}(\mathbf{x})), \quad a.s., \quad (2.4)$$

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

The conditional expectation  $\mathbb{E}[Z(\mathbf{x})|Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_n)]$  is a natural predictor of  $Z(\mathbf{x})$  using  $Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_n)$ , because it is the best linear predictor [26, 5]. It is worth noting that a nice property of the Gaussian process models is that the predictor (2.3) has an explicit expression, which explains why kriging is so useful.

The above simple kriging method can be extended. Instead of using a mean zero Gaussian process, one may introduce extra degrees of freedom by assuming that the Gaussian process has an unknown constant mean, or more generally one may assume the mean function is given by a linear combination of regression functions. The corresponding methods are referred to as ordinary kriging and universal kriging, respectively. A standard prediction scheme then is the best linear unbiased prediction [5, 26]. For the ease of mathematical treatment, we only consider simple kriging in this work. The convergence theory for ordinary and universal kriging requires separate developments. Further discussions are deferred to Section 2.7.

### 2.2.1 Goal of this work

Although kriging has nice and simple predictive distributions, there remain several theoretical issues which have not been addressed.

First, (2.3) and (2.4) only give the predictive distribution at a single point. In many practical problems, we are interested in recovering a whole function rather than predicting for just one point. Therefore, it is natural to ask whether the kriging predictor converges uniformly in the domain of interest. Also, one may want to know whether or how the correlation function has an effect on the kriging prediction power. For example, one may raise this question: between Gaussian random fields with Gaussian and Matérn correlations, which one is easier to predict?

Second, (2.3) and (2.4) said nothing about experimental design. When the design points are controllable, like in computer experiments and related areas, one may want a good

allocation scheme of the design points, to ensure certain overall balancing or optimality properties. In the area of computer experiments, space-filling designs [5, 4], in which the design points spread (approximately) evenly in the experimental region, are commonly used for fitting a kriging model. But no theoretical justification of doing so can be seen from (2.3) and (2.4).

Finally, (2.3) and (2.4) hold only when the correlation function and the variance are known. But this rarely holds true in practice. Therefore, it is natural to ask what would happen if a misspecified correlation function is used. Also, are there correlation functions which are more robust against model misspecification? And what is the cost of gaining robustness?

This paper is devoted to answer the above questions by establishing a uniform error bound for the kriging predictor, given the covariance function and a quantity that measures the space-filling property of the design.

In the function approximation context, the error estimates of the kriging-type interpolants are studied in the literature of radial basis function approximation. We refer to [46] for a comprehensive coverage. Although the mathematical formulations of the interpolants given by kriging and radial basis functions are similar, the two methods are different in their mathematical settings and assumptions. In radial basis function approximation, the underlying function is assumed *fixed*, while kriging utilizes a probabilistic model, driven by a Gaussian random field. Because there is a lack of explicit error bounds for kriging in the literature, in recent years, quite a few authors (e.g., [52, 9, 53, 54]) use error bounds for radial basis functions to justify the predictive behavior of kriging. Because such results from radial basis functions do not directly address the random behavior of kriging, there is an urgent need to establish a uniform convergence theory for kriging.

Kriging with misspecified correlation functions is discussed in [55, 56, 57, 58, 59]. It has been proven in these papers that some correlation functions, especially the Matérn correlation family, are robust against model misspecification. However, explicit rate of

convergence in a general situation has not been obtained. More discussions on this point are given in Section 2.7.

### 2.3 Kriging interpolant

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

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

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

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

The only difference between (2.5) and (2.3) is that we replace the Gaussian process  $Z$  by a function  $f$  here. In other words,

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

As mentioned in Section 2.2, the conditional expectation  $\mathbb{E}[Z(\mathbf{x}) | Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_n)]$  is a natural predictor of  $Z(\mathbf{x})$ . Thus we are interested in bounding the predictive error of the kriging method, given by  $Z(\mathbf{x}) - \mathbb{E}[Z(\mathbf{x}) | Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_n)]$ , which is equal to  $Z(\mathbf{x}) - \mathcal{I}_{\Psi, \mathbf{X}} Z(\mathbf{x})$  almost surely.

Recall from Section 2.2.1 that we are looking for a theory that also works under model misspecification. Because  $\Psi$  is not known, we use another correlation function  $\Phi$  for prediction. We call  $\Psi$  the *true correlation function* and  $\Phi$  the *imposed correlation function*. Under the imposed correlation function, the kriging interpolant of the underlying Gaus-

sian process becomes  $\mathcal{I}_{\Phi, \mathbf{X}}Z(\mathbf{x})$ . In this situation, the interpolant cannot be interpreted as the conditional expectation. With an abuse of terminology, we will still call it the kriging predictor. Thus our aim is to study the approximation power of the kriging predictor. Specifically, we are interested in bounding the quantity

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

where  $\Omega$  is the region of interest, also called the experimental region, and  $\Omega \supset \{\mathbf{x}_1 \dots, \mathbf{x}_n\}$ . The present setting is related to the fixed-domain asymptotic analysis for kriging [55, 56, 58], which studies the asymptotic theory of kriging assuming that the experimental region  $\Omega$  is fixed while the design points become dense over the experimental region.

Because the predictive error in (2.7) is quantified under a uniform metric, the theory to be established can directly address the first theoretical concern raised in Section 2.2.1.

## 2.4 Power function and its upper bounds

In Section 2.3, we have defined the kriging interpolation operator  $\mathcal{I}_{\Phi, \mathbf{X}}$  which can be applied to an arbitrary function. In the area of scattered data approximation, the interpolation using operator  $\mathcal{I}_{\Phi, \mathbf{X}}$  is also called the *radial basis function approximation*. We refer to [46] for details.

A major problem in radial basis functions approximation is to bound the interpolation error  $f(x) - \mathcal{I}_{\Phi, \mathbf{X}}f(x)$  for an arbitrary deterministic function  $f$ . Because  $\Phi$  is not the true correlation function, to use the terminology in applied mathematics and machine learning, we call  $\Phi$  a *kernel function*.

A standard theory of radial basis function approximation works by employing the *reproducing kernel Hilbert space* generated by  $\Phi$ , denoted by  $\mathcal{N}_{\Phi}(\Omega)$ . A definition and some basic properties of the reproducing kernel Hilbert spaces are given in Appendix B.1.1.

If  $f \in \mathcal{N}_\Phi(\Omega)$ , then there is a simple error bound ([46], Theorem 11.4):

$$|f(\mathbf{x}) - \mathcal{I}_{\Phi, \mathbf{X}} f(\mathbf{x})| \leq P_{\Phi, \mathbf{X}}(\mathbf{x}) \|f\|_{\mathcal{N}_\Phi(\Omega)}, \quad (2.8)$$

for each  $\mathbf{x} \in \Omega$ , where  $\|f\|_{\mathcal{N}_\Phi(\Omega)}$  is the norm of  $f$  in the reproducing kernel Hilbert space,  $P_{\Phi, \mathbf{X}}(\mathbf{x})$  is a function independent of  $f$ . The square of  $P_{\Phi, \mathbf{X}}(\mathbf{x})$  is called the *power function*, given by

$$P_{\Phi, \mathbf{X}}^2(\mathbf{x}) = 1 - \mathbf{r}^T(\mathbf{x}) \mathbf{K}^{-1} \mathbf{r}(\mathbf{x}), \quad (2.9)$$

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

The statistical interpretation of the power function is evident. From (2.4) it can be seen that, if  $\Psi = \Phi$ , the power function is the kriging predictive variance for a Gaussian process with  $\sigma^2 = 1$ .

Inequality (2.8) gives an upper bound of the interpolation error, which is the product of two simpler quantities. The first quantity is independent of  $f$ , while the second depends only on  $f$ . To pursue a convergence result under the uniform metric, we define

$$P_{\Phi, \mathbf{X}} := \sup_{\mathbf{x} \in \Omega} P_{\Phi, \mathbf{X}}(\mathbf{x}). \quad (2.10)$$

As in (2.8), we wish to find an upper bound of  $P_{\Phi, \mathbf{X}}$ , in which the effects of the design  $\mathbf{X}$  and the kernel  $\Phi$  can be separated. This step is generally more complicated, but fortunately some upper bounds are available in the literature, especially for the Gaussian and the Matérn kernels. These bounds are given in terms of the *fill distance*, which is a quantity depending only on the design  $\mathbf{X}$ . Given the experimental region  $\Omega$ , the fill distance of a design  $\mathbf{X}$  is defined as

$$h_{\mathbf{X}} := \sup_{\mathbf{x} \in \Omega} \min_{\mathbf{x}_j \in \mathbf{X}} \|\mathbf{x} - \mathbf{x}_j\|. \quad (2.11)$$

Clearly, the fill distance quantifies the space-filling property [5] of a design. A design

having the minimum fill distance among all possible designs with the same number of points is known as a minimax distance design [60].

The upper bounds of  $P_{\Phi, \mathbf{X}}$  in terms of the fill distance for Gaussian and Matérn kernels are given in Theorem 2.4.1 and 2.4.2, respectively.

**Theorem 2.4.1** ([46], Theorem 11.22). *Let  $\Omega = [0, 1]^d$ ;  $\Phi(x)$  be a Gaussian kernel given by (2.1). Then there exist constants  $c, h_0$  depending only on  $\Omega$  and the scale parameter  $\phi$  in (2.1), such that  $P_{\Phi, \mathbf{X}} \leq h_{\mathbf{X}}^{c/h_{\mathbf{X}}}$  provided that  $h_{\mathbf{X}} \leq h_0$ .*

**Theorem 2.4.2** ([61], Theorem 5.14). *Let  $\Omega$  be compact and convex with a positive Lebesgue measure;  $\Phi(x)$  be a Matérn kernel given by (2.2) with the smoothness parameter  $\nu$ . Then there exist constants  $c, h_0$  depending only on  $\Omega, \nu$  and the scale parameter  $\phi$  in (2.2), such that  $P_{\Phi, \mathbf{X}} \leq ch_{\mathbf{X}}^\nu$  provided that  $h_{\mathbf{X}} \leq h_0$ .*

## 2.5 Uniform error bounds for kriging

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

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

**Condition 2.5.1.** *The kernels  $\Psi$  and  $\Phi$  are continuous and integrable on  $\mathbf{R}^d$ , satisfying*

$$\int_{\mathbf{R}^d} \|\omega\| \tilde{\Phi}(\omega) d\omega < +\infty, \quad (2.12)$$

and

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

We will show in Theorem 2.5.1 that, under Condition 2.5.1, the kriging predictor can attain the full convergence rate.

Now we are able to state the main theorem of this paper. Recall that  $\sigma^2$  is the variance of  $Z(\mathbf{x})$ .

**Theorem 2.5.1.** *Suppose Condition 2.5.1 holds and  $P_{\Phi, \mathbf{X}} \leq C \min\{A_1, 1\}$ , where  $P_{\Phi, \mathbf{X}}$  is defined in (2.10) and  $C$  is a constant depending on  $\Omega$ . Then for any  $u > 0$ , with probability at least  $1 - 2 \exp\{-u^2 / (2A_1^2 \sigma^2 P_{\Phi, \mathbf{X}}^2)\}$ , the kriging predictive error has the upper bound*

$$\sup_{\mathbf{x} \in \Omega} |Z(\mathbf{x}) - \mathcal{I}_{\Phi, \mathbf{X}} Z(\mathbf{x})| \leq K \sqrt{1 + A_1 A_1 \sigma P_{\Phi, \mathbf{X}}} \log^{1/2}(1/P_{\Phi, \mathbf{X}}) + u, \quad (2.13)$$

where  $K$  is a constant depending only on  $\Omega$ .

Theorem 2.5.1 presents some non-asymptotic upper bounds for the kriging predictive error. It implies some asymptotic results which are of traditional interests in this area. For instance, suppose we adopt a classic setting of fixed-domain asymptotics [26] in which the probabilistic structure of  $Z(\mathbf{x})$  and the kernel function  $\Phi$  are fixed, and the number of design points increases so that  $P_{\Phi, \mathbf{X}}$  tends to zero. Then from Theorem 2.5.1, it can be seen that, under Condition 2.5.1, the rate of convergence of the kriging predictor is

$$\sup_{\mathbf{x} \in \Omega} |Z(\mathbf{x}) - \mathcal{I}_{\Phi, \mathbf{X}} Z(\mathbf{x})| = O_p(P_{\Phi, \mathbf{X}} \log^{1/2}(1/P_{\Phi, \mathbf{X}})). \quad (2.14)$$

We believe that (2.14) is the full convergence rate because from (2.8) we can see that the convergence rate of the radial basis approximation for deterministic functions in the reproducing kernel Hilbert space is  $O(P_{\Phi, \mathbf{X}})$  and these two rates are nearly the same, except for a logarithmic factor. This is reasonable because the support of a Gaussian process is typically larger than the corresponding reproducing kernel Hilbert space. Specifically, the support of a Gaussian process is equal to the closure of the corresponding reproducing kernel Hilbert space under the uniform metric [62]. As said in Section 2.4, if  $\Psi = \Phi$ ,

$P_{\Phi, \mathbf{X}}$  is the supremum of the pointwise predictive standard deviation. Thus Theorem 2.5.1 implies that, if  $\Psi$  is known, the predictive error of kriging under the uniform metric is not much larger than its pointwise error.

Using the upper bounds of  $P_{\Phi, \mathbf{X}}$  given in Theorems 2.4.1 and 2.4.2, we can further deduce error bounds of the kriging predictor in terms of the fill distance defined in (2.11). Since these upper bounds are functions of the fill distance, we have justified the use of space-filling designs to fit kriging models. We demonstrate these results in Examples 2.5.1-2.5.3.

**Example 2.5.1.** Here we assume  $\Phi$  is a Matérn kernel in (2.2) with smoothness parameter  $\nu$ . It is known that

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

where  $\phi$  is the scale parameter in (2.2). See, for instance, [46, 63]. Suppose  $\Psi$  is a Matérn correlation function with smoothness  $\nu_0$ . It can be verified that Condition 2.5.1 holds if and only if  $1 < \nu \leq \nu_0$ . Therefore, if  $1 < \nu \leq \nu_0$ , we can invoke Theorems 2.4.2 and 2.5.1 to obtain that the kriging predictor converges to the true Gaussian process with a rate at least  $O_p(h_{\mathbf{X}}^\nu \log^{1/2}(1/h_{\mathbf{X}}))$  as  $h_{\mathbf{X}}$  tends to zero. It can be seen that the rate of convergence is maximized at  $\nu = \nu_0$ . In other words, if the true smoothness is known *a priori*, one can obtain the greatest rate of convergence.

**Example 2.5.2.** Suppose  $\Phi$  is the same as in Example 2.5.1, and  $\Psi$  is a Gaussian correlation function in (2.1), with spectral density [5]  $\tilde{\Psi}(\boldsymbol{\omega}) = (\pi/\phi)^{2/d} \exp\{-\|\boldsymbol{\omega}\|^2/(4\phi)\}$ , where  $\phi$  is the scale parameter in (2.1). Then Condition 2.5.1 holds for any choice of  $\nu$ . Then we can invoke Theorems 2.4.2 and 2.5.1 to obtain the same rate of convergence as in Example 2.5.1.

**Example 2.5.3.** Suppose  $\Phi = \Psi$ , and  $\Phi$  is a Gaussian kernel in (2.1). Then we can invoke Theorems 2.4.1 and 2.5.1 to obtain the rate of convergence  $O_p(h_{\mathbf{X}}^{c/h_{\mathbf{X}}-1/2} \log^{1/2}(1/h_{\mathbf{X}}))$

for some constant  $c > 0$ . Note that this rate is faster than the rates obtained in Examples 2.5.1-2.5.3, because it decays faster than any polynomial of  $h_{\mathbf{X}}$ . Such a rate is known as a spectral convergence order [64, 46].

## 2.6 Simulation studies

In Example 2.5.1, we have shown that if  $\Psi$  and  $\Phi$  are Matérn kernels with smoothness parameters  $\nu_0$  and  $\nu$ , respectively, and  $1 < \nu \leq \nu_0$ , then the kriging predictor converges with a rate at least  $O_p(h_{\mathbf{X}}^\nu \log^{1/2}(1/h_{\mathbf{X}}))$ . In this section we report simulation studies that verify that this rate is sharp, i.e., the true convergence rate coincides with that given by the theoretical upper bound.

We denote the expectation of the left-hand side of (2.13) by  $\mathcal{E}$ . If the error bound (2.13) is sharp, we have the approximation

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

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

$$\log \mathcal{E} \approx \nu \log h_{\mathbf{X}} + \frac{1}{2} \log(-\nu \log h_{\mathbf{X}}) + \log c. \quad (2.15)$$

Since  $\log(-\nu \log h_{\mathbf{X}})$  is much smaller than  $\log h_{\mathbf{X}}$ , the effect of  $\log(-\nu \log h_{\mathbf{X}})$  is negligible in (2.15). Consequently, we get our second approximation

$$\log \mathcal{E} \approx \nu \log h_{\mathbf{X}} + \log c. \quad (2.16)$$

As shown in (2.16),  $\log \mathcal{E}$  is approximately a linear function in  $\log h_{\mathbf{X}}$  with slope  $\nu$ . Therefore, to assess whether (2.13) is sharp, we should verify if the regression coefficient (slope) of  $\log \mathcal{E}$  with respect to  $\log h_{\mathbf{X}}$  is close to  $\nu$ .

In our simulation studies, the experimental region is chosen to be  $\Omega = [0, 1]^2$ . To estimate the regression coefficient  $\nu$  in (2.16), we choose 50 different maximin Latin hypercube designs [5] with sample sizes  $10k$ , for  $k = 1, 2, \dots, 50$ . Note that each design corresponds to a specific value of the fill distance  $h_{\mathbf{X}}$ . For each  $k$ , we simulate the Gaussian processes 100 times to reduce the simulation error. For each simulated Gaussian process, we compute  $\sup_{\mathbf{x} \in \Omega_1} |Z(\mathbf{x}) - \mathcal{I}_{\Phi, \mathbf{X}} Z(\mathbf{x})|$  to approximate the sup-error  $\sup_{\mathbf{x} \in \Omega} |Z(\mathbf{x}) - \mathcal{I}_{\Phi, \mathbf{X}} Z(\mathbf{x})|$ , where  $\Omega_1$  is the set of grid points with grid length 0.01. This should give a good approximation since the grid is dense enough. Next, we calculate the average of  $\sup_{\mathbf{x} \in \Omega_1} |Z(\mathbf{x}) - \mathcal{I}_{\Phi, \mathbf{X}} Z(\mathbf{x})|$  over the 100 simulations to approximate  $\mathcal{E}$ . Then the regression coefficient is estimated using the least squares method.

We conduct four simulation studies with different choices of the true and imposed smoothness of the Matérn kernels, denoted by  $\nu_0$  and  $\nu$ , respectively. Their values are shown in Table 2.1.

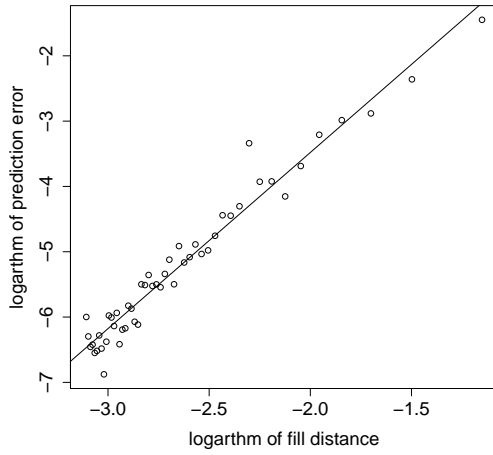
Figure 2.1 shows the relationship between the logarithm of the fill distance (i.e.,  $\log h_{\mathbf{X}}$ ) and the logarithm of the average prediction error (i.e.,  $\log \mathcal{E}$ ) in scatter plots for the four cases. The solid line in each panel shows the linear regression fit calculated from the data.

We summarize the results in Table 2.1.

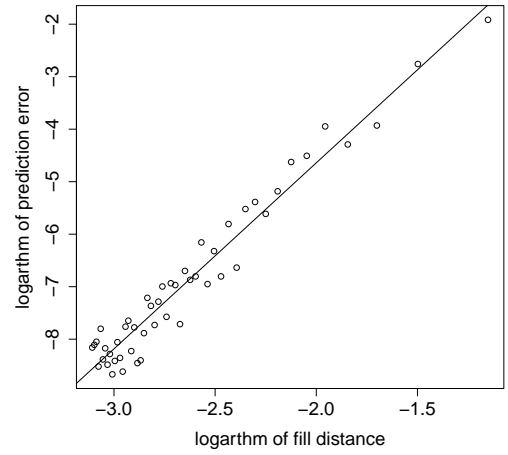
**Table 2.1:** Numerical studies on the convergence rates of kriging prediction. The first two columns show the true and imposed smoothness parameters of the Matérn kernels. The third column shows the convergence rate obtained from the simulation. The fourth column shows the convergence rate given by Theorem 2.5.1. The last column shows the relative difference between the third and the fourth columns, given by  $|\text{Regression coefficient} - \text{Theoretical assertion}| / (\text{Theoretical assertion})$ .

$\nu_0$	$\nu$	Regression coefficient	Theoretical assertion	Relative difference
3	2.5	2.697	2.5	0.0788
5	3.5	3.544	3.5	0.0126
3.5	3.5	3.582	3.5	0.0234
5	5	4.846	5	0.0308

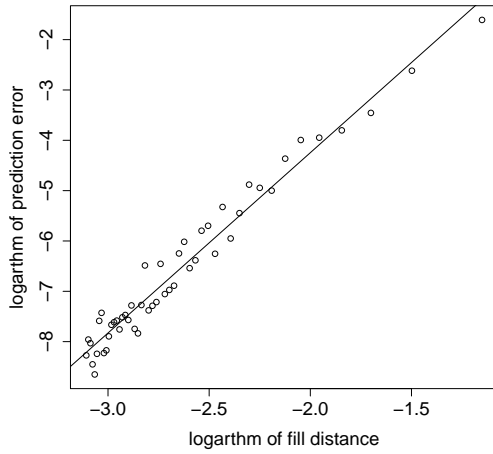
From Figure 2.1, it can be seen that, as the fill distance decreases, the supremum of the



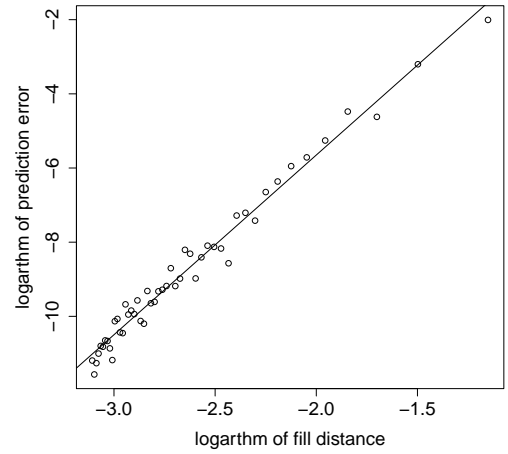
**(a)**  $\nu_0 = 3, \nu = 2.5.$



**(b)**  $\nu_0 = 5, \nu = 3.5.$



**(c)**  $\nu_0 = \nu = 3.5.$



**(d)**  $\nu_0 = \nu = 5.$

**Figure 2.1:** The regression line of  $\log \sup_{\mathbf{x} \in \Omega} \epsilon(\mathbf{x})$  on  $\log h_{\mathbf{X}}$ . Each point denotes one average prediction error for each  $n$ .

kriging prediction error also decreases. From the results in Table 2.1, the regression coefficients are close to the values given by our theoretical analysis, with relative error no more than 0.08. By comparing the third and the fourth rows of Table 2.1, we find that the regression coefficient does not have a significant change when  $\nu$  remains the same, even if  $\nu_0$  changes. On the other hand, the third and the fifth rows show that, the regression coefficient changes significantly as  $\nu$  changes, even if  $\nu_0$  keeps unchanged. This shows convincingly that the convergence rate is independent of the true smoothness of the Gaussian process, and the rate given by Theorem 2.5.1 is sharp. Note that our simulation studies justify the use of the leading term  $\log h_{\mathbf{X}}$  in (2.15) to assess the convergence rate but they do not cover the second term  $\log(-\nu \log h_{\mathbf{X}})$ , which is of lower order.

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

## 2.7 Conclusions and Discussion

We first summarize the statistical implications of this work. We prove that the kriging predictive error converges to zero under a uniform metric, which justifies the use of kriging as a function reconstruction tool. Kriging with a misspecified correlation function is also studied. Theorem 2.5.1 shows that there is a tradeoff between the predictive efficiency and the robustness. Roughly speaking, a less smooth correlation function is more robust against model misspecification. However, we shall lose some predictive efficiency by gaining robustness. With the help of the classic results in radial basis function approximation (in Theorems 2.4.1 and 2.4.2), we find that the predictive error of kriging is associated with the fill distance, which is a space-filling measurement of the design. This justifies the use of space-filling designs for (stationary) kriging models.

Theorem 2.5.1 shows that the predictive error is bounded by a function of  $P_{\Phi, \mathbf{X}}$ . This

inspires us to construct designs using the criterion that minimizes  $P_{\Phi, \mathbf{X}}$ . In fact, it can be proven that, under certain regularity conditions, the first part of Theorem 2.5.1 is still true for non-stationary Gaussian process models if  $\Phi = \Psi$ . Hence this construction of designs can be particularly useful if a non-stationary Gaussian process model is adopted (e.g., [65, 66, 67, 68]). It is known that space-filling designs are justifiable if the underlying Gaussian process is stationary. Therefore, when a non-stationary Gaussian process model is used, it may not be appropriate to continue using space-filling designs. A more general construction of design points should be studied in the non-stationary situation.

In this paper, we only consider Gaussian process models with mean zero, which is referred to as the simple kriging. A natural extension of this work is to include the Gaussian process models with a mean function modeled as a linear combination of a finite set of functions, known as the universal kriging. In this situation, one would consider the best linear unbiased predictor (BLUP) instead of the conditional expectation. The mathematical treatments to obtain new asymptotic theorem is more cumbersome. But we believe that the main idea of this work is still valid, and the general message of the theory remains the same.

We have proved in Theorem 2.5.1 that the kriging predictor is consistent if the true correlation function is smoother than the imposed correlation function. [55] proved that kriging with any Matérn correlation function achieves predictive consistency for any stationary Gaussian processes, although they did not derive the rate of convergence. In light of this result, we may consider extensions of Theorem 2.5.1 in a future work.

In this work, we suppose that the kriging interpolant can be computed exactly. However, this cannot be achieved in reality due to the limit of the machine precision. Specifically, the matrix inversion in (2.5) can be numerically unstable, especially when a Gaussian kernel is used. Thus the numerical error of kriging is generally non-negligible in practice. We refer to [9] for some related theoretical studies. A standard technique to eliminate the numerical instability in kriging is to introduce a nugget term [69, 70]. The convergence theory for

kriging with a nugget term as a numerical stabilizer requires a separate development.

There is a series of papers by [56, 57, 58, 71, 59] investigating the asymptotic efficiency of the kriging predictor. The theory in this work does not yield the assertions about prediction efficiency, although we provide explicit error bounds for kriging predictors with scattered design points in an arbitrary dimension.

Another important topic is the kriging predictive performance when the correlation function is estimated. In this paper, we consider kriging with a misspecified but fixed kernel function. It is shown that the prediction error can be minimized if the imposed kernel has the same smoothness as the true correlation function. A natural question is whether the optimal rate of convergence can be achieved by using a data-driven approach.

**CHAPTER 3**  
**SMOOTHNESS ESTIMATION AND ADAPTIVE KERNEL RIDGE**  
**REGRESSION**

**3.1 Introduction**

In non-parametric regression, the goal is to estimate an underlying function based on its noisy evaluations. One important class of non-parametric regression methods, called the kernel ridge regression, proceeds by minimizing a loss function involving the norm in a reproducing kernel Hilbert space as a regularization term. Some special forms of the kernel ridge regression, like the smoothing splines and the thin-plate smoothing splines, are known for a long time [45]. We refer to [72] for a general discussion of the kernel ridge regression. This methodology has been applied to many areas, including machine learning [73], spatial statistics [74] and biostatistics [75].

The optimal rate of convergence for non-parametric regression is determined by the smoothness of the underlying function [76]. In most practical scenarios, the true smoothness of the underlying function is unknown. This explains why we should consider estimating the smoothness of the underlying function from the data. In addition, estimating the smoothness is also of interest in its own right, because the smoothness itself is an important perspective of a surface in many scientific and engineering contexts [77, 26, 78]. Some estimators of the smoothness are proposed in [79, 80, 81, 82, 77]. However, these smoothness estimators suffer from some deficiencies. They may be subject to strong restrictions on the region of interest or the true smoothness of the underlying function, or may be complicated to compute.

Also, we are interested in obtaining non-parametric estimators of the underlying functions, which can achieve the optimal rate without knowing the true smoothness in advance.

Such estimators are known as *adaptive* ones in the literature [83]. Adaptive estimators have been constructed via kernel estimates [84], thresholding [85, 86], estimators of regularity based on process increments [79, 82] and rescaling a smooth Gaussian random field [83]. Such estimators are usually within a hypercube or regions within the Euclidean space no more than two dimensional. In particular, these estimators do not naturally deduce estimators for the smoothness, although the problems of the smoothness estimation and the adaptive regression are conceptually related. Detailed discussions will be given in Section 3.3.3.

In this work, we propose a method that estimates the underlying function and its smoothness simultaneously. This approach is motivated by the Gaussian process regression method, which has a natural connection with the kernel ridge regression from the computational point of view [45]. The smoothness of a Gaussian process can be parametrized by the smoothness parameter. In the literature, the smoothness parameter is usually estimated using the maximum likelihood estimation [6, 5, 26]. However, to the best of our knowledge, there is no theoretical guarantee of the maximum likelihood estimation of the smoothness. In this article, we propose a new smoothness estimator by maximizing a modified likelihood function. This estimator is proven to be consistent. In addition, we prove that the kernel ridge regression estimator using the estimated kernel function is consistent with a nearly optimal rate of convergence. Compared to the existing methods, the theoretical results for the proposed method are more general and require milder regularity conditions.

The rest of this paper is organized as follows. In Section 3.2, we introduce the problem formulation and the proposed method. In Section 3.3, we state the main results on the consistency and the adaptiveness. Comparison to existing methods is also given in this section. The technical proofs are given in Section 3.4.

## **3.2 Methodology**

In this section, we introduce the problem of interest and the proposed methodology.

### 3.2.1 Problem Setting

Suppose  $f$  is an underlying function defined on a convex and compact set  $\Omega \subset \mathbb{R}^d$  with a positive Lebesgue measure. We assume that the observations are obtained by random sampling. Specifically, we observe pairs  $(x_i, y_i), i = 1, \dots, n$ , given by

$$y_i = f(x_i) + e_i, \quad i = 1, \dots, n, \quad (3.1)$$

where  $x_i$ 's are independent samples from the uniform distribution over  $\Omega$ , and  $e_i$ 's are the measurement error. In this work, we suppose that  $e_i$ 's are independent and identically distributed random variables with zero mean and finite variance. Problems of this kind are encountered in areas such as nonparametric statistics [87, 88, 45], spatial statistics [51, 2, 89], and machine learning [90, 91, 6].

In this article, we mainly concern about two questions. First, we would like to estimate the smoothness of the underlying function  $f$ . For a moment, we loosely say that “a function has smoothness  $m$ ” means that it has  $\lfloor m \rfloor$ -th derivatives but is not  $(\lfloor m \rfloor + 1)$ -th differentiable. Interpreting the non-integer part of the smoothness requires most technical details, which will be given in Section 3.3.1. The second objective of this work is to obtain an estimator of  $f$ , which is a nearly optimal for all degrees of smoothness. An estimator of this kind is called *adaptive* [83].

### 3.2.2 Proposed Method

Before introducing the proposed estimators, we review the kernel ridge regression method. Consider the nonparametric model (3.1). Given  $m > d/2$  and a kernel function  $\Psi_m$  with smoothness  $m$ , the kernel ridge regression reconstruct  $f$  using

$$\hat{f}_m = \operatorname{argmin}_{\hat{f} \in \mathcal{N}_{\Psi_m}(\Omega)} \left( \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \frac{\mu_m}{n} \|\hat{f}\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \right), \quad (3.2)$$

where  $\mu_m$  is the smoothing parameter, and  $\mathcal{N}_{\Psi_m}(\Omega)$  is the reproducing kernel Hilbert space generated by the kernel function  $\Psi_m$ . Under certain conditions, the optimal order of magnitude of  $\mu_m$  is known in the literature [92], given by

$$\mu_m = Cn^{\frac{d}{2m+d}} \quad (3.3)$$

with a constant  $C > 0$ . A prominent class of kernel functions with finite smoothness is the (isotropic) Matérn family [26], after a proper reparametrization, defined as

$$\Psi_m(x) = \frac{1}{\Gamma(m - d/2)2^{m-d/2-1}} \|x\|^{m-d/2} K_{m-d/2}(\|x\|), \quad (3.4)$$

where  $\|\cdot\|$  denotes the Euclidean distance;  $K_{m-d/2}$  is the modified Bessel function of the second kind.

To estimate the true smoothness of the underlying function  $f$  from the data  $X = (x_1, \dots, x_n)^T$  and  $Y = (y_1, \dots, y_n)^T$ , we first define the loss function

$$\ell(m; X, Y, \mu_m) = \frac{mn}{2m+d} \log n - \frac{n}{2} \log(Y^T (K_m + \mu_m I_n)^{-1} K_m (K_m + \mu_m I_n)^{-1} Y), \quad (3.5)$$

where  $K_m := (\Psi_m(x_i - x_j))_{ij}$  is the kernel matrix;  $I_n$  denotes the identity matrix; and  $\mu_m$  is given by (3.3). Now we propose to estimate the smoothness of  $f$  using

$$\hat{m}_n := \operatorname{argmax}_{m > d/2} \ell(m; X, Y, \mu_m). \quad (3.6)$$

Now we turn to the estimation of  $f$ . By plugging (3.6) into (3.2), we suggest using

$$\hat{f}_{\hat{m}_n} = \operatorname{argmin}_{\hat{f} \in \mathcal{N}_{\Psi_{\hat{m}_n}}(\Omega)} \left( \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \frac{\mu_{\hat{m}_n}}{n} \|\hat{f}\|_{\mathcal{N}_{\Psi_{\hat{m}_n}}(\Omega)}^2 \right) \quad (3.7)$$

to recover the underlying function  $f$ . It can be shown that (3.7) is equivalent to

$$\hat{f}_{\hat{m}_n}(x) = \sum_{i=1}^n u_i \Psi_{\hat{m}_n}(x - x_i), \quad (3.8)$$

with  $u = (u_1, \dots, u_n)^T$  determined by the linear system  $Y = (K_{\hat{m}_n} + \mu_{\hat{m}_n} I_n)u$  [88].

We shall call  $\hat{m}_n$  the smoothness estimator and  $\hat{f}_{\hat{m}_n}$  the function estimator. We shall call  $\ell(m; X, Y, \mu_m)$  the modified likelihood function, because  $\ell$  is related to the likelihood function of a Gaussian process model with a Matérn correlation function. Details are given in the next subsection.

### 3.2.3 Some Intuition and Related Methods

The proposed method is partially inspired by the Gaussian process modeling technique, which has been used to recover unknown functions in the areas such as spatial statistics [2, 3], computer experiments [4, 5] and machine learning [6, 7].

Again, consider the regression model given by (3.1). The main idea of Gaussian process modeling is to assume the underlying function is a realization of a Gaussian random field. Specifically, let  $Z(x)$  be a Gaussian random field. It is known that the law of a Gaussian random field is governed by its mean and covariance. We assume that the mean of  $Z(x)$  is zero, and the covariance function is given by a Matérn kernel, i.e.,  $\text{Cov}(Z(x_1), Z(x_2)) = \sigma^2 \Psi_m(x_1 - x_2)$ , where  $\sigma^2$  is the variance and  $\Psi_m$  is as in (3.4). The current use of the Matérn kernel is not much different from a more general case where the (isotropic) Matérn correlation family is indexed by (fixed) scale parameters [26], because we can stretch the region  $\Omega$  to adjust the scale parameters at will.

We use the Matérn correlation family because the parameter  $m$  can determine the smoothness of the associated Gaussian process. It is known that a stationary Gaussian process with a Matérn correlation function in (3.4) has  $p$  times almost surely continuously differentiable sample paths if and only if  $m > p + 1/2$  when  $d = 1$  [49]. In view of this

fact, we would ask whether and how the smoothness of the underlying function can be estimated from the data with the help of the Matérn kernels. It is natural to first consider the maximum likelihood estimate of the Gaussian process models with Matérn correlation functions.

To obtain a tractable likelihood function, we postulate that  $e_i$  follows the normal distribution  $N(0, \mu)$  with  $\mu > 0$ . Here we remark that this assumption does not need to be true in reality. As in a standard Gaussian process model, we assume  $f$  is a realization of a stationary Gaussian process with mean zero, variance  $\sigma^2$  and correlation function  $\Psi_m$  with  $m > d/2$ .

Recall that one of our goal is to estimate the smoothness of the underlying function. That is, to estimate the smoothness parameter from the data. The maximum likelihood method is a widely used method in Gaussian process modeling to estimate unknown parameters [5]. Here we consider using the maximum likelihood method to estimate the smoothness parameter  $m$  from the data  $X = (x_1, \dots, x_n)^T$  and  $Y = (y_1, \dots, y_n)^T$ . Direct calculations show that, up to an additive constant, the log-likelihood function is

$$\begin{aligned} \ell_1(m, \sigma^2; X, Y, \mu) & \quad (3.9) \\ &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2} \log \det(K_m + \mu I_n) - \frac{1}{2\sigma^2} Y^T (K_m + \mu I_n)^{-1} Y, \end{aligned}$$

where  $K_m := (\Psi_m(x_i - x_j))_{ij}$  is the correlation matrix; and  $I_n$  denotes the identity matrix. We refer to [2, 6, 5] for the maximum likelihood method of Gaussian process models.

It is easily seen that, given  $m$ , the maximizer of (3.9) with respect to  $\sigma^2$  is  $\hat{\sigma}^2 = Y^T (K_m + \mu I_n)^{-1} Y / n$ . Substituting  $\hat{\sigma}$  into (3.9), we obtain the profile likelihood function with respect to  $m$  given by

$$\ell_2(m; X, Y, \mu) = -\frac{1}{2} \log(\det(K_m + \mu I_n)) - \frac{n}{2} \log(Y^T (K_m + \mu I_n)^{-1} Y), \quad (3.10)$$

It is reasonable to believe that the minimizer of (3.10) gives a consistent smoothness

estimator. But unfortunately, we cannot proof this result unless we make a modification on the loss function and use

$$\begin{aligned} \ell_2(m; X, Y, \mu) = & -\frac{1}{2} \log(\det(K_m + \mu_m I_n)) \\ & -\frac{n}{2} \log(Y^T (K_m + \mu_m I_n)^{-1} K_m (K_m + \mu_m I_n)^{-1} Y). \end{aligned} \quad (3.11)$$

The second modification we make is to replace the first term in (3.11) to  $(mn)/(2m + d) \log n$  as in (3.5), because they are asymptotically equivalent. See Appendix C.13. This modification reduces the computational cost by waiving the determinant calculation.

### 3.3 Theoretical Results

In this section we present our main theoretical results on the consistency of the smoothness estimator and the nearly optimality of the function estimator. Some comparison with existing results is also provided in this section.

#### 3.3.1 Mathematical Formulation of Smoothness

Before introducing our asymptotic results, we first formalize our notion of smoothness. We define the smoothness of a function using the order of the corresponding (fractional) Sobolev space. Let  $\Omega$  be a subset of  $\mathbb{R}^d$ . For a multi-index  $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ , define its length by  $|\alpha| = \alpha_1 + \dots + \alpha_d$ . Denote  $\alpha$ -th (weak) derivative of a function  $f$  by  $D^\alpha u := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} u$ . For a positive integer  $k$ , the Sobolev space  $H^k(\Omega)$  consists of functions  $u \in L_2(\Omega)$  such that  $\|u\|_{H^k(\Omega)}^2 := \sum_{|\alpha| \leq k} \|D^\alpha u\|_{L_2(\Omega)}^2$  is finite.

For an integrable function  $f$  on  $\mathbb{R}^d$ , its Fourier transform is defined as

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

and with the help of some functional analysis machinery, this definition can be naturally extended to all  $f \in L_2(\mathbb{R}^d)$ . See [93] for the details about this extension.

For  $\Omega = \mathbb{R}^d$ , the Sobolev norm can be expressed using the Fourier transform:

$$\|u\|_{H^k(\mathbb{R}^d)}^2 = \int_{\mathbb{R}^d} |\mathcal{F}(u)(\omega)|^2 (1 + \|\omega\|^2)^k d\omega. \quad (3.12)$$

Identity (3.12) allows us to define the Sobolev spaces with non-integer orders, which are commonly known as the *fractional Sobolev spaces*, denoted by  $H^m(\mathbb{R}^d)$  with a non-integer  $m$ . We refer to [94] for more discussions about Sobolev spaces.

The Fourier transform of the Matérn kernel is [95]

$$\mathcal{F}(\Psi_m)(\omega) = 2^{d/2} \frac{\Gamma(m)}{\Gamma(m - d/2)} (1 + \|\omega\|^2)^{-m}. \quad (3.13)$$

By comparing (3.12) and (3.13), it can be seen that the Sobolev norm is proportional to  $\int_{\mathbb{R}^d} |\mathcal{F}(u)(\omega)|^2 / \mathcal{F}(\Psi_m)(\omega) d\omega$ , which is the norm of the reproducing kernel Hilbert space generated by  $\Psi_m$  [46]. In addition, this relationship implies that  $\Psi_m$  is a reproducing kernel of  $H^m(\mathbb{R}^d)$  up to a constant.

In this article, we say that a function  $u \in L_2(\mathbb{R}^d)$  has a finite degree of smoothness if the quantity

$$\sup\{k \geq 0 : u \in H^k(\mathbb{R}^d)\} \quad (3.14)$$

is finite. We call the quantity (3.14) the smoothness of  $u$ , and we are only interested in the functions with smoothness greater than  $d/2$ , which guarantees the continuity of the function according to the Sobolev embedding theorem [94].

Fractional Sobolev spaces over a bounded region  $\Omega$ , denoted by  $H^k(\Omega)$ , can be defined by the restriction of functions in  $H^k(\mathbb{R}^d)$ , if  $\Omega$  is not too complex, for example, if  $\Omega$  is convex. On the other hand, if  $\Omega$  is convex, there exists an extension operator from  $L_2(\Omega)$  to  $L_2(\mathbb{R}^d)$ , such that the smoothness of each function is maintained [96]. For each  $u \in L_2(\Omega)$ , denote its extended function to the whole space through the proceeding operator by  $u_e \in$

$L_2(\mathbb{R}^d)$ . Hence, we can define the smoothness of a function  $u \in L_2(\Omega)$  by the smoothness of  $u_e \in L_2(\mathbb{R}^d)$  using (3.14).

Clearly, given a function  $u$  with finite smoothness  $m_0$ , there are two cases: 1)  $u \in H^{m_0}(\Omega)$  but  $u \notin H^{m'}(\Omega)$  for any  $m' > m_0$ ; 2)  $u \in H^{m'}(\Omega)$  for any  $m' < m_0$  but  $u \notin H^{m_0}(\Omega)$ . We differentiate these two cases by saying that  $u$  is of type-I or of type-II, respectively.

To the best of our knowledge, the existing work on the smoothness estimation and the adaptive estimation only consider underlying functions of type-I. In this work, we will prove the consistency of proposed smoothness estimator in Section 3.3.2 for both types of functions. The rate of convergence will also be given.

### 3.3.2 Main Theorems

Recall that in Section 3.2.1 we mentioned two objectives. The first objective is to estimate the smoothness of the underlying function, and the second objective is to obtain an estimator of  $f$  which is adaptive. In this section we provide theoretical justification of the smoothness estimator given by (3.15) and the estimator of  $f$  given by (3.17) can achieve these two objectives, respectively.

In order to show the consistency of the smoothness estimator, we assume the true smoothness  $m_0 \in [m_{\min}, m_{\max}]$ , where  $m_{\min}, m_{\max} > d/2$  are known. In this case, the smoothness estimator of  $f$  becomes

$$\hat{m}_n := \operatorname{argmax}_{m \in [m_{\min}, m_{\max}]} \ell(m; X, Y, \mu_m), \quad (3.15)$$

and the function estimator is

$$\hat{f}_{\hat{m}_n} = \operatorname{argmin}_{\hat{f} \in \mathcal{N}_{\Psi_{\hat{m}_n}}(\Omega)} \left( \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \frac{\mu_{\hat{m}_n}}{n} \|\hat{f}\|_{\mathcal{N}_{\Psi_{\hat{m}_n}}(\Omega)}^2 \right) \quad (3.16)$$

which is equivalent to

$$\hat{f}_{\hat{m}_n}(x) = \sum_{i=1}^n u_i \Psi_{\hat{m}_n}(x - x_i), \quad (3.17)$$

with  $u = (u_1, \dots, u_n)^T$  determined by the linear system  $Y = (K_{\hat{m}_n} + \mu_{\hat{m}_n} I_n)u$ .

The technical assumption on the known  $m_{\min}, m_{\max}$  values should be mild in many practical situations, because we can choose  $m_{\max}$  sufficiently large and  $m_{\min}$  sufficiently close to  $d/2$ . We believe that the general smoothness estimator (3.6) is also consistent like the constrained version in (3.15). However, it requires extra efforts to complete such a proof, and needs a separate development.

First, we introduce the following Lemma 3.3.1, which defines some helpful concepts in the description of the rate of convergence.

**Lemma 3.3.1.** *Let  $m_0 \in (d/2, +\infty)$  be the smoothness of  $g$ . If  $g \in H^{m_0}(\mathbb{R}^d)$ , then there exists an increasing positive function  $h_1$  on  $[0, \infty)$  such that*

$$\begin{aligned} \int_{\mathbb{R}^d} |\mathcal{F}(g)(\omega)|^2 h_1(\|\omega\|) (1 + \|\omega\|^2)^{m_0} d\omega &= \infty, \\ \int_{\mathbb{R}^d} |\mathcal{F}(g)(\omega)|^2 h_1(\|\omega\|) (1 + \|\omega\|^2)^{m_0 - \epsilon_1} d\omega &< \infty, \end{aligned} \quad (3.18)$$

and

$$\lim_{x \rightarrow +\infty} \frac{\log h_1(x)}{\log x} = 0, \quad (3.19)$$

for any  $\epsilon_1 > 0$ . If  $g \notin H^{m_0}(\mathbb{R}^d)$ , then there exists an increasing positive function  $h_2$  such that

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{|\mathcal{F}(g)(\omega)|^2}{h_2(\|\omega\|)} (1 + \|\omega\|^2)^{m_0} d\omega &< \infty, \\ \int_{\mathbb{R}^d} \frac{|\mathcal{F}(g)(\omega)|^2}{h_2(\|\omega\|)} (1 + \|\omega\|^2)^{m_0 + \epsilon_2} d\omega &= \infty, \end{aligned} \quad (3.20)$$

and

$$\lim_{x \rightarrow +\infty} \frac{\log h_2(x)}{\log x} = 0, \quad (3.21)$$

for any  $\epsilon_2 > 0$ .

The conditions (3.19) and (3.21) essentially require that  $h_1(x)$  and  $h_2(x)$  increase slower than any power function  $x^\epsilon$  with  $\epsilon > 0$ . The intuition behind Lemma is pretty clear. For example, consider  $f(x) = 1/x$ , we know that  $\int_1^{+\infty} f(x)dx = +\infty$  and  $\int_1^{+\infty} f(x)/x^\epsilon dx < +\infty$  for any  $\epsilon > 0$ . It is easily seen that, the function  $h(x) := \log^2 x$ , which increases slower than any power function  $x^\epsilon$  with  $\epsilon > 0$ , satisfies  $\int_1^{+\infty} f(x)/h(x)dx < +\infty$ . The proof for the general situation is more involved and contains only elementary mathematical analysis. We therefore send the proof of Lemma 3.3.1 to Appendix C.7.

We also need the following assumption on the errors, which means that the error is sub-Gaussian [92].

**Assumption 3.3.1.** *Suppose  $e_i$ 's in (3.1) are i.i.d. random variables satisfying*

$$C_1^2 (E e^{e_i^2/C_1^2} - 1) \leq \sigma_0^2 \quad (3.22)$$

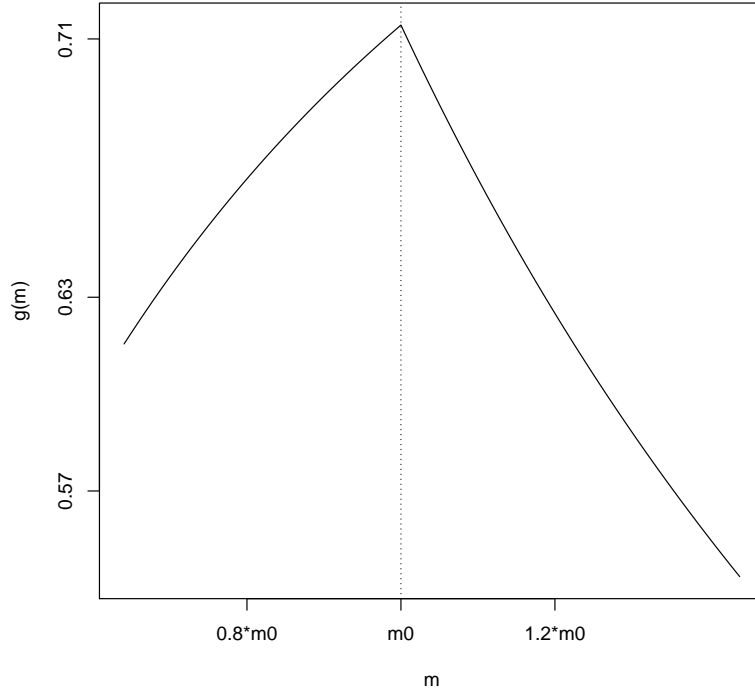
for some constant  $C_1$  and  $\sigma_0^2$ .

According to Theorem C.1.1 in Appendix C.1,  $\ell(m; X, Y, \mu)$  is closely related to the following function

$$g(m) := \begin{cases} \frac{2m}{2m+d}, & m \leq m_0, \\ \frac{2m_0}{2m+d}, & m > m_0, \end{cases}$$

which is maximized at  $m_0$ . See Figure 3.1 for an illustration with  $d = 0.8m_0$ .

With the bounds of  $\ell(m; X, Y, \mu)$ , we have the following theorem, which states the consistency of the smoothness estimator given by (3.15).



**Figure 3.1:** The plot of function  $g(m)$ , where  $d = 0.8m_0$ .

**Theorem 3.3.1.** *Suppose  $m_0$  is the true smoothness of the underlying function  $f$  defined on a compact and convex set  $\Omega \in \mathbb{R}^d$ . Suppose the errors satisfy Assumption 3.3.1. Fix any constant  $C > 0$  and let  $\mu_m = Cn^{\frac{d}{2m+d}}$ . Then for  $n > C_2$ , with probability at least  $1 - C_3 \exp(-C_4 n^{\eta_1})$ ,  $|\hat{m}_n - m_0| \leq 2d(2m_0 + d)^2 s_1$ , where*

$$s_1 = \begin{cases} \log(C_1 h_1(n) \log n) / \log n & \text{for } f \in H^{m_0}(\Omega), \\ \log(C_1 (\log n)^2 h_2(n)) / \log n & \text{for } f \notin H^{m_0}(\Omega), \end{cases}$$

$h_1, h_2$  are defined as in Lemma 3.3.2 and constants  $C_1, C_2, C_3, C_4$  and  $\eta_1$  are positive and depend on only  $f, \Omega$ , and  $C$ .

Theorem 3.3.1 shows the consistency of the smoothness estimator because the quantity  $s_1$  in Theorem 3.3.1 decays to zero because we have  $\log h_i(x) / \log x \rightarrow 0$  for  $i = 1, 2$  according to Lemma 3.3.2.

Now we turn to the estimator of  $f$  given by (3.17). In the following theorem, we show that this estimator is nearly optimal.

**Theorem 3.3.2.** *Suppose  $m_0$  and  $f$  are given in Theorem 3.3.1, and the errors satisfy Assumption 3.3.1. Set  $\mu_m$  as in Theorem 3.3.1. If  $f \in H^{m_0}(\Omega)$ , the estimator  $\hat{f}_{\hat{m}_n}$  given by (3.17) satisfies  $\|\hat{f}_{\hat{m}_n} - f\|_2 = O_{\mathbb{P}}(n^{-m_0/(2m_0+d)}(h_3(n)))$ , where  $h_3(n)$  depends on  $h_1(n)$  (defined in (3.18)) and  $m_{\max}$  if  $f \in H^{m_0}(\Omega)$ , and  $h_2(n)$  (defined in (3.20)) and  $m_{\max}$  if  $f \notin H^{m_0}(\Omega)$ . In both cases  $h_3(n)$  satisfies  $\lim_{n \rightarrow \infty} \frac{\log h_3(n)}{\log n} = 0$ .*

Recall that, when the true smoothness  $m_0$  is known, the optimal convergence rate is  $n^{-m_0/(2m_0+d)}$ . Theorem 3.3.2 implies that the estimator  $\hat{f}_n$  given by (3.2) is nearly optimal without knowing  $m_0$  in advance, because  $h_3$  increases slower than any  $n^\alpha$  with  $\alpha > 0$ .

### 3.3.3 Comparison with Existing Results

We compare the proposed method with the existing ones in the literature in the following directions.

First, the proposed method works on any compact and convex region  $\Omega$  within any dimension  $d$ , given the smoothness is greater than  $d/2$ . The previous methods of constructing the smoothness estimation are usually considered on regions within Euclidean space  $\mathbb{R}^d$  with  $d \leq 2$ . For instance, [79, 80, 82, 97] construct the smoothness estimator with equispaced data on a line transect, where the dimension is one. [98, 77] estimate the smoothness via quadratic variation, in which the area is within a compact domain in  $\mathbb{R}^d$  where  $d \in \{1, 2\}$ . Some of the adaptive function estimators have the same problem. For instance, [86, 85, 97, 99, 100] construct adaptive estimators on  $[0, 1]$  with equispaced data. [101] constructs an adaptive estimator based on the random data points on  $[0, 1]$ . [102, 83] are able to construct an adaptive estimator with the input space of the underlying function is  $[0, 1]^d$ , which is a hypercube, for  $d > 0$ .

Second, the proposed method is able to estimate any smoothness which is greater than  $d/2$ . In contrast with the proposed method, [79, 80, 82] estimate the smoothness within

$(1/2, 3/2)$ , while [81] assumes that the smoothness is within  $(D + 1/2, D + 3/2)$  for some known integer  $D$ .

Third, the conditions on the underlying function is milder than some of the adaptive function estimators. For instance, [102, 83, 99] consider the underlying function that is within some Hölder class, which is more restrictive than Sobolev spaces (?).

Last, we consider the two cases,  $f \in H^{m_0}(\Omega)$  and  $f \notin H^{m_0}(\Omega)$ , separately. The estimators are assuming that  $f \in H^{m_0}(\Omega)$ . To the best of our knowledge, our work is the first result obtaining the function estimator with convergence rate  $o_{\mathbb{P}}(n^{-m_0/(2m_0+d)+\epsilon})$  for any  $\epsilon > 0$ .

### 3.4 Proofs

In this section we prove Theorems 3.3.1 and 3.3.2. Some necessary lemmas are introduced, where the proofs are in Appendix. In this section and Appendix, we define the empirical norm of a function  $g$  as

$$\|g\|_n^2 = \frac{1}{n} \sum_{i=1}^n g(x_i)^2.$$

Let  $H_B(\delta_n, \mathcal{G}, \|\cdot\|_\infty)$  denote the bracket entropy number of the metric space  $(\mathcal{G}, \|\cdot\|_\infty)$ . For detailed discussion of the bracket entropy number, see [92]. Let  $e = (e_1, \dots, e_n)^T$ . For notational simplicity, we will use  $C_1, C_2, \dots$  and  $\eta_0, \eta_1, \dots$  to denote the constants, of which the values can change from line to line. We will use  $\langle \cdot, \cdot \rangle_n$  to denote the empirical inner product, which is defined by

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

for two functions  $f$  and  $g$ . In particular, let

$$\langle e, f \rangle_n = \frac{1}{n} \sum_{i=1}^n e_i f(x_i)$$

for a function  $f$ .

### 3.4.1 Proof of Theorem 3.3.1

This theorem is a direct corollary of Theorem C.1.1. More specifically, by taking the lower bound of  $l_3(m_0; X, Y, \mu)$  larger than the upper bound of  $l_3(m; X, Y, \mu)$  for all  $m \in [m_{\min}, m_0 - \epsilon_n] \cup [m_0 + \epsilon_n, m_{\max}]$ , we have the following results for  $n > C_0$ , where  $C_0$  is a constant.

**Case 1:**  $f \in H^{m_0}(\Omega)$ . Let  $\epsilon_n = \max\{(4m_0 s_1)/(1 - 2s_1), 2d(2m_0 + d)^2 s_2\}$ , where  $s_1 = \log(C_1 h_1(n) \log n) / \log n$ , and  $s_2 = \log(C_2 \log n) / \log n$ . With probability at least  $1 - C_3 \exp(-C_4 n^{\eta_1})$ ,  $\hat{m}_n \in [m_0 - \epsilon_n, m_0 + \epsilon_n]$ .

**Case 2:**  $f \notin H^{m_0}(\Omega)$ : Let  $\epsilon_n = \max\{(4m_0 s_1)/(1 - 2s_1), 2d(2m_0 + d)^2 s_2\}$ , where  $s_1 = \log(C_1 (\log n)^2 h_2(n)) / \log n$ , and  $s_2 = \log(C_2 \log n h_2(n)) / \log n$ . With probability at least  $1 - C_3 \exp(-C_4 n^{\eta_1})$ ,  $\hat{m}_n \in [m_0 - \epsilon_n, m_0 + \epsilon_n]$ .

By noting that: (i)  $1 - 2s_1 > 1/2$  and  $s_1 > s_2$  for  $n > C_5$  with some constant  $C_5$ , and (ii)  $m_0 > d/2 \geq 1/2$ , we can take  $\epsilon_n = 2d(2m_0 + d)^2 s_1$ , and obtain the desired results.

### 3.4.2 Proof of Theorem 3.3.2

Let  $t_{m_0 - \epsilon_n} = C_0 n^{-2(m_0 - \epsilon_n)/(2(m_0 - \epsilon_n) + d)} (\log n)^{2m_{\max}}$  if  $f \in H^{m_0}$ , and  $t = C_0 n^{-2(m_0 - \epsilon_n)/(2(m_0 - \epsilon_n) + d)} h_2(n)$  if  $f \notin H^{m_0}$ , where  $C_0$  is a constant, and  $\epsilon_n$  is as in Theorem 3.3.1. From the proof of Theorem C.1.1, it can be seen that

$$\frac{\mu_m}{n} Y^T (K_m + \mu_m I_n)^{-1} K_m (K_m + \mu_m I_n)^{-1} Y = \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2.$$

Let  $\lambda_1 = \mu_{\hat{m}_n}/n$  and  $\lambda_2 = C_1\lambda_1$ . Let  $f_{\hat{m}_n}^*$  be the minimizer of optimization problem

$$\min_{\hat{f} \in \mathcal{N}_{\Psi_{\hat{m}_n}}(\Omega)} \|f - \hat{f}\|_2^2 + \lambda_2 \|\hat{f}\|_{\mathcal{N}_{\Psi_{\hat{m}_n}}(\Omega)}^2.$$

By (3.16), we have

$$\begin{aligned} & \|f - \hat{f}_{\hat{m}_n}\|_n^2 + \lambda_1 \|\hat{f}_{\hat{m}_n}\|_{\mathcal{N}_{\Psi_{\hat{m}_n}}(\Omega)}^2 \\ & \leq 2\langle e, \hat{f}_{\hat{m}_n} - f_{\hat{m}_n}^* \rangle + \|f - f_{\hat{m}_n}^*\|_n^2 + \lambda_2 \|f_{\hat{m}_n}^*\|_{\mathcal{N}_{\Psi_{\hat{m}_n}}(\Omega)}^2 + \lambda_1 \|\hat{f}_{\hat{m}_n}\|_{\mathcal{N}_{\Psi_{\hat{m}_n}}(\Omega)}^2. \end{aligned}$$

By the similar approach in the proof of Theorem C.1.1, it can be shown that there exists a constant  $c_1$  such that when  $n > c_1$ ,

$$2\langle e, \hat{f}_{\hat{m}_n} - f_{\hat{m}_n}^* \rangle = n^{-\eta_1} t_{m_0 - \epsilon_n}, \quad (3.23)$$

$$\frac{\mu_{\hat{m}_n}}{n} \|\hat{f}_{\hat{m}_n}\|_{\mathcal{N}_{\Psi_{\hat{m}_n}}(\Omega)}^2 \leq t_{m_0 - \epsilon_n}, \quad (3.24)$$

and

$$\|f - f_{\hat{m}_n}^*\|_n^2 + \lambda_2 \|f_{\hat{m}_n}^*\|_{\mathcal{N}_{\Psi_{\hat{m}_n}}(\Omega)}^2 \leq t_{m_0 - \epsilon_n}. \quad (3.25)$$

with probability at least  $1 - C_2 \exp(-C_3 n^{\eta_2})$ , where  $C_2$ ,  $C_3$ , and  $\eta_i$ 's are positive constants. Therefore, combining (3.23), (3.24) and (3.25), we have that with probability at least  $1 - C_4 \exp(-C_5 n^{\eta_3})$ ,

$$\|f - \hat{f}_{\hat{m}_n}\|_2^2 \leq t_{m_0 - \epsilon_n}.$$

Since  $\epsilon_n$  converges to zero, there exists a constant  $c_2$  such that when  $n > c_2$ ,  $\epsilon_n < 2m_0$ .

Therefore, by direct calculation, it can be shown that

$$\begin{aligned} t_{m_0-\epsilon_n} &= n^{-2m_0/(2m_0+d) + \frac{2d\epsilon_n}{(2m_0+d)(2m_0-2\epsilon_n+d)}} (\log n)^{2m_{\max}}, \\ &\leq n^{-2m_0/(2m_0+d)} (n^{\epsilon_n})^{\frac{2}{2m_0+d}} (\log n)^{2m_{\max}} \end{aligned}$$

By taking  $h_3(n) = (n^{\epsilon_n})^{\frac{2}{2m_0+d}} (\log n)^{2m_{\max}}$ , we finish the proof of Theorem 3.3.2.

# Appendices

## APPENDIX A

### APPENDIX OF CHAPTER 1

#### A.1 Proof of Theorem 1.3.1

Consider a location of interest  $x \in \Omega$  and the nearest *design point*  $x_i \in \bar{X}$ . The uppermost terms in (1.4) can be expressed as

$$\begin{aligned}
 & \Psi_\theta(x, x) - \Psi_\theta(x, \bar{X})[\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1}\Psi_\theta(\bar{X}, x) \\
 &= \Psi_\theta(x, x) \\
 & \quad - [(\Psi_\theta(x, \bar{X}) - \Psi_\theta(x_i, \bar{X}) - \sigma_i^2 e_i^T)[\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1}(\Psi_\theta(\bar{X}, x) - \Psi_\theta(\bar{X}, x_i) - \sigma_i^2 e_i) \\
 & \quad + 2(\Psi_\theta(x_i, \bar{X}) + \sigma_i^2 e_i^T)[\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1}\Psi_\theta(\bar{X}, x) \\
 & \quad - (\Psi_\theta(x_i, \bar{X}) + \sigma_i^2 e_i^T)[\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1}(\Psi_\theta(\bar{X}, x_i) + \sigma_i^2 e_i)] \\
 &= \Psi_\theta(x, x) - 2e_i^T \Psi_\theta(\bar{X}, x) + e_i^T (\Psi_\theta(\bar{X}, x_i) + \sigma_i^2 e_i) \\
 & \quad - (\Psi_\theta(x, \bar{X}) - \Psi_\theta(x_i, \bar{X}) - \sigma_i^2 e_i^T)[\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1}(\Psi_\theta(\bar{X}, x) - \Psi_\theta(\bar{X}, x_i) - \sigma_i^2 e_i) \\
 &= \Psi_\theta(x, x) + \Psi_\theta(x_i, x_i) + \sigma_i^2 - 2\Psi_\theta(x_i, x) \\
 & \quad - (\Psi_\theta(x, \bar{X}) - \Psi_\theta(x_i, \bar{X}) - \sigma_i^2 e_i^T)[\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1}(\Psi_\theta(\bar{X}, x) - \Psi_\theta(\bar{X}, x_i) - \sigma_i^2 e_i),
 \end{aligned} \tag{A.1}$$

where  $e_i$  denotes the  $i^{\text{th}}$  column of an  $n \times n$  identity matrix. The fourth term on the right-hand side of (A.1) can be bounded as

$$\begin{aligned}
& - (\Psi_\theta(x, \bar{X}) - \Psi_\theta(x_i, \bar{X}) - \sigma_i^2 e_i^T) [\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1} (\Psi_\theta(\bar{X}, x) - \Psi_\theta(\bar{X}, x_i) - \sigma_i^2 e_i) \\
\leq & - \frac{\|\Psi_\theta(\bar{X}, x) - \Psi_\theta(\bar{X}, x_i) - \sigma_i^2 e_i\|_2^2}{\lambda_{\max}[\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]} \\
\leq & - \frac{(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i) - \sigma_i^2)^2}{\lambda_{\max}[\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]} \\
\leq & - \frac{(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i))^2 - 2\sigma_i^2(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i)) + \sigma_i^4}{\lambda_{\max}[\Psi_\theta(\bar{X}, \bar{X})] + \lambda_{\max}(\bar{\Sigma}_\epsilon)} \\
\leq & - \frac{(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i))^2 - 2\sigma_i^2(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i)) + \sigma_i^4}{n \sup_{u,v \in \Omega} \Psi_\theta(u, v) + \lambda_{\max}(\bar{\Sigma}_\epsilon)} \\
= & - \frac{(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i))^2}{n \sup_{u,v \in \Omega} \Psi_\theta(u, v) + \lambda_{\max}(\bar{\Sigma}_\epsilon)} + \frac{2\sigma_i^2(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i)) - \sigma_i^4}{n \sup_{u,v \in \Omega} \Psi_\theta(u, v) + \lambda_{\max}(\bar{\Sigma}_\epsilon)}, \tag{A.2}
\end{aligned}$$

where the first inequality is true because for any vector  $d$  and matrix  $G$ ,  $d^T G^{-1} d \geq \lambda_{\min}(G^{-1}) \|d\|_2^2$  and  $\lambda_{\min}(G^{-1}) = 1/\lambda_{\max}(G)$ , the second inequality is true because the sum of squares  $\|\cdot\|_2^2$  is larger than any one of its elements squared, the third inequality is true because the maximum eigenvalue of a sum is at most the sum of the maximum eigenvalues, and the final inequality is true because Gershgorin's theorem [25] implies

$$\lambda_{\max}(\Psi_\theta(\bar{X}, \bar{X})) \leq \max_j \sum_{i=1}^n \Psi_\theta(x_i, x_j) \leq n \sup_{u,v \in \Omega} \Psi_\theta(u, v). \tag{A.3}$$

Combining (A.1) and (A.2) gives

$$\begin{aligned}
& \Psi_\theta(x, x) - \Psi_\theta(x, \bar{X}) [\Psi_\theta(\bar{X}, \bar{X}) + \bar{\Sigma}_\epsilon]^{-1} \Psi_\theta(\bar{X}, x) \\
\leq & \Psi_\theta(x, x) + \Psi_\theta(x_i, x_i) - 2\Psi_\theta(x_i, x) \\
& - \frac{(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i))^2}{n \sup_{u,v \in \Omega} \Psi_\theta(u, v) + \lambda_{\max}(\bar{\Sigma}_\epsilon)} + \sigma_i^2 + \frac{2\sigma_i^2(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i)) - \sigma_i^4}{n \sup_{u,v \in \Omega} \Psi_\theta(u, v) + \lambda_{\max}(\bar{\Sigma}_\epsilon)}. \tag{A.4}
\end{aligned}$$

Consider the concave, quadratic function

$$f_1(t) = t + \frac{2t(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i)) - t^2}{n \sup_{u,v \in \Omega} \Psi_\theta(u, v) + \lambda_{\max}(\bar{\Sigma}_\epsilon)},$$

where  $t \in [0, \lambda_{\max}(\bar{\Sigma}_\epsilon)]$ .  $f_1(\cdot)$  has axis of symmetry

$$\begin{aligned} t &= \frac{n \sup_{u,v \in \Omega} \Psi_\theta(u, v) + \lambda_{\max}(\bar{\Sigma}_\epsilon) + 2(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i))}{2} \\ &\geq \frac{(n-2) \sup_{u,v \in \Omega} \Psi_\theta(u, v) + \lambda_{\max}(\bar{\Sigma}_\epsilon)}{2}, \end{aligned}$$

where the last inequality is true because  $\Psi_\theta(x_i, x) \geq 0$  and  $\Psi_\theta(x_i, x_i) < \sup_{u,v \in \Omega} \Psi_\theta(u, v)$ .

If  $(n-2) \sup_{u,v \in \Omega} \Psi_\theta(u, v) > \lambda_{\max}(\bar{\Sigma}_\epsilon)$ , then the axis of symmetry lies to the right of the interval  $[0, \lambda_{\max}(\bar{\Sigma}_\epsilon)]$  and  $f_1(t)$  is increasing in  $[0, \lambda_{\max}(\bar{\Sigma}_\epsilon)]$ . This indicates

$$\begin{aligned} f_1(t) &\leq \lambda_{\max}(\bar{\Sigma}_\epsilon) + \frac{2\lambda_{\max}(\bar{\Sigma}_\epsilon)(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i)) - \lambda_{\max}(\bar{\Sigma}_\epsilon)^2}{n \sup_{u,v \in \Omega} \Psi_\theta(u, v) + \lambda_{\max}(\bar{\Sigma}_\epsilon)} \\ &= \frac{\lambda_{\max}(\bar{\Sigma}_\epsilon)(n \sup_{u,v \in \Omega} \Psi_\theta(u, v) + 2(\Psi_\theta(x_i, x) - \Psi_\theta(x_i, x_i)))}{n \sup_{u,v \in \Omega} \Psi_\theta(u, v) + \lambda_{\max}(\bar{\Sigma}_\epsilon)}. \end{aligned} \tag{A.5}$$

Plugging (A.5) into (A.4), gives the result.

## A.2 Assumptions for Theorem 1.4.1

**Assumption A.2.1.** Assume  $\kappa(\Psi_\theta(\bar{X}, \bar{X}) + \Sigma_\epsilon) = r/\delta$  with  $r < 1$ , and

$$\begin{aligned} \|h(x) - \tilde{h}(x)\|_2 &\leq \delta \|h(x)\|_2, \|\hat{f}(\bar{X}) - \tilde{f}(\bar{X})\|_2 \leq \delta \|\hat{f}(\bar{X})\|_2, \\ \|\Psi_\theta(\bar{X}, \bar{X}) + \Sigma_\epsilon - (\tilde{\Psi}_\theta(\bar{X}, \bar{X}) + \tilde{\Sigma}_\epsilon)\|_2 &\leq \delta \|\Psi_\theta(\bar{X}, \bar{X}) + \Sigma_\epsilon\|_2, \quad \text{and} \\ \|\Psi_\theta(x, \bar{X}) - \tilde{\Psi}_\theta(x, \bar{X})\|_2 &\leq \delta \|\Psi_\theta(x, \bar{X})\|_2. \end{aligned}$$

### A.3 Proof of Theorem 1.5.1

Here, the derivatives of the log-likelihood and emulator are expressed in terms of the equivalent parameters  $\vartheta = (\beta', \sigma^2, \rho, \gamma)'$ , where  $\gamma_i = \text{Var}(\epsilon(x_i))/\sigma^2$ . The vector of derivatives of the emulator with respect to the parameters  $\frac{\partial \hat{f}(x)}{\partial \vartheta}$  has block components

$$\begin{aligned}
c_1 &= \frac{\partial \hat{f}(x)}{\partial \beta} = h(x) - H(X)^T (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} \Phi_\rho(X, x), \\
c_2 &= \frac{\partial \hat{f}(x)}{\partial \sigma^2} = 0, \\
(c_3)_j &= \frac{\partial \hat{f}(x)}{\partial \rho_j} = \frac{\partial \Phi_\rho(x, X)}{\partial \rho_j} (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} (f(X) - H(X)\beta) \\
&\quad - \Phi_\rho(x, X) (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} \frac{\partial \Phi_\rho(X, X)}{\partial \rho_j} (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} (f(X) - H(X)\beta), \\
(c_4)_t &= \frac{\partial \hat{f}(x)}{\partial \tau_t} \Phi_\rho(x, X) (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} \text{diag} \left\{ \frac{\partial \gamma_1}{\partial \tau_t} I_{k_1}, \dots, \frac{\partial \gamma_i}{\partial \tau_t} I_{k_i}, \dots, \frac{\partial \gamma_n}{\partial \tau_t} I_{k_n} \right\} \\
&\quad \times (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} (y(X) - H(X)\beta),
\end{aligned}$$

where  $\Sigma_\gamma = \text{diag}(\gamma_1 I_{k_1}, \dots, \gamma_n I_{k_n})$ . The vector of derivatives of the log-likelihood with respect to the parameters  $\frac{\partial l}{\partial \vartheta}$  has block components

$$\begin{aligned}
\frac{\partial l}{\partial \beta} &= \frac{1}{\sigma^2} (X)^T [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} (f(X) - H(X)\beta), \\
\frac{\partial l}{\partial \sigma^2} &= -\frac{m}{2\sigma^2} + \frac{1}{2\sigma^4} (f(X) - H(X)\beta)^T (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} (f(X) - H(X)\beta), \\
\frac{\partial l}{\partial \rho_j} &= -\frac{1}{2} \text{trace} \left( [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} \frac{\partial \Phi_\rho(X, X)}{\partial \rho_j} \right) \\
&\quad + \frac{1}{2\sigma^2} (f(X) - H(X)\beta)^T [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} \frac{\partial \Phi_\rho(X, X)}{\partial \rho_j} [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} (f(X) - H(X)\beta), \\
\frac{\partial l}{\partial \tau_t} &= -\frac{1}{2\sigma^2} \text{trace} \left( [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} \text{diag} \left\{ \frac{\partial \gamma_1}{\partial \tau_t} I_{k_1}, \dots, \frac{\partial \gamma_i}{\partial \tau_t} I_{k_i}, \dots, \frac{\partial \gamma_n}{\partial \tau_t} I_{k_n} \right\} \right) \\
&\quad + \frac{1}{2\sigma^4} (f(X) - H(X)\beta)^T [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} \text{diag} \left( \frac{\partial \gamma_1}{\partial \tau_t} I_{k_1}, \dots, \frac{\partial \gamma_i}{\partial \tau_t} I_{k_i}, \dots, \frac{\partial \gamma_n}{\partial \tau_t} I_{k_n} \right) \\
&\quad \times [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} (f(X) - H(X)\beta).
\end{aligned}$$

So, the information matrix has block components

$$\begin{aligned}
\mathbb{E} - \frac{\partial^2 l}{\partial \beta^2} &= \frac{1}{\sigma^2} (X)^T [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} H(X), \\
\mathbb{E} - \frac{\partial^2 l}{\partial \beta \partial \sigma^2} &= 0, \\
\mathbb{E} - \frac{\partial^2 l}{\partial \beta \partial \tau_t} &= 0, \\
\mathbb{E} - \frac{\partial^2 l}{\partial \beta \partial \rho_j} &= 0, \\
\mathbb{E} - \frac{\partial^2 l}{\partial \sigma^2 \partial \sigma^2} &= \frac{m}{2\sigma^4} \\
\mathbb{E} - \frac{\partial^2 l}{\partial \sigma^2 \partial \rho_j} &= \frac{1}{2} \text{trace} \left( [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} \frac{\partial \Phi_\rho(X, X)}{\partial \rho_j} \right), \\
\mathbb{E} - \frac{\partial^2 l}{\partial \rho_{j_1} \partial \rho_{j_2}} &= \frac{1}{2} \text{trace} \left( [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} \frac{\partial \Phi_\rho(X, X)}{\partial \rho_{j_2}} [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} \frac{\partial \Phi_\rho(X, X)}{\partial \rho_{j_1}} \right), \\
\mathbb{E} - \frac{\partial^2 l}{\partial \tau_t \partial \sigma^2} &= \frac{1}{2\sigma^2} \text{trace} \left( [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} \text{diag} \left\{ \frac{\partial \gamma_1}{\partial \tau_t} I_{k_1}, \dots, \frac{\partial \gamma_i}{\partial \tau_t} I_{k_i}, \dots, \frac{\partial \gamma_n}{\partial \tau_t} I_{k_n} \right\} \right), \\
\mathbb{E} - \frac{\partial^2 l}{\partial \tau_t \partial \rho_j} &= \frac{1}{2} \text{trace} \left( [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} \frac{\partial \Phi_\rho(X, X)}{\partial \rho_j} [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} \right. \\
&\quad \left. \times \text{diag} \left\{ \frac{\partial \gamma_1}{\partial \tau_t} I_{k_1}, \dots, \frac{\partial \gamma_i}{\partial \tau_t} I_{k_i}, \dots, \frac{\partial \gamma_n}{\partial \tau_t} I_{k_n} \right\} \right), \\
\mathbb{E} - \frac{\partial^2 l}{\partial \tau_{t_1} \partial \tau_{t_2}} &= \frac{1}{2} \text{trace} \left( [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} \text{diag} \left\{ \frac{\partial \gamma_1}{\partial \tau_{t_1}} I_{k_1}, \dots, \frac{\partial \gamma_i}{\partial \tau_{t_1}} I_{k_i}, \dots, \frac{\partial \gamma_n}{\partial \tau_{t_1}} I_{k_n} \right\} \right. \\
&\quad \left. \times [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} \text{diag} \left\{ \frac{\partial \gamma_1}{\partial \tau_{t_2}} I_{k_1}, \dots, \frac{\partial \gamma_i}{\partial \tau_{t_2}} I_{k_i}, \dots, \frac{\partial \gamma_n}{\partial \tau_{t_2}} I_{k_n} \right\} \right).
\end{aligned}$$

Building from (1.14) and the block representations above gives

$$\begin{aligned}
\mathbb{E}\{\hat{f}_{\vartheta_*}(x) - \hat{f}_{\hat{\vartheta}}(x)\}^2 &\approx (c_1^T, c_2^T, c_3^T, c_4^T) \begin{pmatrix} a_{11}^{-1} & 0 \\ 0 & \mathcal{I}^{-1} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} \\
&= c_1^T a_{11}^{-1} c_1 + (c_2^T, c_3^T, c_4^T) \mathcal{I}^{-1} \begin{pmatrix} c_2 \\ c_3 \\ c_4 \end{pmatrix} \\
&= \text{Part(I)} + \text{Part(II)},
\end{aligned}$$

where

$$a_{11} = \frac{\partial^2 l}{\partial \beta^2} \quad \text{and} \quad \mathcal{I} = \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix}$$

with

$$\begin{aligned}
\mathcal{I}_{11} &= -\mathbb{E} \frac{\partial^2 l}{\partial \sigma^2 \partial \sigma^2}, \quad \mathcal{I}_{12} = -(\mathbb{E} \frac{\partial^2 l}{\partial \sigma^2 \partial \rho'}, \mathbb{E} \frac{\partial^2 l}{\partial \sigma^2 \partial \tau'}), \quad \mathcal{I}_{21} = \mathcal{I}_{12}^T, \quad \mathcal{I}_{22} = \begin{pmatrix} D_1 & D_2^T \\ D_2 & D_3 \end{pmatrix}, \\
D_1 &= -\mathbb{E} \frac{\partial^2 l}{\partial \rho \partial \rho'}, \quad D_2 = -\mathbb{E} \frac{\partial^2 l}{\partial \tau \partial \rho'}, \quad \text{and} \quad D_3 = -\mathbb{E} \frac{\partial^2 l}{\partial \tau \partial \tau'}. \tag{A.6}
\end{aligned}$$

Applying block matrix inverse results [17] and noticing that  $c_2 = 0$  gives

$$\text{Part(II)} = c^T B_1^{-1} c,$$

where  $B_1 = \mathcal{I}_{22} - \mathcal{I}_{21} \mathcal{I}_{11}^{-1} \mathcal{I}_{12}$ , and  $c = (c_3^T, c_4^T)^T$ . With the aim of bounding Part(II), the

following notation is introduced. Let

$$\begin{aligned} a_j &= \text{vec} \left( \sigma^2 \frac{\partial \Phi_\rho(X, X)}{\partial \rho_j} \right), \\ b_t &= \text{vec} \left( \text{diag} \left\{ \frac{\partial \gamma_1}{\partial \tau_t} I_{k_1}, \dots, \frac{\partial \gamma_i}{\partial \tau_t} I_{k_i}, \dots, \frac{\partial \gamma_n}{\partial \tau_t} I_{k_n} \right\} \right), \\ A_1 &= (a_1, \dots, a_{|\rho|}, b_1, \dots, b_{|\tau|}). \end{aligned}$$

Then,

$$\begin{aligned} B_1 &= \frac{1}{\sigma^4} A_1^T ((\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} \otimes (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} \\ &\quad - \frac{1}{n} \text{vec}([\Phi_\rho(X, X) + \Sigma_\gamma]^{-1}) \text{vec}([\Phi_\rho(X, X) + \Sigma_\gamma]^{-1})^T) A_1. \end{aligned}$$

For simplicity, let

$$W_1 = \Phi_\rho(X, X) + \Sigma_\gamma \quad \text{and} \quad w = \frac{\text{vec}(W_1)}{\|\text{vec}(W_1)\|_2}.$$

The matrix inside the quadratic form has eigenvector  $w$  with corresponding eigenvalue 0.

Following the approach in [10], the minimum eigenvalue of  $B_1$  can be bounded below by

$$\frac{1}{\sigma^4} \lambda_{\min}(A_1^T (I - ww^T) A_1) \times \lambda_2 \left( (W_1^{-1} \otimes W_1^{-1} - \frac{1}{m} \text{vec}(W_1^{-1}) \text{vec}(W_1^{-1})^T) \right),$$

where  $\lambda_2$  denotes the second smallest eigenvalue of its argument. Weyl's theorem [103] implies that the second smallest eigenvalue can be bounded below by

$$\frac{1}{\sigma^4} \lambda_{\min}(W_1^{-1} \otimes W_1^{-1}) = \frac{1}{\sigma^4 \lambda_{\max}(\Psi_\theta(X, X) + \Sigma_\epsilon)^2} \geq \frac{1}{\sigma^4 (m \sup_{u,v \in \Omega} \Phi_\rho(u, v) + \lambda_{\max}(\Sigma_\gamma))^2}.$$

For  $\lambda_{\min}(A_1^T (I - ww^T) A_1)$ , an approximate lower bound is given. Let  $\xi = (\rho, \tau)$ . Notice

that

$$\begin{aligned}
& A_1^T (I - ww^T) A_1 \\
&= \left[ \sum_{i,j} \frac{\partial W_1(x_i, x_j)}{\partial \xi} \frac{\partial W_1(x_i, x_j)}{\partial \xi'} \right. \\
&\quad \left. - \frac{1}{\|\text{vec}(W_1)\|_2^2} \left( \sum_{i,j} \frac{\partial W_1(x_i, x_j)}{\partial \xi} W_1(x_i, x_j) \right) \left( \sum_{i,j} \frac{\partial W_1(x_i, x_j)}{\partial \xi'} W_1(x_i, x_j) \right) \right] \\
&\approx m^2 \left[ \int \frac{\partial W_1(x, y)}{\partial \xi} \frac{\partial W_1(x, y)}{\partial \xi'} dF^2(x, y) \right. \\
&\quad \left. - \frac{1}{\|W_1\|_{L_2(F^2)}^2} \left( \int \frac{\partial W_1(x, y)}{\partial \xi} W_1(x, y) dF^2(x, y) \right) \left( \int \frac{\partial W_1(x, y)}{\partial \xi'} W_1(x, y) dF^2(x, y) \right) \right] \\
&\succeq m^2 s_1, \tag{A.7}
\end{aligned}$$

where  $W_1(x, y) = \Phi_\rho(x, y) + \frac{\sigma_x^2(x)}{\sigma^2} \mathbb{I}_{\{x=y\}}$  and  $F^2$  denotes the large sample distribution of point pairs. Applying a version of the Cauchy-Schwarz inequality for random vectors [104], gives  $s_1 \geq 0$  with  $s_1 > 0$  unless

$$\frac{\partial W_1(x, y)}{\partial \xi} a = W_1(x, y) b$$

with probability 1 with respect to large sample distribution of point pairs  $F^2$  for some vectors  $a$  and  $b$ . So, Part(II) has approximate upper bound

$$\frac{\sigma^4 (m \sup_{u,v \in \Omega} \Phi_\rho(u, v) + \lambda_{\max}(\Sigma_\gamma))^2}{m^2 s_1} \|c\|_2^2. \tag{A.8}$$

Also,

$$\begin{aligned}
\text{Part(I)} &\leq \frac{\sigma^2 \|c_1\|_2^2}{\lambda_{\min}(H(X)^T [\Phi_\rho(X, X) + \Sigma_\gamma]^{-1} H(X))} \\
&\leq \frac{\sigma^2 \|c_1\|_2^2 \lambda_{\max}(\Phi_\rho(X, X) + \Sigma_\gamma)}{\lambda_{\min}(H(X)^T H(X))}. \tag{A.9}
\end{aligned}$$

Following development similar to above,  $\lambda_{\min}(H(X)^T H(X))$  admits approximation

$$\lambda_{\min}(H(X)^T H(X)) = \lambda_{\min}\left(\sum_{i=1}^n h(x_i)h(x_i)'\right) \approx m\lambda_{\min}\left(\int h(y)h(y)'dF(y)\right) = ms_2, \quad (\text{A.10})$$

with respect to the large sample distribution of the input locations,  $F$ . Further,  $s_2 \geq 0$  with equality if and only if there exists  $a \neq 0$  such that  $h(y)'a = 0$  with probability 1.

Combining (A.8) and (A.9) gives approximate upper bound for Part(I) + Part(II)

$$\begin{aligned} & \frac{\sigma^2 \|c_1\|_2^2 \lambda_{\max}(\Phi_\rho(X, X) + \Sigma_\gamma)}{ms_2} + \frac{\sigma^4 \|c\|_2^2 (m \sup_{u,v \in \Omega} \Phi_\rho(u, v) + \lambda_{\max}(\Sigma_\gamma))^2}{m^2 s_1}, \\ \leq & \frac{\sigma^2 \|c_1\|_2^2 (m \sup_{u,v \in \Omega} \Phi_\rho(u, v) + \lambda_{\max}(\Sigma_\gamma))}{ms_2} + \frac{\sigma^4 \|c\|_2^2 (m \sup_{u,v \in \Omega} \Phi_\rho(u, v) + \lambda_{\max}(\Sigma_\gamma))^2}{m^2 s_1}, \end{aligned} \quad (\text{A.11})$$

finishing the proof of Theorem 1.5.1.

#### A.4 Proof of Proposition 1.5.1

Recall that

$$\begin{aligned} (c_4)_t &= \frac{\partial \hat{f}(x)}{\partial \tau_t} = \Phi_\rho(x, X) (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} \text{diag}\left(\frac{\partial \gamma_1}{\partial \tau_t} I_{k_1}, \dots, \frac{\partial \gamma_i}{\partial \tau_t} I_{k_i}, \dots, \frac{\partial \gamma_n}{\partial \tau_t} I_{k_n}\right) \\ & \quad \times (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} (y(X) - H(X)\beta). \end{aligned}$$

In this section we would give an upper bound of  $(c_4)_t$ . Without loss of generality, we can suppose  $\Phi_\rho(x, x) = 1$ . Let

$$\Phi_\rho(X, X) + \Sigma_\gamma = \begin{bmatrix} B_1 + \Sigma_{\gamma_1} & R^T \\ R & B_2 + \Sigma_{\gamma_2} \end{bmatrix},$$

where

$$\begin{aligned}
B_1 &= 11^T, \\
\Sigma_{\gamma_1} &= \sigma_1^2 I_{k_1}, \\
R &= \Phi_\rho(X_2, x_1)1^T, \\
B_2 &= \Phi_\rho(X_2, X_2).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
(\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} &= \begin{bmatrix} B_1 + \Sigma_{\gamma_1} & R^T \\ R & B_2 + \Sigma_{\gamma_2} \end{bmatrix}^{-1} \\
&= \begin{bmatrix} B_{22}^{-1} & -B_{22}^{-1}R^T(B_2 + \Sigma_{\gamma_2})^{-1} \\ -(B_2 + \Sigma_{\gamma_2})^{-1}RB_{22}^{-1} & (B_2 + \Sigma_{\gamma_2})^{-1} + (B_2 + \Sigma_{\gamma_2})^{-1}RB_{22}^{-1}R^T(B_2 + \Sigma_{\gamma_2})^{-1} \end{bmatrix}
\end{aligned}$$

where  $B_{22} = B_1 + \Sigma_{\gamma_1} - R^T(B_2 + \Sigma_{\gamma_2})^{-1}R$ . Notice that

$$(B_1 + \Sigma_{\gamma_1})^{-1}1 = \frac{1}{k_1 + \sigma_1^2}1,$$

we have

$$\begin{aligned}
B_{22}^{-1} &= (B_1 + \Sigma_{\gamma_1} - R^T(B_2 + \Sigma_{\gamma_2})^{-1}R)^{-1} \\
&= (B_1 + \Sigma_{\gamma_1})^{-1}((B_1 + \Sigma_{\gamma_1})^{-1} - (B_1 + \Sigma_{\gamma_1})^{-1}R^T(B_2 + \Sigma_{\gamma_2})^{-1}R(B_1 + \Sigma_{\gamma_1})^{-1})^{-1}(B_1 + \Sigma_{\gamma_1})^{-1} \\
&= (B_1 + \Sigma_{\gamma_1})^{-1} \left( (B_1 + \Sigma_{\gamma_1})^{-1} - \frac{1}{(k_1 + \sigma_1^2)^2} 1\Phi_\rho(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}\Phi_\rho(X_2, x_1)1^T \right)^{-1} (B_1 + \Sigma_{\gamma_1})^{-1} \\
&= (B_1 + \Sigma_{\gamma_1})^{-1} \left( (B_1 + \Sigma_{\gamma_1})^{-1} - \frac{\Phi_\rho(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}\Phi_\rho(X_2, x_1)11^T}{(k_1 + \sigma_1^2)^2} \right)^{-1} (B_1 + \Sigma_{\gamma_1})^{-1}.
\end{aligned}$$

By binomial inverse theorem,

$$\begin{aligned}
& \left( (B_1 + \Sigma_{\gamma_1})^{-1} - \frac{\Phi_{\rho}(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}\Phi_{\rho}(X_2, x_1)}{(k_1 + \sigma_1^2)^2} 11^T \right)^{-1} \\
&= B_1 + \Sigma_{\gamma_1} + \frac{\Phi_{\rho}(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}\Phi_{\rho}(X_2, x_1)}{(k_1 + \sigma_1^2)^2} \frac{(B_1 + \Sigma_{\gamma_1})11^T(B_1 + \Sigma_{\gamma_1})}{1 - \frac{\Phi_{\rho}(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}\Phi_{\rho}(X_2, x_1)}{(k_1 + \sigma_1^2)^2} 1^T(B_1 + \Sigma_{\gamma_1})1} \\
&= B_1 + \Sigma_{\gamma_1} + \frac{\Phi_{\rho}(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}\Phi_{\rho}(X_2, x_1)}{(k_1 + \sigma_1^2)^2} \frac{(k_1 + \sigma_1^2)^2 11^T}{1 - \frac{\Phi_{\rho}(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}\Phi_{\rho}(X_2, x_1)}{(k_1 + \sigma_1^2)^2} (k_1 + \sigma_1^2)k_1} \\
&= B_1 + \Sigma_{\gamma_1} + \Phi_{\rho}(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}\Phi_{\rho}(X_2, x_1) \frac{11^T}{1 - \frac{\Phi_{\rho}(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}\Phi_{\rho}(X_2, x_1)}{k_1 + \sigma_1^2} k_1}.
\end{aligned}$$

Let  $d = \Phi_{\rho}(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}\Phi_{\rho}(X_2, x_1)$ . Thus,

$$\begin{aligned}
B_{22}^{-1} &= (B_1 + \Sigma_{\gamma_1})^{-1} \left( B_1 + \Sigma_{\gamma_1} + d \frac{11^T}{1 - \frac{d}{k_1 + \sigma_1^2} k_1} \right) (B_1 + \Sigma_{\gamma_1})^{-1} \\
&= (B_1 + \Sigma_{\gamma_1})^{-1} + d \frac{(B_1 + \Sigma_{\gamma_1})^{-1} 11^T (B_1 + \Sigma_{\gamma_1})^{-1}}{1 - \frac{d}{k_1 + \sigma_1^2} k_1} \\
&= (B_1 + \Sigma_{\gamma_1})^{-1} + d \frac{\frac{1}{(k_1 + \sigma_1^2)^2} 11^T}{1 - \frac{d}{k_1 + \sigma_1^2} k_1}.
\end{aligned}$$

Since

$$\Phi_{\rho}(x, X) = (\Phi_{\rho}(x, x_1)1^T, \Phi_{\rho}(x, X_2)),$$

we have

$$\begin{aligned}
& \Phi_\rho(x, X)(\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} \\
&= (\Phi_\rho(x, x_1)1^T, \Phi_\rho(x, X_2)) \\
& \quad \times \begin{bmatrix} B_{22}^{-1} & -B_{22}^{-1}R^T(B_2 + \Sigma_{\gamma_2})^{-1} \\ -(B_2 + \Sigma_{\gamma_2})^{-1}RB_{22}^{-1} & (B_2 + \Sigma_{\gamma_2})^{-1} + (B_2 + \Sigma_{\gamma_2})^{-1}RB_{22}^{-1}R^T(B_2 + \Sigma_{\gamma_2})^{-1} \end{bmatrix}, \\
&= \left( \Phi_\rho(x, x_1)1^T B_{22}^{-1} - \Phi_\rho(x, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}RB_{22}^{-1}, \right. \\
& \quad - \Phi_\rho(x, x_1)1^T B_{22}^{-1}R^T(B_2 + \Sigma_{\gamma_2})^{-1} + \Phi_\rho(x, X_2)(B_2 + \Sigma_{\gamma_2})^{-1} \\
& \quad \left. + (B_2 + \Sigma_{\gamma_2})^{-1}RB_{22}^{-1}R^T(B_2 + \Sigma_{\gamma_2})^{-1} \right).
\end{aligned}$$

Notice that

$$\begin{aligned}
B_{22}^{-1}1 &= (B_1 + \Sigma_{\gamma_1})^{-1}1 + d \frac{\frac{1}{(k_1 + \sigma_1^2)^2} 11^T 1}{1 - \frac{d}{k_1 + \sigma_1^2} k_1} \\
&= (B_1 + \Sigma_{\gamma_1})^{-1}1 + d \frac{\frac{1}{(k_1 + \sigma_1^2)^2} 11^T 1}{1 - \frac{d}{k_1 + \sigma_1^2} k_1} \\
&= \left( \frac{1}{k_1 + \sigma_1^2} + \frac{\frac{dk_1}{(k_1 + \sigma_1^2)^2}}{1 - \frac{dk_1}{k_1 + \sigma_1^2}} \right) 1,
\end{aligned}$$

we have (let  $d_1 = \left( \frac{1}{k_1 + \sigma_1^2} + \frac{\frac{dk_1}{(k_1 + \sigma_1^2)^2}}{1 - \frac{dk_1}{k_1 + \sigma_1^2}} \right) = \frac{1}{k_1 + \sigma_1^2 - dk_1}$ )

$$\begin{aligned}
& \Phi_\rho(x, x_1)1^T B_{22}^{-1} - \Phi_\rho(x, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}RB_{22}^{-1} \\
&= \Phi_\rho(x, x_1)d_1 1^T - d_1 \Phi_\rho(x, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}\Phi_\rho(X_2, x_1)1^T,
\end{aligned}$$

and

$$\begin{aligned}
& -\Phi_\rho(x, x_1)1^T B_{22}^{-1}R^T(B_2 + \Sigma_{\gamma_2})^{-1} + \Phi_\rho(x, X_2)(B_2 + \Sigma_{\gamma_2})^{-1} + (B_2 + \Sigma_{\gamma_2})^{-1}RB_{22}^{-1}R^T(B_2 + \Sigma_{\gamma_2})^{-1} \\
&= -\Phi_\rho(x, x_1)k_1d_1\Psi_\theta(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1} + \Phi_\rho(x, X_2)(B_2 + \Sigma_{\gamma_2})^{-1} \\
&+ \Phi_\rho(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}\Phi_\rho(X_2, x_1)k_1d_1\Phi_\rho(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1} \\
&= -\Phi_\rho(x, x_1)k_1d_1\Phi_\rho(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1} + \Phi_\rho(x, X_2)(B_2 + \Sigma_{\gamma_2})^{-1} \\
&+ dk_1d_1\Phi_\rho(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}.
\end{aligned}$$

With the same procedure, we have

$$\begin{aligned}
& (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1}(y(X) - H(X)\beta) \\
&= \left( (y(x_1) - H(x_1)\beta)d_11 - d_1\Phi_\rho(x_1, X_2)(B_2 + \Sigma_{\gamma_2})^{-1}(y(X_2) - H(X_2)\beta)1, \right. \\
&\quad \left. - (y(x_1) - H(x_1)\beta)k_1d_1(B_2 + \Sigma_{\gamma_2})^{-1}\Phi_\rho(X_2, x_1) + (B_2 + \Sigma_{\gamma_2})^{-1}(y(X_2) - H(X_2)\beta) \right. \\
&\quad \left. + dk_1d_1(B_2 + \Sigma_{\gamma_2})^{-1}\Phi_\rho(X_2, x_1) \right).
\end{aligned}$$

Thus,

$$\begin{aligned}
(c_4)_t &= \frac{\partial \hat{f}(x)}{\partial \tau_t} = \Phi_\rho(x, X)(\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} \text{diag}\left(\frac{\partial \gamma_1}{\partial \tau_t} I_{k_1}, \dots, \frac{\partial \gamma_i}{\partial \tau_t} I_{k_i}, \dots, \frac{\partial \gamma_n}{\partial \tau_t} I_{k_n}\right) \\
&\quad \times (\Phi_\rho(X, X) + \Sigma_\gamma)^{-1} (y(X) - H(X)\beta) \\
&= k_1 \frac{\partial \gamma_1}{\partial \tau_t} (\Phi_\rho(x, x_1) d_1 - d_1 \Phi_\rho(x, X_2) (B_2 + \Sigma_{\gamma_2})^{-1} \Phi_\rho(X_2, x_1)) ((y(x_1) - H(x_1)\beta) d_1 \\
&\quad - d_1 \Phi_\rho(x_1, X_2) (B_2 + \Sigma_{\gamma_2})^{-1} (y(X_2) - H(X_2)\beta)) \\
&\quad + (-\Phi_\rho(x, x_1) k_1 d_1 \Phi_\rho(x_1, X_2) (B_2 + \Sigma_{\gamma_2})^{-1} + \Phi_\rho(x, X_2) (B_2 + \Sigma_{\gamma_2})^{-1} \\
&\quad + dk_1 d_1 \Phi_\rho(x_1, X_2) (B_2 + \Sigma_{\gamma_2})^{-1}) \text{diag}\left(\frac{\partial \gamma_2}{\partial \tau_t} I_{k_2}, \dots, \frac{\partial \gamma_i}{\partial \tau_t} I_{k_i}, \dots, \frac{\partial \gamma_n}{\partial \tau_t} I_{k_n}\right) \\
&\quad \times (-(y(x_1) - H(x_1)\beta) k_1 d_1 (B_2 + \Sigma_{\gamma_2})^{-1} \Phi_\rho(X_2, x_1) + (B_2 + \Sigma_{\gamma_2})^{-1} (y(X_2) - H(X_2)\beta) \\
&\quad + dk_1 d_1 (B_2 + \Sigma_{\gamma_2})^{-1} \Phi_\rho(X_2, x_1))
\end{aligned}$$

Let  $d_2 = \frac{1}{1 + \sigma_1^2/k_1 - d}$ , we have

$$\begin{aligned}
(c_4)_t &= \frac{1}{k_1} \frac{\partial \gamma_1}{\partial \tau_t} (\Phi_\rho(x, x_1) d_2 - d_2 \Phi_\rho(x, X_2) (B_2 + \Sigma_{\gamma_2})^{-1} \Phi_\rho(X_2, x_1)) ((y(x_1) - H(x_1)\beta) d_2 \\
&\quad - d_2 \Phi_\rho(x_1, X_2) (B_2 + \Sigma_{\gamma_2})^{-1} (y(X_2) - H(X_2)\beta)) \\
&\quad + (-\Phi_\rho(x, x_1) d_2 \Phi_\rho(x_1, X_2) (B_2 + \Sigma_{\gamma_2})^{-1} + \Phi_\rho(x, X_2) (B_2 + \Sigma_{\gamma_2})^{-1} \\
&\quad + dd_2 \Phi_\rho(x_1, X_2) (B_2 + \Sigma_{\gamma_2})^{-1}) \text{diag}\left(\frac{\partial \gamma_2}{\partial \tau_t} I_{k_2}, \dots, \frac{\partial \gamma_i}{\partial \tau_t} I_{k_i}, \dots, \frac{\partial \gamma_n}{\partial \tau_t} I_{k_n}\right) \\
&\quad \times (-(y(x_1) - H(x_1)\beta) d_2 (B_2 + \Sigma_{\gamma_2})^{-1} \Phi_\rho(X_2, x_1) + (B_2 + \Sigma_{\gamma_2})^{-1} (y(X_2) - H(X_2)\beta) \\
&\quad + dd_2 (B_2 + \Sigma_{\gamma_2})^{-1} \Phi_\rho(X_2, x_1)) \\
&= \Phi_\rho(x, X') (\Phi_\rho(X', X') + \Sigma_\gamma)^{-1} \text{diag}\left(\frac{1}{k_1} \frac{\partial \gamma_1}{\partial \tau_t}, \dots, \frac{\partial \gamma_i}{\partial \tau_t} I_{k_i}, \dots, \frac{\partial \gamma_n}{\partial \tau_t} I_{k_n}\right) \\
&\quad \times (\Phi_\rho(X', X') + \Sigma_\gamma)^{-1} (y(X') - H(X')\beta),
\end{aligned}$$

where  $X' = (x_1, X_2)$ . Thus, by continuing this procedure, we have

$$|(c_4)_t| \leq \frac{\|\Phi_\rho(x, \bar{X})\|_2 \|\bar{Y} - H(\bar{X})\beta\|_2}{(\lambda_{\min}(\Phi_\rho(\bar{X}, \bar{X}) + \Sigma_\gamma))^2} \max_{i: x_i \in \bar{X}} \left| \frac{1}{k_i} \frac{\partial \gamma_i}{\partial \tau_t} \right|.$$

### A.5 Proof of Theorem 1.6.1

The following lemma, which describes the accuracy of solving linear systems [105], will be used to develop a bound on the numeric error.

**Lemma A.5.1.** *Suppose  $Ax = b$  and  $\tilde{A}\tilde{x} = \tilde{b}$  with  $\|\tilde{A} - A\|_2 \leq \delta\|A\|_2$ ,  $\|\tilde{b} - b\|_2 \leq \delta\|b\|_2$ , and  $\kappa(A) = r/\delta < 1/\delta$  for some  $\delta > 0$ . Then,  $\tilde{A}$  is non-singular,*

$$\begin{aligned} \frac{\|\tilde{x}\|_2}{\|x\|_2} &\leq \frac{1+r}{1-r}, \\ \frac{\|\tilde{x} - x\|_2}{\|x\|_2} &\leq \frac{2\delta}{1-r} \kappa(A), \end{aligned} \tag{A.12}$$

where  $\kappa(A) = \|A\|_2\|A^{-1}\|_2$ .

Further, for conformable  $A$ ,  $b$ ,  $\tilde{A}$ , and  $\tilde{b}$ , we have

$$\begin{aligned} \|Ab - \tilde{A}\tilde{b}\|_2 &= \|A(b - \tilde{b}) - (\tilde{A} - A)\tilde{b}\|_2 \\ &\leq \|A(b - \tilde{b})\|_2 + \|(\tilde{A} - A)\tilde{b}\|_2 \leq \|A\|_2\|(b - \tilde{b})\|_2 + \|(\tilde{A} - A)\|_2\|\tilde{b}\|_2. \end{aligned} \tag{A.13}$$

In order to satisfy the conditions of Lemma A.5.1, we make a few assumptions *in addition to Assumption A.2.1*, in particular, with regard to the accuracy of numeric optimization.

**Assumption A.5.1.** *Assume  $\kappa(\hat{A}) = r/\delta$  with  $r < 1$  and*

$$\|\hat{A} - \tilde{A}\|_2 \leq \delta\|\hat{A}\|_2, \|\Psi_{\hat{\theta}}(\bar{X}, x) - \Psi_{\tilde{\theta}}(\bar{X}, x)\|_2 \leq \delta\|\Psi_{\hat{\theta}}(\bar{X}, x)\|_2.$$

Note that this assumption does not concern the parameter estimates themselves, but instead the accuracy of the solution to the optimization problem. If the optimization problem

is solved with sufficient accuracy, then this assumption will be satisfied. However, as we will see in the following, the regression function coefficients  $\beta$  have great potential to cause problems. Briefly, in order to control parameter estimation numeric error, we need that numeric properties are even more tightly controlled, in particular, an even smaller condition number of  $\Psi_\theta(\bar{X}, \bar{X}) + \Sigma_\epsilon$ , which is stated in the following assumption.

**Assumption A.5.2.**

$$\delta\kappa(\hat{A})\kappa(H(\bar{X})^T H(\bar{X})) \left( 1 + (1 + \delta)^2 + \frac{(1 + \delta)^2}{1 - r} \kappa(\hat{A}) \right) < 1.$$

Assumption A.5.2 is a strong assumption, since it requires  $\delta\kappa(\hat{A})^2$  to be relatively small, at least smaller than 1. However, since our goal is to make  $\kappa(\hat{A})$  small, in practice this condition is not too difficult to be achieved, since we can control the condition number of  $\hat{A}$ .

The following lemma states that if Assumption A.5.2 holds, combining Assumption A.5.1, the conditions of Lemma A.5.1 holds.

**Lemma A.5.2.** *Let*

$$\begin{aligned} r_1 &= \delta\kappa(\hat{A})\kappa(H(\bar{X})^T H(\bar{X})) \left( 1 + (1 + \delta)^2 + \frac{(1 + \delta)^2}{1 - r} \kappa(\hat{A}) \right) \\ &\quad + \frac{1}{2} \min\{\delta, 1 - \delta\kappa(\hat{A})\kappa(H(\bar{X})^T H(\bar{X})) \left( 1 + (1 + \delta)^2 + \frac{(1 + \delta)^2}{1 - r} \kappa(\hat{A}) \right)\} \\ \delta_1 &= \frac{r_1}{\kappa(H(\bar{X})^T \hat{A}^{-1} H(\bar{X}))}. \end{aligned} \tag{A.14}$$

*Suppose Assumptions A.2.1, A.5.1, and A.5.2 hold, we have  $r_1 < 1$  and*

$$\|H(\bar{X})^T \hat{A}^{-1} H(\bar{X}) - \tilde{H}(\bar{X})^T \tilde{A}^{-1} \tilde{H}(\bar{X})\|_2 < \delta_1 \|H(\bar{X})^T \hat{A}^{-1} H(\bar{X})\|_2. \tag{A.15}$$

Thus, we have all tools to give an upper bound of  $|\hat{f}_{\hat{\theta}} - \hat{f}_{\tilde{\theta}}|$ . Using Assumption 2,

$$\begin{aligned}
& |\hat{f}_{\hat{\theta}} - \hat{f}_{\tilde{\theta}}| \\
&= |h(x)^T \hat{\beta} + \Psi_{\hat{\theta}}(x, \bar{X}) \hat{A}^{-1} (f(\bar{X}) - H(\bar{X}) \hat{\beta}) - (h(x)^T \tilde{\beta} + \Psi_{\tilde{\theta}}(x, \bar{X}) \tilde{A}^{-1} (f(\bar{X}) - H(\bar{X}) \tilde{\beta}))| \\
&= |h(x)^T (\hat{\beta} - \tilde{\beta}) + f(\bar{X})^T (\hat{A}^{-1} \Psi_{\hat{\theta}}(\bar{X}, x) - \tilde{A}^{-1} \Psi_{\tilde{\theta}}(\bar{X}, x)) \\
&\quad - [\Psi_{\hat{\theta}}(x, \bar{X}) \hat{A}^{-1} H(\bar{X}) \hat{\beta} - \Psi_{\tilde{\theta}}(x, \bar{X}) \tilde{A}^{-1} H(\bar{X}) \tilde{\beta}]| \\
&\leq \|h(x)\|_2 \|\hat{\beta} - \tilde{\beta}\|_2 + \|f(\bar{X})\|_2 \|\hat{A}^{-1} \Psi_{\hat{\theta}}(\bar{X}, x) - \tilde{A}^{-1} \Psi_{\tilde{\theta}}(\bar{X}, x)\|_2 \\
&\quad + \|\Psi_{\hat{\theta}}(x, \bar{X}) \hat{A}^{-1} H(\bar{X}) \hat{\beta} - \Psi_{\tilde{\theta}}(x, \bar{X}) \tilde{A}^{-1} H(\bar{X}) \tilde{\beta}\|_2 \\
&= \text{Part}(i) + \text{Part}(ii) + \text{Part}(iii). \tag{A.16}
\end{aligned}$$

Part(ii) can be bounded using Lemma A.5.1 as

$$\text{Part}(ii) \leq \|f(\bar{X})\|_2 \frac{2\delta}{1-r} \kappa(\hat{A}) \|\hat{A}^{-1} \Psi_{\hat{\theta}}(\bar{X}, x)\|_2. \tag{A.17}$$

Similarly, Part(iii) can be bounded using (A.13) and Lemma A.5.1 as

$$\begin{aligned}
\text{Part}(iii) &\leq \|H(\bar{X})\|_2 \|\hat{\beta}\|_2 \frac{2\delta}{1-r} \kappa(\hat{A}) \|\hat{A}^{-1} \Psi_{\hat{\theta}}(\bar{X}, x)\|_2 + \|H(\bar{X})\|_2 \|\hat{\beta} - \tilde{\beta}\|_2 \|\tilde{A}^{-1} \Psi_{\tilde{\theta}}(\bar{X}, x)\|_2 \\
&\leq \|H(\bar{X})\|_2 \|\hat{\beta}\|_2 \frac{2\delta}{1-r} \kappa(\hat{A}) \|\hat{A}^{-1} \Psi_{\hat{\theta}}(\bar{X}, x)\|_2 + \|H(\bar{X})\|_2 \|\hat{\beta} - \tilde{\beta}\|_2 \frac{1+r}{1-r} \|\tilde{A}^{-1} \Psi_{\tilde{\theta}}(\bar{X}, x)\|_2.
\end{aligned} \tag{A.18}$$

Combining (A.16), (A.17) and (A.18) gives

$$\begin{aligned}
|\hat{f}_{\hat{\theta}} - \hat{f}_{\tilde{\theta}}| &\leq \frac{2\delta \kappa(\hat{A})}{(1-r) \lambda_{\min}(\hat{A})} \|\Psi_{\hat{\theta}}(\bar{X}, x)\|_2 (\|f(x)\|_2 + \|H(\bar{X})\|_2 \|\hat{\beta}\|_2) \\
&\quad + \|\hat{\beta} - \tilde{\beta}\|_2 (\|h(x)\|_2 + \frac{1+r}{(1-r) \lambda_{\min}(\hat{A})} \|H(\bar{X})\|_2 \|\Psi_{\tilde{\theta}}(\bar{X}, x)\|_2). \tag{A.19}
\end{aligned}$$

Notice that the first term in (A.19) can be controlled by restraining  $g(\Sigma_M, \Sigma_\epsilon)$ , as defined in (1.13). The second part can be controlled by, in addition, restraining  $\|\hat{\beta} - \tilde{\beta}\|_2$ . Recall

that

$$\begin{aligned}\hat{\beta} &= (H(\bar{X})^T \hat{A}^{-1} H(\bar{X}))^{-1} H(\bar{X})^T \hat{A}^{-1} f(\bar{X}), \\ \tilde{\beta} &= (\tilde{H}(\bar{X})^T \tilde{A}^{-1} \tilde{H}(\bar{X}))^{-1} \tilde{H}(\bar{X})^T \tilde{A}^{-1} \tilde{f}(\bar{X}).\end{aligned}$$

Since by Lemma A.5.2, the condition of Lemma A.5.1 holds. Thus, by Lemma A.5.1, we have

$$\|\hat{\beta} - \tilde{\beta}\|_2 \leq \frac{2\delta_1}{1-r_1} \kappa(H(\bar{X})^T \hat{A}^{-1} H(\bar{X})) \|\hat{\beta}\|_2 = \frac{2r_1}{1-r_1} \|\hat{\beta}\|_2.$$

By plugging in (A.14), we have

$$\|\hat{\beta} - \tilde{\beta}\|_2 \leq 2\delta \left( \kappa(\hat{A}) \kappa(H(\bar{X})^T H(\bar{X})) \left( 1 + (1+\delta)^2 + \frac{(1+\delta)^2}{1-r} \kappa(\hat{A}) \right) + 1 \right) \|\hat{\beta}\|_2. \quad (\text{A.20})$$

Combining (A.19) and (A.20), we finish the proof.

## A.6 Proof of Lemma A.5.2

Notice that if Assumption A.5.2 holds, we have  $r_1 < 1$ . We only need to prove (A.15).

Notice that

$$\begin{aligned}& \|H(\bar{X})^T \hat{A}^{-1} H(\bar{X}) - \tilde{H}(\bar{X})^T \tilde{A}^{-1} \tilde{H}(\bar{X})\|_2 \\ & \leq \|H(\bar{X})^T \hat{A}^{-1} H(\bar{X}) - \tilde{H}(\bar{X})^T \hat{A}^{-1} H(\bar{X})\|_2 + \|\tilde{H}(\bar{X})^T \hat{A}^{-1} H(\bar{X}) - \tilde{H}(\bar{X})^T \tilde{A}^{-1} \tilde{H}(\bar{X})\|_2 \\ & \leq \delta \|H(\bar{X})\|_2 \|\hat{A}^{-1} H(\bar{X})\|_2 + \|\tilde{H}(\bar{X})^T \hat{A}^{-1} H(\bar{X}) - \tilde{H}(\bar{X})^T \tilde{A}^{-1} \tilde{H}(\bar{X})\|_2 \\ & \leq \delta \|H(\bar{X})\|_2^2 \|\hat{A}^{-1}\|_2 + \|\tilde{H}(\bar{X})^T \hat{A}^{-1} H(\bar{X}) - \tilde{H}(\bar{X})^T \tilde{A}^{-1} \tilde{H}(\bar{X})\|_2,\end{aligned} \quad (\text{A.21})$$

where the first inequality is true because of the triangle inequality, the second inequality is true because of Assumption A.5.1, and the third inequality is true because  $\|G^{-1}d\|_2 \leq$

$\|G^{-1}\|_2\|d\|$  for any vector  $d$  and non-singular matrix  $G$ . The second term in (A.21) has

$$\begin{aligned}
& \|\tilde{H}(\bar{X})^T \hat{A}^{-1} H(\bar{X}) - \tilde{H}(\bar{X})^T \tilde{A}^{-1} \tilde{H}(\bar{X})\|_2 \\
& \leq \|\tilde{H}(\bar{X})^T \hat{A}^{-1} H(\bar{X}) - \tilde{H}(\bar{X})^T \hat{A}^{-1} \tilde{H}(\bar{X})\|_2 + \|\tilde{H}(\bar{X})^T \hat{A}^{-1} \tilde{H}(\bar{X}) - \tilde{H}(\bar{X})^T \tilde{A}^{-1} \tilde{H}(\bar{X})\|_2 \\
& \leq \delta \|\tilde{H}(\bar{X})\|_2 \|\hat{A}^{-1} \tilde{H}(\bar{X})\|_2 + \|\tilde{H}(\bar{X})^T (\hat{A}^{-1} - \tilde{A}^{-1}) \tilde{H}(\bar{X})\|_2 \\
& \leq \delta \|\tilde{H}(\bar{X})\|_2^2 \|\hat{A}^{-1}\|_2 + \|\hat{A}^{-1} - \tilde{A}^{-1}\|_2 \|\tilde{H}(\bar{X})\|_2^2 \\
& \leq \delta(1 + \delta)^2 \|H(\bar{X})\|_2^2 \|\hat{A}^{-1}\|_2 + (1 + \delta)^2 \|H(\bar{X})\|_2^2 \|\hat{A}^{-1} - \tilde{A}^{-1}\|_2, \tag{A.22}
\end{aligned}$$

where the first inequality is true is because of the triangle inequality, the second inequality is true because of Assumption A.5.1, the third inequality is true because  $\|G^{-1}d\|_2 \leq \|G^{-1}\|_2\|d\|$ , and the last inequality is true because by Assumption A.2.1,  $\|\tilde{H}(\bar{X})\|_2 \leq (1 + \delta)\|H(\bar{X})\|_2$ . Next,  $\|\hat{A}^{-1} - \tilde{A}^{-1}\|_2$  is bounded.

For any  $x \in \mathbb{R}^n$  such that  $\|x\|_2 = 1$ , let  $y_1, y_2 \in \mathbb{R}^n$  such that  $\hat{A}y_1 = x$  and  $\tilde{A}y_2 = x$ . Let  $\delta_A = \tilde{A} - \hat{A}$ . Thus,  $(\hat{A} + \delta_A)y_2 = x$ . Notice that by assumption,

$$\|\hat{A}^{-1}\delta_A\|_2 \leq \delta \|\hat{A}^{-1}\|_2 \|\hat{A}\|_2 = r < 1 \quad \text{and} \quad (I + \hat{A}^{-1}\delta_A)y_2 = y_1.$$

The following Lemma from [105] will be used.

**Lemma A.6.1.** *Suppose  $F \in \mathbb{R}^{n \times n}$ ,  $\|F\|_2 < 1$ . Then  $I - F$  is invertible and*

$$\|(I - F)^{-1}\|_2 \leq \frac{1}{1 - \|F\|_2},$$

where  $I$  is identity matrix in  $\mathbb{R}^{n \times n}$ .

By Lemma A.6.1, we have

$$\|y_2\|_2 \leq \|(I + \hat{A}^{-1}\delta_A)^{-1}\|_2 \|y_2\|_2 \leq \frac{1}{1 - r} \|y_1\|_2 \quad \text{and} \quad y_1 - y_2 = \hat{A}^{-1}\delta_A y_2.$$

So,

$$\|y_1 - y_2\|_2 \leq \|\hat{A}^{-1}\delta_A\|_2\|y_2\|_2 \leq \delta\|\hat{A}^{-1}\|_2\|\hat{A}\|_2\|y_2\|_2 = \frac{\delta}{1-r}\kappa(\hat{A})\|y_1\|_2.$$

Plugging in  $y_1$  and  $y_2$  gives

$$\|(\hat{A}^{-1} - \tilde{A}^{-1})x\|_2 \leq \frac{\delta}{1-r}\kappa(\hat{A})\|\hat{A}^{-1}x\|_2 \leq \frac{\delta}{1-r}\kappa(\hat{A})\|\hat{A}^{-1}\|_2 = \frac{\delta}{1-r}\kappa(\hat{A})\frac{1}{\lambda_{\min}(\hat{A})}, \quad (\text{A.23})$$

indicating

$$\|\hat{A}^{-1} - \tilde{A}^{-1}\|_2 \leq \frac{\delta}{1-r}\kappa(\hat{A})\frac{1}{\lambda_{\min}(\hat{A})}, \quad (\text{A.24})$$

since (A.23) is true for any  $x$  with  $\|x\|_2 = 1$ . Combining (A.21), (A.22), and (A.24) gives

$$\begin{aligned} & \|H(\bar{X})^T \hat{A}^{-1} H(\bar{X}) - \tilde{H}(\bar{X})^T \tilde{A}^{-1} \tilde{H}(\bar{X})\|_2 \\ & \leq \delta \|H(\bar{X})\|_2^2 \|\hat{A}^{-1}\|_2 + \delta(1+\delta)^2 \|H(\bar{X})\|_2^2 \|\hat{A}^{-1}\|_2 + (1+\delta)^2 \|H(\bar{X})\|_2^2 \|\hat{A}^{-1} - \tilde{A}^{-1}\|_2 \\ & \leq \delta \frac{\|H(\bar{X})\|_2^2}{\lambda_{\min}(\hat{A})} + \delta(1+\delta)^2 \frac{\|H(\bar{X})\|_2^2}{\lambda_{\min}(\hat{A})} + \frac{\delta(1+\delta)^2}{1-r} \kappa(\hat{A}) \frac{\|H(\bar{X})\|_2^2}{\lambda_{\min}(\hat{A})} \\ & = \frac{\delta \|H(\bar{X})\|_2^2}{\lambda_{\min}(\hat{A})} \left( 1 + (1+\delta)^2 + \frac{(1+\delta)^2}{1-r} \kappa(\hat{A}) \right). \end{aligned} \quad (\text{A.25})$$

Thus, (A.15) holds if

$$\frac{\delta \|H(\bar{X})\|_2^2}{\lambda_{\min}(\hat{A})} \left( 1 + (1+\delta)^2 + \frac{(1+\delta)^2}{1-r} \kappa(\hat{A}) \right) < \delta_1 \|H(\bar{X})^T \hat{A}^{-1} H(\bar{X})\|_2,$$

or equivalently

$$\frac{\delta \|H(\bar{X})\|_2^2}{\lambda_{\min}(\hat{A}) \lambda_{\min}(H(\bar{X})^T \hat{A}^{-1} H(\bar{X}))} \left( 1 + (1+\delta)^2 + \frac{(1+\delta)^2}{1-r} \kappa(\hat{A}) \right) < \delta_1 \kappa(H(\bar{X})^T \hat{A}^{-1} H(\bar{X})). \quad (\text{A.26})$$

Next, we simplify (A.26). Notice that the left-hand side of (A.26) has

$$\begin{aligned}
& \frac{\delta \|H(\bar{X})\|_2^2}{\lambda_{\min}(\hat{A})\lambda_{\min}(H(\bar{X})^T \hat{A}^{-1} H(\bar{X}))} \left( 1 + (1 + \delta)^2 + \frac{(1 + \delta)^2}{1 - r} \kappa(\hat{A}) \right) \\
& \leq \frac{\lambda_{\max}(\hat{A})}{\lambda_{\min}(H(\bar{X})^T H(\bar{X}))} \frac{\delta \|H(\bar{X})\|_2^2}{\lambda_{\min}(\hat{A})} \left( 1 + (1 + \delta)^2 + \frac{(1 + \delta)^2}{1 - r} \kappa(\hat{A}) \right) \\
& = \delta \kappa(\hat{A}) \kappa(H(\bar{X})^T H(\bar{X})) \left( 1 + (1 + \delta)^2 + \frac{(1 + \delta)^2}{1 - r} \kappa(\hat{A}) \right),
\end{aligned}$$

so, if

$$\delta \kappa(\hat{A}) \kappa(H(\bar{X})^T H(\bar{X})) \left( 1 + (1 + \delta)^2 + \frac{(1 + \delta)^2}{1 - r} \kappa(\hat{A}) \right) < \delta_1 \kappa(H(\bar{X})^T \hat{A}^{-1} H(\bar{X})), \tag{A.27}$$

(A.15) holds. By plugging in (A.14), we have (A.27) holds, which finishes the proof.

**APPENDIX B**  
**APPENDIX OF CHAPTER 2**

**B.1 Auxiliary tools**

In this section, we review some mathematical tools which are used in the proof of Theorem 2.5.1 in Appendix B.2.

B.1.1 Reproducing kernel Hilbert spaces

There are several equivalent ways to define the reproducing kernel Hilbert spaces. See, for example, [45, 88, 46, 62]. Here we adopt the one using the Fourier transform. See [106] and Theorem 10.12 of [46]. Let  $L_2(\mathbf{R}^d)$  be the space of complex-valued square integrable functions on  $\mathbf{R}^d$ , and  $C(\mathbf{R}^d)$  be the space of continuous real-valued functions on  $\mathbf{R}^d$ .

**Definition B.1.1.** *Let  $\Phi$  be a positive definite kernel function which is continuous and integrable in  $\mathbf{R}^d$ . Define the reproducing kernel Hilbert space  $\mathcal{N}_\Phi(\mathbf{R}^d)$  as*

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

*with the inner product*

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

A reproducing kernel Hilbert space can also be defined on a subset  $\Omega \subset \mathbf{R}^d$ , denoted by  $\mathcal{N}_\Phi(\Omega)$ . The only thing that matters in this work is that the norm  $\|f\|_{\mathcal{N}_\Phi(\Omega)}$  is the minimum value among the norms of all possible extensions of  $f$  to the whole space, i.e.,

$$\|f\|_{\mathcal{N}_\Phi(\Omega)} = \inf\{\|f_E\|_{\mathcal{N}_\Phi(\mathbf{R}^d)} : f_E \in \mathcal{N}_\Phi(\mathbf{R}^d), f_E|_\Omega = f\}, \quad (\text{B.1})$$

where  $f_E|_{\Omega}$  denotes the restriction of  $f_E$  to  $\Omega$ . See Theorem 10.48 of [46] for details.

### B.1.2 A Maximum inequality for Gaussian processes

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

The theory of bounding the maximum value of a Gaussian process is well-established in the literature. The main step of finding an upper bound is to calculate the *covering number* of the index space. Here we review the main results. Detailed discussions can be found in [47, 48].

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

$$d(t_1, t_2) = \sqrt{\mathbb{E}(Z_{t_1} - Z_{t_2})^2}. \quad (\text{B.2})$$

The  $\epsilon$ -covering number of the metric space  $(T, d)$ , denoted as  $N(\epsilon, T, d)$ , is the minimum integer  $N$  so that there exist  $N$  distinct balls in  $(T, d)$  with radius  $\epsilon$ , and the union of these balls covers  $T$ . Let  $D$  be the diameter of  $T$ . The supremum of a Gaussian process is closely tied to a quantity called the *entropy integral*, defined as

$$\int_0^{D/2} \sqrt{\log N(\epsilon, T, d)} d\epsilon. \quad (\text{B.3})$$

Theorem B.1.1 gives a maximum inequality for Gaussian processes, which is an equivalent statement of Corollary 2.2.8 of [47]. Also see Theorems 1.3.3 and 2.1.1 of [48].

**Theorem B.1.1.** *Let  $Z_t$  be a centered separable Gaussian process on a  $d$ -compact  $T$ ,  $d$  the metric, and  $N$  the  $\epsilon$ -covering number. Then there exists a universal constant  $K$  such that*

for all  $u > 0$ ,

$$\mathbb{P}(\sup_{t \in T} |Z_t| > K \int_0^{D/2} \sqrt{\log N(\epsilon, T, d)} d\epsilon + u) \leq 2e^{-u^2/2\sigma_T^2}, \quad (\text{B.4})$$

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

## B.2 Proof of Theorem 2.5.1

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

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

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

The rest of our proof consists of the following steps. In step 1, we bound the covering number  $N(\epsilon, \Omega, d)$ . Next we bound the diameter  $D$ . In step 3, we invoke Theorem B.1.1 to obtain a bound for the entropy integral. In the last step, we use (B.4) to obtain the desired results.

### Step 1: Bounding the covering number

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

can verified that

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

By Condition 2.5.1,  $h \in \mathcal{N}_{\Phi}(\mathbf{R}^d)$ , since  $\Psi(\mathbf{x} - \cdot) \in \mathcal{N}_{\Phi}(\mathbf{R}^d)$  for any  $\mathbf{x} \in \Omega$ . Thus, by (2.8),

$$d^2(\mathbf{x}, \mathbf{x}') \leq 2P_{\Phi, \mathbf{x}}(\|h\|_{\mathcal{N}_{\Phi}(\mathbf{R}^d)} + \|h_1\|_{\mathcal{N}_{\Phi}(\mathbf{R}^d)}). \quad (\text{B.5})$$

By Definition B.1.1,

$$\|h\|_{\mathcal{N}_{\Phi}(\mathbf{R}^d)}^2 = (2\pi)^{-d} \int_{\mathbf{R}^d} \frac{|\tilde{h}(\boldsymbol{\omega})|^2}{\tilde{\Phi}(\boldsymbol{\omega})} d\boldsymbol{\omega}. \quad (\text{B.6})$$

Under Condition 2.5.1, by (B.6),

$$\|h\|_{\mathcal{N}_{\Phi}(\mathbf{R}^d)}^2 = (2\pi)^{-d} \int_{\mathbf{R}^d} \frac{|\tilde{h}(\boldsymbol{\omega})|^2}{\tilde{\Phi}(\boldsymbol{\omega})} d\boldsymbol{\omega} \leq A_1^2 (2\pi)^{-d} \int_{\mathbf{R}^d} \frac{|\tilde{h}(\boldsymbol{\omega})|^2}{\tilde{\Psi}(\boldsymbol{\omega})} d\boldsymbol{\omega} = A_1^2 \|h\|_{\mathcal{N}_{\Psi}(\mathbf{R}^d)}^2. \quad (\text{B.7})$$

However, since  $h(\cdot) = \Psi(\mathbf{x} - \cdot) - \Psi(\mathbf{x}' - \cdot)$ ,  $\|h\|_{\mathcal{N}_{\Psi}(\mathbf{R}^d)}^2 = \Psi(\mathbf{x} - \mathbf{x}) - 2\Psi(\mathbf{x}' - \mathbf{x}) + \Psi(\mathbf{x}' - \mathbf{x}')$ . Thus, by Fourier transform and the mean value theorem,

$$\begin{aligned} \|h\|_{\mathcal{N}_{\Psi}(\mathbf{R}^d)}^2 &= \Psi(\mathbf{x} - \mathbf{x}) - 2\Psi(\mathbf{x}' - \mathbf{x}) + \Psi(\mathbf{x}' - \mathbf{x}') \\ &= 2(2\pi)^{-d} \int_{\mathbf{R}^d} (1 - e^{i(\mathbf{x} - \mathbf{x}')^T \boldsymbol{\omega}}) \tilde{\Psi}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\ &\leq \left( 2(2\pi)^{-d} \int_{\mathbf{R}^d} \|\boldsymbol{\omega}\| \tilde{\Psi}(\boldsymbol{\omega}) d\boldsymbol{\omega} \right) \|\mathbf{x} - \mathbf{x}'\|. \end{aligned} \quad (\text{B.8})$$

By Condition 2.5.1, there exists a constant  $C_1$  such that

$$\left(2(2\pi)^{-d} \int_{\mathbf{R}^d} \|\omega\| \tilde{\Psi}(\omega) d\omega\right) \leq C_1,$$

which implies

$$\|h\|_{\mathcal{N}_\Psi(\mathbf{R}^d)}^2 \leq C_1 \|\mathbf{x} - \mathbf{x}'\|. \quad (\text{B.9})$$

Now we consider  $h_1(\cdot)$ . It follows from a similar argument that  $\|h_1\|_{\mathcal{N}_\Phi(\mathbf{R}^d)}^2 \leq A_1^2 \|h_1\|_{\mathcal{N}_\Psi(\mathbf{R}^d)}^2$ .

Since  $h_1(\cdot) = \mathbf{r}^T(\mathbf{x})\mathbf{K}^{-1}\mathbf{r}_1(\cdot) - \mathbf{r}^T(\mathbf{x}')\mathbf{K}^{-1}\mathbf{r}_1(\cdot)$ ,  $\|h_1\|_{\mathcal{N}_\Psi(\mathbf{R}^d)}^2 = (\mathbf{r}(\mathbf{x}') - \mathbf{r}(\mathbf{x}))^T \mathbf{K}^{-1} \mathbf{K}_1 \mathbf{K}^{-1} (\mathbf{r}(\mathbf{x}') - \mathbf{r}(\mathbf{x}))$ .

For any  $\mathbf{u} = (u_1, \dots, u_n)$ ,

$$\begin{aligned} & \sum_{j,k=1}^n u_j \bar{u}_k \Psi(\mathbf{x}_j - \mathbf{x}_k) \\ &= \frac{1}{(2\pi)^d} \int_{\mathbf{R}^d} \sum_{j,k=1}^n u_j \bar{u}_k e^{i(\mathbf{x}_j - \mathbf{x}_k)^T \omega} \tilde{\Psi}(\omega) d\omega \\ &= \frac{1}{(2\pi)^d} \int_{\mathbf{R}^d} \left| \sum_{j=1}^n u_j e^{i\mathbf{x}_j^T \omega} \right|^2 \tilde{\Psi}(\omega) d\omega. \end{aligned} \quad (\text{B.10})$$

Thus, by Condition 2.5.1,

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

We plug in  $\mathbf{u} = \mathbf{K}^{-1}(\mathbf{r}(\mathbf{x}') - \mathbf{r}(\mathbf{x}))$  to get

$$\|h_1\|_{\mathcal{N}_\Psi(\mathbf{R}^d)}^2 \leq A_1^2(\mathbf{r}(\mathbf{x}') - \mathbf{r}(\mathbf{x}))^T \mathbf{K}^{-1}(\mathbf{r}(\mathbf{x}') - \mathbf{r}(\mathbf{x})). \quad (\text{B.11})$$

Let  $h_2(\cdot) = \Phi(\cdot - \mathbf{x}') - \Phi(\cdot - \mathbf{x})$ , thus,  $\mathcal{I}_{\Phi, \mathbf{X}} h_2(\cdot) = \mathbf{r}^T(\cdot) \mathbf{K}^{-1}(\mathbf{r}(\mathbf{x}') - \mathbf{r}(\mathbf{x}))$ . By (2.8) and note that  $\|h_2\|_{\mathcal{N}_\Phi(\mathbf{R}^d)}^2 = \Phi(\mathbf{x} - \mathbf{x}) - 2\Phi(\mathbf{x}' - \mathbf{x}) + \Phi(\mathbf{x}' - \mathbf{x}')$ ,

$$\begin{aligned} & (\mathbf{r}(\mathbf{x}') - \mathbf{r}(\mathbf{x}))^T \mathbf{K}^{-1}(\mathbf{r}(\mathbf{x}') - \mathbf{r}(\mathbf{x})) \\ & \leq |h_2(\mathbf{x}') - \mathcal{I}_{\Phi, \mathbf{X}} h_2(\mathbf{x}')| + |h_2(\mathbf{x}) - \mathcal{I}_{\Phi, \mathbf{X}} h_2(\mathbf{x})| + |h_2(\mathbf{x}')| + |h_2(\mathbf{x})| \\ & \leq 2P_{\Phi, \mathbf{X}} \sqrt{\Phi(\mathbf{x} - \mathbf{x}) - 2\Phi(\mathbf{x}' - \mathbf{x}) + \Phi(\mathbf{x}' - \mathbf{x}')} + 2(\Phi(\mathbf{x} - \mathbf{x}) - 2\Phi(\mathbf{x}' - \mathbf{x}) + \Phi(\mathbf{x}' - \mathbf{x}')) \\ & \leq 2(P_{\Phi, \mathbf{X}} + \sqrt{\Phi(\mathbf{x} - \mathbf{x}) - 2\Phi(\mathbf{x}' - \mathbf{x}) + \Phi(\mathbf{x}' - \mathbf{x}')})) \sqrt{\Phi(\mathbf{x} - \mathbf{x}) - 2\Phi(\mathbf{x}' - \mathbf{x}) + \Phi(\mathbf{x}' - \mathbf{x}')} \\ & \leq 2(P_{\Phi, \mathbf{X}} + 2) \sqrt{\Phi(\mathbf{x} - \mathbf{x}) - 2\Phi(\mathbf{x}' - \mathbf{x}) + \Phi(\mathbf{x}' - \mathbf{x}')} \end{aligned} \quad (\text{B.12})$$

Thus, if  $P_{\Phi, \mathbf{X}} < 1$ , by a similar argument in (B.8), and together with (B.11) and (B.12), we have

$$\|h_1\|_{\mathcal{N}_\Psi(\mathbf{R}^d)}^2 \leq C_2 A_1^4 \|\mathbf{x} - \mathbf{x}'\|^{1/2}, \quad (\text{B.13})$$

where  $C_2$  is a constant.

In view of (B.5), (B.9) and (B.13), there exists a constant  $C_3$  such that

$$d^2(\mathbf{x}, \mathbf{x}') \leq C_3 (A_1 + A_1^2) P_{\Phi, \mathbf{X}} \|\mathbf{x} - \mathbf{x}'\|^{1/4}, \quad (\text{B.14})$$

provided that  $\|\mathbf{x} - \mathbf{x}'\| < 1$ .

Therefore, the covering number can be bounded as

$$\log N(\epsilon, \Omega, d) \leq \log N\left(\frac{\epsilon^8}{C_3^2 (A_1 + A_1^2)^2 P_{\Phi, \mathbf{X}}^2}, \Omega, \|\cdot\|\right). \quad (\text{B.15})$$

The right side of (B.15) involves the covering number of a Euclidean ball, which is well understood in the literature. See Lemma 2.5 of [92]. This result leads to the bound

$$\log N(\epsilon, \Omega, d) \leq C_4 \log \left( 1 + \frac{C_5 (A_1 + A_1^2)^{1/4} P_{\Phi, \mathbf{X}}^{1/4}}{\epsilon} \right), \quad (\text{B.16})$$

provided that

$$\epsilon^4 < C_3 (A_1 + A_1^2) P_{\Phi, \mathbf{X}}, \quad (\text{B.17})$$

where  $C_4$  and  $C_5$  are two constants.

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

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

$$\begin{aligned} d^2(\mathbf{x}, \mathbf{x}') &= \mathbb{E}(g(\mathbf{x}) - g(\mathbf{x}'))^2 \leq 4 \sup_{\mathbf{x} \in \Omega} \mathbb{E}(g(\mathbf{x}))^2 \\ &= 4 \sup_{\mathbf{x} \in \Omega} \mathbb{E}(Z(\mathbf{x}) - \mathcal{I}_{\Phi, \mathbf{X}} Z(\mathbf{x}))^2 \\ &= 4 \sup_{\mathbf{x} \in \Omega} (\Psi(\mathbf{x} - \mathbf{x}) - 2\mathbf{r}_1^T(\mathbf{x})\mathbf{K}^{-1}\mathbf{r}(\mathbf{x}) + \mathbf{r}^T(\mathbf{x})\mathbf{K}^{-1}\mathbf{K}_1\mathbf{K}^{-1}\mathbf{r}(\mathbf{x})), \end{aligned} \quad (\text{B.18})$$

where  $\mathbf{r}$ ,  $\mathbf{r}_1$ ,  $\mathbf{K}$  and  $\mathbf{K}_1$  are defined in the beginning of Appendix B.2.

Combining identity (B.10) with

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

for any  $\mathbf{u} = (u_1, \dots, u_n)$ , under Condition 2.5.1, we have

$$\begin{aligned}
& \mathbf{u}^T \mathbf{K}_1 \mathbf{u} - 2\mathbf{u}^T \mathbf{r}_1(\mathbf{x}) + \Psi(\mathbf{x} - \mathbf{x}) \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{i\mathbf{x}_j^T \boldsymbol{\omega}} - e^{i\mathbf{x}^T \boldsymbol{\omega}} \right|^2 \tilde{\Psi}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\
&\leq \frac{A_1^2}{(2\pi)^d} \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{i\mathbf{x}_j^T \boldsymbol{\omega}} - e^{i\mathbf{x}^T \boldsymbol{\omega}} \right|^2 \tilde{\Phi}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\
&= A_1^2 (\mathbf{u}^T \mathbf{K} \mathbf{u} - 2\mathbf{u}^T \mathbf{r}(\mathbf{x}) + \Phi(\mathbf{x} - \mathbf{x})). \tag{B.19}
\end{aligned}$$

We can combine (B.19) with (B.18) by substituting  $\mathbf{u}$  in (B.19) by  $\mathbf{K}^{-1} \mathbf{r}(\mathbf{x})$  and arrive at

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

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

$$D \leq 2A_1 P_{\Phi, \mathbf{X}}. \tag{B.20}$$

### Step 3: Bounding the entropy integral

Under Condition 2.5.1, if  $P_{\Phi, \mathbf{X}} < \min\{1, \sqrt[3]{C_3(1 + A_1)}/A_1, \sqrt[3]{C_5^4(1 + A_1)^2}/A_1, A_1/(2\sqrt[3]{C_5^4(1 + A_1)^2})\} := C$ , (B.17) is satisfied for all  $\epsilon \in [0, D/2]$ . Thus, by (B.16) and (B.20),

$$\begin{aligned}
\int_0^{D/2} \sqrt{\log N(\epsilon, \Omega, d)} d\epsilon &\leq \int_0^{A_1 P_{\Phi, \mathbf{X}}} \sqrt{C_4 \log \left( 1 + \frac{C_5 (A_1 + A_1^2)^{1/4} P_{\Phi, \mathbf{X}}^{1/4}}{\epsilon} \right)} d\epsilon \\
&\leq C_6 \sqrt{1 + A_1} A_1 P_{\Phi, \mathbf{X}} \sqrt{\log(C_5 (1 + A_1)^{1/2} / (A_1^{3/4} P_{\Phi, \mathbf{X}}^{3/4}) + 1)} \\
&\leq C_7 \sqrt{1 + A_1} A_1 P_{\Phi, \mathbf{X}} \sqrt{\log(1/P_{\Phi, \mathbf{X}})}, \tag{B.21}
\end{aligned}$$

where  $C_6$  and  $C_7$  are constants and the second inequality follows from the Cauchy-Schwartz

inequality.

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

Noting that  $\sup_{\mathbf{x} \in \Omega} \mathbb{E}(Z(\mathbf{x}) - \mathcal{I}_{\Phi, \mathbf{x}} Z(\mathbf{x}))^2 = D^2$ , by plugging (B.20) into (B.4), we obtain the desired inequality, which completes the proof.

## APPENDIX C

### APPENDIX OF CHAPTER 3

#### C.1 Upper and Lower Bounds of the modified likelihood function

**Theorem C.1.1.** *Suppose  $m_0$  and  $f$  as in Theorem 3.3.1. Let  $\epsilon_n$  be defined in Theorem 3.3.1. Then for some constant  $C_0$ , when  $n \geq C_0$ , the following statements are true.*

1. *If  $f \in H^{m_0}(\Omega)$ , then with probability at least  $1 - C_1 \exp(-C_2 n^{\eta_1})$ ,*

$$\ell(m; X, Y, \mu) \begin{cases} \leq \frac{m_0}{2m+d} n \log n + \frac{n}{2} \log C_3 h_1(n) & \text{for all } m \in [m_0 + \epsilon_n, m_{\max}], \\ \leq \frac{m}{2m+d} n \log n + \frac{n}{2} \log(C_4 \log n) & \text{for all } m \in [m_{\min}, m_0 - \epsilon_n], \\ \geq \frac{m_0}{2m_0+d} n \log n - \frac{n}{2} \log(C_5 \log n) & \text{for } m = m_0. \end{cases}$$

2. *If  $f \notin H^{m_0}(\Omega)$ , then with probability at least  $1 - C_1 \exp(-C_2 n^{\eta_1})$ ,*

$$\ell(m; X, Y, \mu) \begin{cases} \leq \frac{m_0}{2m+d} n \log n + n \log(C_3 \log n) & \text{for all } m \in [m_0 + \epsilon_n, m_{\max}], \\ \leq \frac{m}{2m+d} n \log n + \frac{n}{2} \log(C_4 \log n) & \text{for all } m \in [m_{\min}, m_0 - \epsilon_n], \\ \geq \frac{m_0}{2m_0+d} n \log n - \frac{n}{2} \log C_5 h_2(n) & \text{for } m = m_0. \end{cases}$$

*In the above statements,  $C_i$ 's and  $\eta_i$ 's are constants depending on  $f$ ,  $\Omega$  and  $C$ .  $h_1(n)$  and  $h_2(n)$  are defined in (3.18) and (3.20), respectively.*

Before the proof of Theorem C.1.1, we introduce some lemmas used in this section. Lemma C.1.1 states the lower bound of  $\frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2$  when  $m \in [m_0 + \epsilon_n, m_{\max}]$ . Lemma C.1.2 states the lower bound of  $\frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2$  when  $m \in [m_{\min}, m_0 - \epsilon_n]$ . Lemma C.1.3 states the upper bound of  $\frac{\mu_{m_0}}{n} \|\hat{f}_{m_0}\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^2$ . The proofs of Lemmas C.1.1, C.1.2, and C.1.3 can be found in Appendices C.2, C.3, and C.4, respectively.

**Lemma C.1.1.** Suppose  $m \in [m_0 + \epsilon_n, m_{\max}]$ . Let  $\mu_m = Cn^{\frac{d}{2m+d}}$ , where  $C$  is any fixed constant. Let  $t_m = C_0n^{-2m_0/(2m+d)}/h_1(n)$  if  $f \in H^{m_0}$ , and  $t_m = C_0n^{-2m_0/(2m+d)}/(\log n)^2$  if  $f \notin H^{m_0}$  for any constant  $C_0 < c$ , where  $h_1(n)$  is defined in (3.18). Let  $\mathcal{C}$  denote the class  $\{\hat{f}_m : \forall m \in [m_0 + \epsilon_n, m_{\max}], \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq t_m\}$ . Under class  $\mathcal{C}$ , with probability at least  $1 - C_1 \exp(-C_2n^{\eta_1})$ ,

$$\frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \geq \begin{cases} C_3n^{\frac{-2m_0}{2m+d}}/h_1(n) & \text{if } f \in H^{m_0}, \\ C_3n^{\frac{-2m_0}{2m+d}}/(\log n)^2 & \text{if } f \notin H^{m_0}, \end{cases}$$

where  $C_i$  for  $i = 1, 2, 3, c$ , and  $\eta_1$  are constants depending on  $f$  and  $\Omega$ .

**Lemma C.1.2.** Suppose  $m \in [m_{\min}, m_0 - \epsilon_n]$ . Let  $\mu_m = Cn^{\frac{d}{2m+d}}$ , where  $C$  is any fixed constant. Let  $t_m = C_0n^{-2m/(2m+d)}/\log n$  for any constant  $C_0 < c$ . Let  $\mathcal{C} = \{\hat{f}_m : \forall m \in [m_{\min}, m_0 - \epsilon_n], \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq t_m\}$ . Under class  $\mathcal{C}$ , with probability at least  $1 - C_1 \exp(-C_2n^{\eta_1})$ ,

$$\frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \geq C_3n^{-2m/(2m+d)}/\log n,$$

where  $C_i$  for  $i = 1, 2, 3, c$ , and  $\eta_1$  are constants depending on  $f$  and  $\Omega$ .

**Lemma C.1.3.** Let  $\mu_{m_0} = Cn^{\frac{d}{2m_0+d}}$ , where  $C$  is any fixed constant. Let  $t = C_0n^{-2m_0/(2m_0+d)} \log n$  if  $f \in H^{m_0}$ , and  $t = C_0n^{-2m_0/(2m_0+d)}h_2(n)$  if  $f \notin H^{m_0}$  for any constant  $C_0 < c$ , where  $h_2(n)$  is defined in (3.20). Let  $\mathcal{C} = \{\hat{f}_{m_0} : \frac{\mu_{m_0}}{n} \|\hat{f}_{m_0}\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^2 \geq t\}$ . Under class  $\mathcal{C}$ , with probability at least  $1 - C_1 \exp(-C_2n^{d/(4(2m_0+d))})$ ,

$$\frac{\mu_{m_0}}{n} \|\hat{f}_{m_0}\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^2 \leq \begin{cases} C_3n^{\frac{-2m_0}{2m_0+d}} & \text{if } f \in H^{m_0}, \\ C_3n^{\frac{-2m_0}{2m_0+d}}h_2(n) & \text{if } f \notin H^{m_0}, \end{cases}$$

where  $C_i$  for  $i = 1, 2, 3$ , and  $c$  are constants depending on  $f$  and  $\Omega$ .

Now we are ready to present the proof of Theorem C.1.1.

*Proof of Theorem C.1.1:*

Note that

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

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

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

Let  $\tilde{Y} = (\tilde{y}_1, \dots, \tilde{y}_n)^T$  be the solution to the right-hand side of (C.1). It can be verified that

$$\begin{aligned} & \frac{\mu_m}{n} Y^T (K_m + \mu_m I_n)^{-1} K_m (K_m + \mu_m I_n)^{-1} Y \\ &= \min_{\hat{u}} \frac{1}{n} (Y - K_m \hat{u})^T (Y - K_m \hat{u}) + \frac{\mu_m}{n} \hat{u}^T K_m \hat{u} - \frac{1}{n} \sum_{i=1}^n (\tilde{y}_i - y_i)^2. \end{aligned} \quad (\text{C.2})$$

Therefore, by the representer theorem and (3.1), the right-hand side of (C.2) is the same as

$$\min_{\hat{f} \in \mathcal{N}_{\Psi_m}(\Omega)} \left( \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \frac{\mu_m}{n} \|\hat{f}\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \right) - \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}_m(x_i))^2, \quad (\text{C.3})$$

where  $\hat{f}_m$  is defined in (3.2) and  $\|\cdot\|_{\mathcal{N}_{\Psi_m}(\Omega)}$  denotes the norm of reproducing kernel Hilbert space with kernel function  $\Psi_m(\cdot, \cdot)$ . By Corollary 10.48 in [46],  $\|\cdot\|_{\mathcal{N}_{\Psi_m}(\Omega)}$  is equivalent to  $\|\cdot\|_{H^m(\Omega)}$ .

Combining

$$\begin{aligned} & \min_{\hat{f} \in \mathcal{N}_{\Psi_m}(\Omega)} \left( \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \frac{\mu_m}{n} \|\hat{f}\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \right) - \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}_m(x_i))^2 \\ &= \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \end{aligned}$$

with (C.1), (C.2) and (C.3), it suffices to obtain the bounds of  $\|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2$ .

Next we consider the three cases described in Theorem C.1.1.

**Case 1:**  $m \in [m_0 + \epsilon_n, m_{\max}]$

Let  $t_m = C_0 n^{-2m_0/(2m+d)}/h_1(n)$  if  $f \in H^{m_0}$ , and  $t_m = C_0 n^{-2m_0/(2m+d)}/(\log n)^2$  if  $f \notin H^{m_0}$ , where  $C_0$  is some constant determined later, and  $h_1(n)$  is defined in (3.18).

In order to show the bounds stated in Theorem C.1.1, it is enough to show that the probability

$$P\left(\frac{\mu}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq t_m\right)$$

can be bounded by some small number.

By Lemma C.1.1 and taking  $C_0 = \min\{C_3, c\}/2$ , where  $c$  is as in Lemma C.1.1, we have

$$P\left(\frac{\mu}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq t_m\right) \leq p_0, \quad (\text{C.4})$$

where  $p_0 = C_4 \exp(-C_5 n^{\eta_1})$ .

**Case 2:**  $m \in [m_{\min}, m_0 - \epsilon_n]$

Let  $t_m = C_0 n^{-2m/(2m+d)}/\log n$  for some constant  $C_0$  which will be determined later.

Similar to Case 1, it suffices to show the probability

$$P\left(\frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq t_m\right)$$

is bounded by some small number.

By Lemma C.1.2 and taking  $C_0 = \min\{C_3, c\}/2$ , we have

$$P\left(\frac{\mu}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq t_m\right) \leq p_1, \quad (\text{C.5})$$

where  $p_1 = C_6 \exp(-C_7 n^{\eta_2})$ .

**Case 3:**  $m = m_0$

Let  $t = C_0 n^{-2m_0/(2m_0+d)} \log n$  if  $f \in H^{m_0}$ , and  $t = C_0 n^{-2m_0/(2m_0+d)} h_2(n)$  if  $f \notin H^{m_0}$ , where  $C_0$  is a constant determined later. In this case, it suffices to show the probability

$$P\left(\frac{\mu_{m_0}}{n} \|\hat{f}_{m_0}\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^2 \geq t\right)$$

is bounded by some small number.

By Lemma C.1.3, if we choose  $C_0 \geq 2C_3$ ,

$$P\left(\frac{\mu_{m_0}}{n} \|\hat{f}_{m_0}\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^2 \geq t\right) \leq p_3, \quad (\text{C.6})$$

where  $p_3 = C_8 \exp(-C_9 n^{n_3})$ .

By combining (C.4), (C.5) and (C.6), we finish the proof.

## C.2 Proof of Lemma C.1.1

We first present the lemmas and the theorem used in the proof of Lemma C.1.1. Lemma C.2.1 is from [107], which states the discrepancy of the empirical norm and the  $L_2$  norm. Lemma C.2.2 states that the absolute value of the empirical inner product  $|\langle e, f - \hat{f}_m \rangle_n|$  is small when  $m \in [m_0 + \epsilon_n, m_{\max}]$ , where  $\hat{f}_m$  is defined in (3.2). The proof of Lemma C.2.2 can be found in Appendix C.5.

In the rest of Appendix 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.

**Lemma C.2.1** (Theorem 2.1 in [107]). *Let  $R := \sup_{f \in \mathcal{F}} \|f\|_2$ ,  $K := \sup_{f \in \mathcal{F}} \|f\|_\infty$ , where  $\mathcal{F}$  is a class. Then for all  $t > 0$ , with probability at least  $1 - \exp(-t)$ ,*

$$\sup_{f \in \mathcal{F}} \left| \|f\|_n^2 - \|f\|_2^2 \right| \leq C_1 \left( \frac{2RJ_\infty(K, \mathcal{F}) + RK\sqrt{t}}{\sqrt{n}} + \frac{4J_\infty^2(K, \mathcal{F}) + K^2t}{n} \right),$$

where  $C_1$  is a constant, and

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

with  $C_2$  another constant.

**Lemma C.2.2.** *Suppose  $m \in [m_0 + \epsilon_n, m_{\max}]$ . Let  $\mu_m = C n^{\frac{d}{2m+d}}$ , where  $C$  is an any fixed constant. Let  $t_m = C_0 n^{-2m_0/(2m+d)}/h_1(n)$  if  $f \in H^{m_0}$ , and  $t_m = C_0 n^{-2m_0/(2m+d)}/(\log n)^2$  if  $f \notin H^{m_0}$ , where  $C_0$  is some constant, and  $h_1(n)$  is defined in (3.18). Let  $\mathcal{C}$  denote the class  $\{\hat{f}_m : \forall m \in [m_0 + \epsilon_n, m_{\max}], \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq t_m\}$ . Under class  $\mathcal{C}$ , with probability at least  $1 - C_1 \exp(-C_2 n^{\eta_0})$ ,*

$$2|\langle e, f - \hat{f}_m \rangle_n| \leq n^{-\eta_1/2} t_m,$$

where  $C_1, C_2, \eta_0$ , and  $\eta_1$  are constants depending on  $f$  and  $\Omega$ . In particular,  $|\langle e, f - \hat{f}_m \rangle_n| = o_p(t_m)$ .

Now we are ready to prove Lemma C.1.1.

*Proof of Lemma C.1.1:*

Let  $\lambda_2 = \mu_m/n$ , and  $\lambda_1 = C_1 \lambda_2$ , where  $C_1$  is a constant which will be determined later.

Let

$$f_1 = \operatorname{argmin}_{\hat{f} \in \mathcal{N}_{\Psi_m}(\Omega)} \|f - \hat{f}\|_n^2 + \lambda_1 \|\hat{f}\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2,$$

and

$$f_m^* = \operatorname{argmin}_{\hat{f} \in \mathcal{N}_{\Psi_m}(\mathbb{R}^d)} \|f - \hat{f}\|_2^2 + \frac{C_2 \mu_m}{n} \|\hat{f}\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2, \quad (\text{C.7})$$

where  $C_2$  is a constant. From (3.2), it can be seen that

$$\begin{aligned} & \|f - f_1\|_n^2 + \lambda_1 \|f_1\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq \|f - \hat{f}_m\|_n^2 + \lambda_1 \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \\ & \leq 2\langle e, \hat{f}_m - f_m^* \rangle + \|f - f_m^*\|_n^2 + \lambda_2 \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 + \lambda_1 \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2, \end{aligned} \quad (\text{C.8})$$

which indicates

$$\lambda_1 \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \geq \|f - f_1\|_n^2 + \lambda_1 \|f_1\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 - (\|f - f_m^*\|_n^2 + \lambda_2 \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2) - 2\langle e, \hat{f}_m - f_m^* \rangle \quad (\text{C.9})$$

by rearrangement of (C.8). Therefore, it is enough to show that under class  $\mathcal{C}$ , with probability at least  $1 - C_1 \exp(-C_2 n^m)$ ,

$$\begin{aligned} & \|f - f_1\|_n^2 + \frac{C_1 \mu_m}{n} \|f_1\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 - (\|f - f_m^*\|_n^2 + \frac{\mu_m}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2) - 2\langle e, \hat{f}_m - f_m^* \rangle \\ & \geq \begin{cases} C_4 n^{\frac{-2m_0}{2m+a}} / h_1(n) & \text{if } f \in H^{m_0}, \\ C_5 n^{\frac{-2m_0}{2m+a}} / (\log n)^2 & \text{if } f \notin H^{m_0}, \end{cases} \end{aligned}$$

where  $C_i$ 's, and  $\eta_1$  are constants.

Let  $m_0 + \epsilon_n = m_1 < m_2 < \dots < m_p = m_{\max}$  be a partition of  $[m_0 + \epsilon_n, m_{\max}]$ , and  $\mathcal{G}_i = \{g : g = f - \hat{f}_m, \forall m \in [m_i, m_{i+1}], \hat{f}_m \in \mathcal{C}\}$ . Let  $\mathcal{G} = \bigcup \mathcal{G}_i$ . We will use Lemma C.2.1. First, we calculate the quantity  $J_\infty^2(K, \mathcal{G}'_i)$ , in which we need to calculate the bracket entropy of class  $\mathcal{G}_i$ . Consider the class  $\mathcal{F}_i = \{g : g = \hat{f}_m, \forall m \in [m_i, m_{i+1}], \hat{f}_m \in \mathcal{C}\}$ , which has the same bracket entropy as class  $\mathcal{G}_i$ . Note that  $\mathcal{F}_i \subset H^{m_i}$ . Let  $t_i = t_{m_i}$  and  $\mu_i = \mu_{m_i}$ . Define  $\rho_m = (nt_m/\mu_m)^{1/2}$  and  $\rho_i = (nt_i/\mu_i)^{1/2}$ . Therefore,  $\|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq \rho_m^2$  for all  $m$ . Let  $\mathcal{F}'_i = \{g : g = \hat{f}_m/\rho_m, \forall m \in [m_i, m_{i+1}], g \in \mathcal{F}'_i\}$  and  $\mathcal{G}'_i = \{g : g = (\hat{f}_m - f_0)/\rho_m, \forall m \in [m_i, m_{i+1}], \hat{f}_m \in \mathcal{C}\}$ .

Since  $\mathcal{F}'_i \subset H^{m_i}$ , the bracket entropy satisfies

$$H_B(\delta_n/V(\Omega), \mathcal{F}'_i, \|\cdot\|_\infty) \leq C_3 \left( \frac{1}{\delta_n} \right)^{d/m_i}.$$

Therefore, since  $\mathcal{G}'_i$  and  $\mathcal{F}'_i$  have the same bracket entropy number, by the relation between the entropy number and the bracket entropy number, we have

$$\begin{aligned} J_\infty^2(K, \mathcal{G}'_i) &= C_4^2 \inf_{\delta > 0} E \left[ K \int_\delta^1 \sqrt{H(uK/2, \mathcal{F}'_i, \|\cdot\|_{n,\infty})} du + \sqrt{n}\delta z \right]^2 \\ &\leq C_5^2 \left[ K \int_0^1 \left( \frac{1}{uK} \right)^{d/(2m_i)} du \right]^2 \\ &= C_6^2 K^2 \left( \frac{1}{K} \right)^{d/m_i}. \end{aligned}$$

By the interpolation inequality and the boundedness of  $\|\hat{f}_m\|_2$ ,

$$\|\hat{f}_m\|_\infty \leq C_7 \|\hat{f}_m\|_2^{1-\frac{d}{2m}} \|\hat{f}_m\|_{H_m}^{\frac{d}{2m}} \leq C_8 \|\hat{f}_m\|_{\mathcal{N}_{K_m}(\Omega)}^{\frac{d}{2m}} \leq C_8 \rho_m^{\frac{d}{2m}}.$$

Since  $\rho_m \rightarrow \infty$  as  $n \rightarrow \infty$ , and  $\rho_m^{\frac{d-2m}{2m}}$  decreases when  $m$  increases, we have

$$\|g\|_\infty \leq C_9 \max_{m \in [m_i, m_{i+1}]} \rho_m^{\frac{d-2m}{2m}} \leq C_9 \rho_i^{\frac{d-2m_i}{2m_i}},$$

for  $g \in \mathcal{G}'_i$ . Therefore, we have  $K \leq C_9 \rho_i^{\frac{d-2m_i}{2m_i}}$ . By Lemma C.2.1, we have

$$\begin{aligned} \sup_{g \in \mathcal{G}_i} \left| \|g\|_n^2 - \|g\|_2^2 \right| &\leq \rho_{i+1}^2 \sup_{g \in \mathcal{G}'_i} \left| \|g\|_n^2 - \|g\|_2^2 \right| \\ &\leq C_{10} \rho_{i+1}^2 \left( \frac{2RJ_\infty(K, \mathcal{G}'_i) + RK\sqrt{t}}{\sqrt{n}} + \frac{4J_\infty^2(K, \mathcal{G}'_i) + K^2t}{n} \right) \end{aligned}$$

with probability at least  $1 - \exp(-t)$ .

By Lemma C.2.2 and its proof, there exist constants  $\eta_i$ 's and  $C_j$ 's such that with prob-

ability at least  $1 - C_{11} \exp(-C_{12}n^{\eta_1}) - C_{13} \exp(-C_{14}n^{\eta_0})$ ,

$$\begin{aligned}
\eta_2 \|f - \hat{f}_m\|_2^2 &\leq \frac{1}{n} \sum_{i=1}^n (f(x_i) - \hat{f}_m(x_i))^2 + \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \\
&\leq \frac{1}{n} \sum_{i=1}^n (f(x_i) - f_m^*(x_i))^2 + \frac{2}{n} \sum_{i=1}^n e_i(\hat{f}_m(x_i) - f_m^*(x_i)) + \frac{\mu_m}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \\
&\leq \eta_3 \|f - f_m^*\|_2^2 + \frac{2}{n} \sum_{i=1}^n e_i(\hat{f}_m(x_i) - f_m^*(x_i)) + \frac{\mu}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \\
&\leq C_{15} n^{-\frac{2m_{\min}}{2m+d}},
\end{aligned}$$

where  $f_m^*$  is defined in (C.7). Therefore, we have

$$R = \sup_{g \in \mathcal{G}'_i} \|g\|_2 \leq \sup_{m \in [m_i, m_{i+1}]} C_{16} n^{-\frac{m_{\min}}{2m+d}} / \rho_m = C_{16} n^{-\frac{m_{\min}}{2m_{i+1}+d}} / \rho_i.$$

Since  $J_\infty(K, \mathcal{G}'_i) \leq C_6 K^{\frac{2m_i-d}{2m_i}}$  and  $K \leq C_9 \rho_i^{\frac{d-2m_i}{2m_i}}$ , we have  $J_\infty(K, \mathcal{G}'_i) \leq C_{17} \rho_i^{-\left(\frac{d-2m_i}{2m_i}\right)^2}$ .

By taking  $m_{i+1} - m_i \leq C_{18} \log h_1(n) / \log n$  if  $f \in H^{m_0}$  and  $m_{i+1} - m_i \leq C_{19} \log \log n / \log n$  if  $f \notin H^{m_0}$ , it can be verified that there exist constant  $\eta_2$  and  $\eta_3$  such that

$$\rho_{i+1}^2 \frac{R J_\infty(K, \mathcal{G}'_i)}{\sqrt{n}} \leq t_i n^{-\eta_2}, \text{ and } \rho_{i+1}^2 \frac{J_\infty^2(K, \mathcal{G}'_i)}{n} \leq t_i n^{-\eta_3}.$$

By taking  $t = \left(\frac{1}{K}\right)^{d/m_i} \leq n^{\eta_4}$ , where  $\eta_4$  is another constant, we have with probability at least  $1 - \exp(-n^{\eta_4})$ ,

$$\sup_{g \in \mathcal{G}'_i} \left| \|g\|_n^2 - \|g\|_2^2 \right| \leq t_i n^{-\min(\eta_2, \eta_3)}.$$

Similar results can be applied to  $f - \hat{f}_m^*$  and  $f - f_1^*$ , where  $f_1^*$  is defined as in (C.12). Let

$p_{1,i}$  and  $p_{2,i}$  be the corresponding probability. Let

$$\begin{aligned} p_0 &= 1 - \sum_{i=1}^p \left( C_{11} \exp(-C_{12}n^{\eta_1}) - C_{13} \exp(-C_{14}n^{\eta_0}) - \exp(-n^{\eta_4}) - (1 - p_{1,i}) - (1 - p_{2,i}) \right) \\ &\leq 1 - C_{17} \exp(-C_{18}n^{\eta_5}), \end{aligned} \quad (\text{C.10})$$

where  $\eta_5$  is a constant.

Using the extension theorem, with probability at least  $p_0$ ,

$$\begin{aligned} &\|f - f_1\|_n^2 + \lambda_1 \|f_1\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 - (\|f - f_m^*\|_n^2 + \lambda_2 \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2) \\ &\geq \|f - f_1\|_2^2 + \lambda_1 \|f_1\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 - (\|f - f_m^*\|_2^2 + \lambda_2 \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2) - t_i n^{-\eta_6} \\ &\geq \|f - f_1^*\|_2^2 + C_3 \lambda_1 \|f_1^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 - (\|f - f_m^*\|_2^2 + C_4 \lambda_2 \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2) - t_i n^{-\eta_6}, \end{aligned} \quad (\text{C.11})$$

where  $\lambda_1 = C_1 \lambda_2$ , and  $f_1^*$  is defined as

$$f_1^* = \operatorname{argmin}_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)} \|f - \hat{f}\|_2^2 + C_3 \lambda_1 \|\hat{f}\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2. \quad (\text{C.12})$$

By Fourier transform, we have

$$\begin{aligned} \|f - f_m^*\|_2^2 + C_4 \lambda_2 \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 &= \int |\mathcal{F}(f)(\omega) - \mathcal{F}(f_m^*)(\omega)|^2 d\omega + C_4 \lambda_2 \int |\mathcal{F}(f_m^*)(\omega)|^2 (1 + |\omega|^2)^m d\omega \\ &= \int |\mathcal{F}(f)(\omega) - \mathcal{F}(f_m^*)(\omega)|^2 + C_4 \lambda_2 |\mathcal{F}(f_m^*)(\omega)|^2 (1 + |\omega|^2)^m d\omega \\ &= \int \frac{C_4 \lambda_2 (1 + |\omega|^2)^m}{1 + C_4 \lambda_2 (1 + |\omega|^2)^m} |\mathcal{F}(f)(\omega)|^2 d\omega, \end{aligned} \quad (\text{C.13})$$

and

$$\|f - f_1^*\|_2^2 + C_3 \lambda_1 \|f_1^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 = \int \frac{C_3 \lambda_1 (1 + |\omega|^2)^m}{1 + C_3 \lambda_1 (1 + |\omega|^2)^m} |\mathcal{F}(f)(\omega)|^2 d\omega. \quad (\text{C.14})$$

By choosing  $C_1 = 2C_4/C_3$ , and combining (C.11), (C.13) and (C.14), we have

$$\begin{aligned}
& \|f - f_1^*\|_2^2 + C_3\lambda_1\|f_1^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 - (\|f - f_m^*\|_2^2 + C_4\lambda_2\|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2) \\
&= \int \frac{C_3\lambda_1(1+|\omega|^2)^m}{1+C_3\lambda_1(1+|\omega|^2)^m} |\mathcal{F}(f)(\omega)|^2 d\omega - \int \frac{C_4\lambda_2(1+|\omega|^2)^m}{1+C_4\lambda_2(1+|\omega|^2)^m} |\mathcal{F}(f)(\omega)|^2 d\omega \\
&= \int \left( \frac{C_1C_3\lambda_2(1+|\omega|^2)^m}{1+C_1C_3\lambda_2(1+|\omega|^2)^m} - \frac{C_4\lambda_2(1+|\omega|^2)^m}{1+C_4\lambda_2(1+|\omega|^2)^m} \right) |\mathcal{F}(f)(\omega)|^2 d\omega \\
&\geq \int_{C_4\lambda_2(1+|\omega|^2)^m < 1} \left( \frac{C_1C_3\lambda_2(1+|\omega|^2)^m}{1+C_1C_3\lambda_2(1+|\omega|^2)^m} - \frac{C_4\lambda_2(1+|\omega|^2)^m}{1+C_4\lambda_2(1+|\omega|^2)^m} \right) |\mathcal{F}(f)(\omega)|^2 d\omega \\
&\geq \frac{1}{6} \int_{C_4\lambda_2(1+|\omega|^2)^m < 1} (1+|\omega|^2)^m |\mathcal{F}(f)(\omega)|^2 d\omega \\
&\geq \frac{C_4\lambda_2}{6} \int_{C_4\lambda_2(1+|\omega|^2)^m < 1} (1+|\omega|^2)^m |\mathcal{F}(f)(\omega)|^2 d\omega.
\end{aligned}$$

By similar approach as in the proof of Lemma C.5.4, which is presented in the supplementary materials, the results hold.

### C.3 Proof of Lemma C.1.2

We need the following lemma, which states that the absolute value of the empirical inner product  $|\langle e, f - \hat{f}_m \rangle_n|$  is small when  $m \in [m_{\min}, m_0 - \epsilon_n]$ , where  $\hat{f}_m$  is defined in (3.2).

The proof can be found in Appendix C.6.

**Lemma C.3.1.** *Suppose  $m \in [m_{\min}, m_0 - \epsilon_n]$ . Let  $t_m = C_0 n^{-2m/(2m+d)}/\log n$  and  $\mu_m = C n^{\frac{d}{2m+d}}$  for any fixed constants  $C$  and  $C_0$ . Let  $\mathcal{C} = \{\hat{f}_m : \forall m \in [m_{\min}, m_0 - \epsilon_n], \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq t_m\}$ . Under class  $\mathcal{C}$ , with probability at least  $1 - C_1 \exp(-C_2 n^{\eta_0})$ ,  $2|\langle e, f - \hat{f}_m \rangle_n| \leq n^{-\eta_1/2} t_m$ , where  $C_1, C_2, \eta_0$ , and  $\eta_1$  are positive constants depending on  $f$  and  $\Omega$ . In particular,  $|\langle e, f - \hat{f}_m \rangle_n| = o_P(t_m)$ .*

*Proof of Lemma C.1.2:*

Since  $m \in [m_{\min}, m_0 - \epsilon_n]$ , we have  $f \in H^m$ . Under class  $\mathcal{C}$ , since  $t_m = C_0 n^{-2m/(2m+d)}/\log n$ , we have  $\|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq C_1/\log n$  for any  $m \in [m_{\min}, m_0 - \epsilon_n]$  and  $C_1$  is a constant.

The rest of the proof is similar as the proof of Lemma C.1.1, where the differences are:

(i) Calculating  $J_\infty^2(K, \mathcal{G}'_i)$  in Lemma C.2.1; and (ii) Applying Lemma C.3.1. We will not present the procedure for the conciseness of this article.

#### C.4 Proof of Lemma C.1.3

We first present the following lemma, whose proof is provided in Appendix ?.

**Lemma C.4.1.** *Define  $f_{m_0}^*$  as*

$$f_{m_0}^* = \operatorname{argmin}_{\hat{f} \in \mathcal{N}_{\Psi_{m_0}}(\Omega)} \|\hat{f} - f\|_2^2 + \frac{C_1 \mu_{m_0}}{n} \|\hat{f}\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^2. \quad (\text{C.15})$$

If  $f \in H^{m_0}$ ,

$$\|f - f_{m_0}^*\|_2^2 + \frac{C_1 \mu_{m_0}}{n} \|f_{m_0}^*\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^2 \leq \begin{cases} C_2 n^{\frac{-2m_0}{2m_0+d}} & \text{if } f \in H^{m_0}, \\ C_2 n^{\frac{-2m_0}{2m_0+d}} h_2(n) & \text{if } f \notin H^{m_0}, \end{cases}$$

where  $h_2(n)$  is defined in (3.20).

Now we are able to prove Lemma C.1.3.

*Proof of Lemma C.1.3:*

Let  $\mathcal{G} = \{g : g = (f - \hat{f}_{m_0}) / \|\hat{f}_{m_0}\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}\}$ , and  $\mathcal{F} = \{g : g = \hat{f}_{m_0} / \|\hat{f}_{m_0}\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}\}$ .

Since  $\mathcal{F} \subset H^{m_0}$ , the bracket entropy of  $\mathcal{F}$  can be bounded by

$$H_B(\delta_n/V(\Omega), \mathcal{F}, \|\cdot\|_\infty) \leq C_1 \left(\frac{1}{\delta_n}\right)^{d/m_0}.$$

Applying Lemma C.5.1, we have

$$P\left(\sup_{g \in \mathcal{G}} \frac{\left|\frac{1}{n} \sum_{i=1}^n e_i g(x_i)\right|}{\|g\|_n^{1-\frac{d}{2m_0}} \|\hat{f}_{m_0}\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^{\frac{d}{2m_0}}} \geq T n^{-1/2}\right) \leq C_4 \exp(-C_5 T^2), \quad (\text{C.16})$$

where  $T^2 = n^{\frac{d}{4(2m_0+d)}}$ . Similar result holds for  $f_{m_0}^*$ .

Since  $\hat{f}_{m_0}$  is the minimizer of (3.2), we have

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n (f(x_i) - \hat{f}_{m_0}(x_i))^2 + \frac{\mu_{m_0}}{n} \|\hat{f}_{m_0}\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^2 \\
& \leq \frac{1}{n} \sum_{i=1}^n (f(x_i) - f_{m_0}^*(x_i))^2 + \frac{2}{n} \sum_{i=1}^n e_i (f(x_i) - f_{m_0}^*(x_i)) + \frac{\mu_{m_0}}{n} \|f_{m_0}^*\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^2 \\
& \quad - \frac{2}{n} \sum_{i=1}^n e_i (f(x_i) - \hat{f}_{m_0}(x_i)) \\
& \leq \frac{1}{n} \sum_{i=1}^n (f(x_i) - f_{m_0}^*(x_i))^2 + \frac{\mu_{m_0}}{n} \|f_{m_0}^*\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^2 + 2Tn^{-1/2} \|f - f_{m_0}^*\|_n^{1-\frac{d}{2m_0}} \|f_{m_0}^*\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^{\frac{d}{2m_0}} \\
& \quad + 2Tn^{-1/2} \|f - \hat{f}_{m_0}\|_n^{1-\frac{d}{2m_0}} \|\hat{f}_{m_0}\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^{\frac{d}{2m_0}}.
\end{aligned}$$

Similar to the proof of Theorem C.2.2, we have

$$2Tn^{-1/2} \|f - \hat{f}_{m_0}\|_n^{1-\frac{d}{2m_0}} \|\hat{f}_{m_0}\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^{\frac{d}{2m_0}} + 2Tn^{-1/2} \|f - f_{m_0}^*\|_n^{1-\frac{d}{2m_0}} \|f_{m_0}^*\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^{\frac{d}{2m_0}} \leq tn^{-\eta_1},$$

where  $\eta_1 > 0$  is a constant.

Notice that

$$E \left( \frac{1}{n} \sum_{i=1}^n (f(x_i) - f_{m_0}^*(x_i))^2 - \|f - f_{m_0}^*\|_2^2 / V(\Omega) \right)^2 \leq 4 \|f - f_{m_0}^*\|_\infty^2 \|f - f_{m_0}^*\|_2^2 / (V(\Omega)),$$

and

$$\|f - f_{m_0}^*\|_\infty \leq \|f\|_\infty + \|f_{m_0}^*\|_\infty \leq \|f\|_\infty + C_4 n^{\frac{2m_0}{2m_0+d}} t,$$

where the second inequality is because of the interpolation inequality. By Bernstein's in-

equality, we have

$$\begin{aligned}
& P\left(\frac{1}{n}\sum_{i=1}^n(f(x_i) - f_{m_0}^*(x_i))^2 - \|f - f_{m_0}^*\|_2^2/V(\Omega) \geq \|f - f_{m_0}^*\|_2^2\right) \\
& \leq \exp\left(-\frac{n\|f - f_{m_0}^*\|_2^4/2}{4\|f - f_{m_0}^*\|_\infty^2\|f - f_{m_0}^*\|_2^2/(V(\Omega)) + 3\|f - f_{m_0}^*\|_\infty^2\|f - f_{m_0}^*\|_2^2}\right) \\
& \leq \exp\left(-C_9\frac{nt}{tn^{\frac{2m_0}{2m_0+d}}}\right) = \exp(-C_9n^{\frac{d}{2m_0+d}}).
\end{aligned}$$

Therefore, by combining the probability in (C.16), with probability at least  $1 - C_{10} \exp(-C_{11}n^{\frac{d}{4(2m_0+d)}})$ , we have

$$\begin{aligned}
\frac{\mu_{m_0}}{n}\|\hat{f}_{m_0}\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^2 & \leq \frac{1}{n}\sum_{i=1}^n(f(x_i) - f_{m_0}^*(x_i))^2 + \frac{\mu_{m_0}}{n}\|f_{m_0}^*\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^2 + tn^{-\eta_1} \quad (\text{C.17}) \\
& \leq C_{12}(\|f - f_{m_0}^*\|_2^2 + \frac{\mu_{m_0}}{n}\|\hat{f}_{m_0}^*\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^2).
\end{aligned}$$

Combining (C.17) with Lemma C.4.1, we finish the proof.

## C.5 Proof of Lemma C.2.2

Before the proof, we present the lemmas used in the proof of Lemma C.2.2. Lemma C.5.1 is Lemma 8.4 in [92], which states the property of the absolute value of inner product  $|\langle e, g \rangle|$ . Lemma C.5.2 states the uniform bound of the ratio between the empirical norm and the  $L_2$  norm. Lemma C.5.3 states the  $L_2$  norm of  $\hat{f}_m$  (defined in (3.2)) is bounded. Lemma C.5.4 is needed for the  $L_2$  condition of Lemma C.5.2. The proofs of Lemmas C.5.2, C.5.3 and C.5.4 can be found in Appendix C.10, C.8, and C.11, respectively.

**Lemma C.5.1.** *Suppose that for class  $\mathcal{G}$  there exists a constant  $A$  and  $\alpha \in (0, 2)$  such that*

$$H(\delta, \mathcal{G}, \|\cdot\|_\infty) \leq A\delta^{-\alpha},$$

for all  $\delta > 0$ , and  $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq R$  for some constant  $R$ . Furthermore, suppose  $e_i$ 's

satisfy Assumption 3.3.1. Then for some constant  $c$  depending on  $A, \alpha, R, K$  and  $\sigma_0$ , for all  $T > c$ ,

$$P\left(\sup_{g \in \mathcal{G}} \frac{\left|\frac{1}{\sqrt{n}} \sum_{i=1}^n e_i g(x_i)\right|}{\|g\|_n^{1-\frac{\alpha}{2}}} \geq T\right) \leq c \exp(-T^2/c^2).$$

**Lemma C.5.2.** Assume for class  $\mathcal{G}$ ,  $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq K < 1$ ,  $H_B(\delta_n/V(\Omega), \mathcal{G}, \|\cdot\|_\infty) \leq \frac{n\delta_n^2}{1200V(\Omega)K^2}$ , and  $n\delta_n^2 \rightarrow \infty$ , where  $V(\Omega)$  denotes the volume of  $\Omega$  and  $0 < \delta_n < 1$ . Then we have

$$P\left(\inf_{\|g\|_2 \geq 2\delta_n, g \in \mathcal{G}} \frac{\|g\|_n^2}{\|g\|_2^2} < \eta_1\right) \leq C_1 \exp(-C_2 n \delta_n^2 / K^2),$$

and

$$P\left(\sup_{\|g\|_2 \geq 2\delta_n, g \in \mathcal{G}} \frac{\|g\|_n^2}{\|g\|_2^2} > \eta_2\right) \leq C_3 \exp(-C_4 n \delta_n^2 / K^2),$$

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

**Lemma C.5.3.** Suppose  $m \in [m_0 + \epsilon_n, m_{\max}]$ . Let  $\mu_m = Cn^{\frac{d}{2m+d}}$ ,  $\mu_{m_0} = Cn^{\frac{d}{2m_0+d}}$ ,  $t_m = C_0 n^{-2m_0/(2m+d)} / h_1(n)$  if  $f \in H^{m_0}$ , and  $t_m = C_0 n^{-2m_0/(2m+d)} / (\log n)^2$  if  $f \notin H^{m_0}$ , where  $C$  and  $C_0$  are any fixed constants, and  $h_1(n)$  is defined in (3.18). Let  $\mathcal{C}$  denote the class  $\{\hat{f}_m : \forall m \in [m_0 + \epsilon_n, m_{\max}], \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq t_m\}$ . Under class  $\mathcal{C}$ , with probability at least  $1 - C_1 \exp(-C_2 \mu_{m_0})$ ,  $\|\hat{f}_m\|_2^2 \leq C_3$ , where  $C_1, C_2$ , and  $C_3$  are constants related to  $f, \Omega, C$  and  $m_{\max}$ .

**Lemma C.5.4.** Fix  $\mu_m = Cn^{\frac{d}{2m+d}}$  with any constant  $C$ . Define  $f_m^*$  as

$$f_m^* = \operatorname{argmin}_{\hat{f} \in \mathcal{N}_{\Psi_m}(\mathbb{R}^d)} \|f - \hat{f}\|_2^2 + \frac{C_1 \mu_m}{n} \|\hat{f}\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2. \quad (\text{C.18})$$

If the extended function  $f \in H^{m_0}(\mathbb{R}^d)$ , we have

$$\|f - f_m^*\|_2^2 + \frac{C_1 \mu_m}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 \geq C_2 n^{\frac{-2m_0}{2m+d}} / h_1(n),$$

where  $h_1(\cdot)$  is defined in (3.18).

If the extended function  $f \notin H^{m_0}(\mathbb{R}^d)$ , we have

$$\|f - f_m^*\|_2^2 + \frac{C_1 \mu_m}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 \geq C_2 n^{\frac{-2m_0}{2m+d}} / (\log n)^2.$$

Now we are ready to prove Lemma C.2.2.

*Proof of Lemma C.2.2:*

We use the following notations. Let  $m_0 + \epsilon_n = m_1 < m_2 < \dots < m_p = m_{\max}$  be a partition of  $[m_0 + \epsilon_n, m_{\max}]$ ,  $\mathcal{G}_i = \{g : g = f - \hat{f}_m, \forall m \in [m_i, m_{i+1}], \hat{f}_m \in \mathcal{C}\}$  and  $\mathcal{G} = \bigcup \mathcal{G}_i$ . Let class  $\mathcal{F}_i = \{g : g = \hat{f}_m, \forall m \in [m_i, m_{i+1}], \hat{f}_m \in \mathcal{C}\}$ . Therefore, the bracket entropy of class  $\mathcal{G}_i$  is the same as class  $\mathcal{F}_i$ , and  $\mathcal{F}_i \subset H^{m_i}$ . Let  $t_i = t_{m_i}$  and  $\mu_i = \mu_{m_i}$ . Define  $\rho_m = (nt_m/\mu_m)^{1/2}$  and  $\rho_i = (nt_i/\mu_i)^{1/2}$ . Since  $\hat{f}_m \in \mathcal{C}$ , we have  $\|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq C\rho_m^2$  for all  $m$ . Let class  $\mathcal{F}'_i = \{g : g = \hat{f}_m/\rho_m, \forall m \in [m_i, m_{i+1}], \hat{f}_m \in \mathcal{C}\}$ , and  $\mathcal{G}'_i = \{g : g = (f - \hat{f}_m)/\rho_m, \forall m \in [m_i, m_{i+1}], \hat{f}_m \in \mathcal{C}\}$ .

The proof consists of four steps. In Step 1 we apply Lemma C.5.1, which will be presented later, to class  $\mathcal{G}_i$  to link the empirical inner product  $\langle e, g \rangle_n$  with the empirical norm  $\|g\|_n$  for  $g \in \mathcal{G}_i$ . Step 2 is to apply Lemma C.5.2 to  $\mathcal{G}'_i$  (which is a scaled class of  $\mathcal{G}_i$ ) to obtain the relation between  $\|g\|_n$  and  $\|g\|_2$  for  $g \in \mathcal{G}_i$ . In Step 3, the  $L_2$  norm of  $\|g\|_2$  is bounded. In Step 4, by combining results obtained from Steps 1, 2, and 3, we complete the proof.

**Step 1:**

It can be checked that normal random noise satisfy (3.22). By interpolation inequality,

$$\|\hat{f}_m\|_{H^{m_i}} \leq C_1 \|\hat{f}_m\|_2^{1-\frac{m_i}{m}} \|\hat{f}_m\|_{H^m}^{\frac{m_i}{m}},$$

where  $C_1$  is a constant.

Since  $\mathcal{F}'_i \subset H^{m_i}$ , the bracket entropy numbr can be controlled by [94]

$$H_B(\delta_n/V(\Omega), \mathcal{F}'_i, \|\cdot\|_\infty) \leq C_2 \left(\frac{1}{\delta_n}\right)^{d/m_i}.$$

Therefore, by Lemma C.5.1, we have

$$P\left(\sup_{g \in H^m(\rho_m)} \frac{|\frac{1}{n} \sum_{i=1}^n e_i g(x_i)|}{\|g\|_n^{1-\frac{d}{2m}} \rho_m^{\frac{d}{2m}}} \geq T n^{-1/2}\right) \leq C_3 \exp(-C_4 T^2), \quad (\text{C.19})$$

where  $H^m(\rho_m)$  denotes the Sobolev space with radius  $\rho_m$ , and  $T$  is a constant given in Lemma C.5.1.

By Lemma C.5.3,  $\|g\|_2 \leq C_5$  for all  $g \in \mathcal{G}$  with some constant  $C_5$ . Therefore, we can normalize  $g$  by  $g/C_5$ , thus  $\|g\|_2 \leq 1$ .

Consider class  $\mathcal{G}_i$ . By interpolation inequality, it can be shown that  $\rho_m \leq \rho_{i+1}$ , which indicates

$$\frac{|\frac{1}{n} \sum_{i=1}^n e_i g(x_i)|}{\|g\|_n^{1-\frac{d}{2m}} \rho_m^{\frac{d}{2m}}} \geq \frac{|\frac{1}{n} \sum_{i=1}^n e_i g(x_i)|}{\|g\|_n^{1-\frac{d}{2m_i}} \rho_{i+1}^{\frac{d}{2m_i}}}.$$

Therefore, by (C.19), we have

$$P\left(\sup_{g \in \mathcal{G}_i} \frac{|\frac{1}{n} \sum_{i=1}^n e_i g(x_i)|}{\|g\|_n^{1-\frac{d}{2m_i}} \rho_{i+1}^{\frac{d}{2m_i}}} \geq T n^{-1/2}\right) \leq C_3 \exp(-C_4 T^2). \quad (\text{C.20})$$

## Step 2:

In order to apply Lemma C.5.2 to  $\mathcal{G}'_i$ , we need to check the conditions of Lemma C.5.2

hold for  $\mathcal{G}'_i$ . The first condition is the  $L_2$  norm of  $g$  is bounded away from zero for any  $g \in \mathcal{G}'_i$ , which we refer as  $L_2$  condition, and the second condition is the entropy condition. By (C.18) and extension,  $(\mu_m/n)\|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 \leq (C_1\mu_m/n)\|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq C_1t_m$ , for some constant  $C_1$ , we have

$$\begin{aligned} \|f - f_m^*\|_2^2 + \frac{C_5\mu_m}{n}\|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 &\leq \|f - \hat{f}_m\|_2^2 + \frac{C_5\mu_m}{n}\|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 \\ &\leq \|f - \hat{f}_m\|_2^2 + C_6t_m. \end{aligned}$$

Therefore, by Lemma C.5.4,

$$\begin{aligned} \|f - \hat{f}_m\|_2^2 &\geq \|f - f_m^*\|_2^2 + \frac{C_5\mu_m}{n}\|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 - C_6t_m \\ &\geq (C_7/C_0 - C_6)t_m. \end{aligned} \tag{C.21}$$

It need to be noticed that  $t_m$  takes different value for  $f \in H^{m_0}(\mathbb{R}^d)$  and  $f \notin H^{m_0}(\mathbb{R}^d)$ . Note the  $L_2$  norm is taken in  $\mathbb{R}^d$ . We can choose  $C_5$  and  $C_0$  accordingly such that  $C_7/C_0 - C_6 > 0$ . By restriction, it can be shown that  $\|f - \hat{f}_m\|_2^2 \geq C_8t_m$ , where the  $L_2$  norm is taken in  $\Omega$  with  $C_8 > 0$ .

By (C.21), we have

$$\|g\|_2^2 \geq \min_{m \in [m_i, m_{i+1}]} C_8 \frac{t_m}{\rho_m^2} = C_8 \min_{m \in [m_i, m_{i+1}]} \frac{\mu_m}{n} = C_8 \frac{\mu_{i+1}}{n} \tag{C.22}$$

for  $g \in \mathcal{G}'_i$ , which indicates  $L_2$  condition holds for  $\mathcal{G}'_i$ .

Next, we turn to the entropy condition. By the interpolation inequality and the boundedness of  $\|\hat{f}_m\|_2$ , we have

$$\|\hat{f}_m\|_\infty \leq C_9 \|\hat{f}_m\|_2^{1-\frac{d}{2m}} \|\hat{f}_m\|_{H^m}^{\frac{d}{2m}} \leq C_{10} \|\hat{f}_m\|_{H^m}^{\frac{d}{2m}} \leq C_{10} \rho_m^{\frac{d}{2m}}$$

for some constant  $C_{10}$ . Therefore,  $\|g\|_\infty$  can be bounded by

$$\|g\|_\infty \leq C_{10} \max_{m \in [m_i, m_{i+1}]} \rho_m^{\frac{d-2m}{2m}} \leq C_{10} \rho_i^{\frac{d-2m_i}{2m_i}}, \quad (\text{C.23})$$

for  $g \in \mathcal{G}_i$  since  $\rho_m \rightarrow \infty$  as  $n \rightarrow \infty$ , and  $\rho_m^{\frac{d-2m}{2m}}$  decreases when  $m$  increases.

Since  $\mathcal{G}'_i \subset H^{m_i}$ , the bracket entropy can be bounded by

$$H_B(\delta_n/V(\Omega), \mathcal{G}'_i, \|\cdot\|_\infty) \leq C_{11} \left( \frac{1}{\delta_n} \right)^{d/m_i}.$$

The entropy condition  $H_B(\delta_n/V(\Omega), \mathcal{G}', \|\cdot\|_\infty) \leq \frac{n\delta_n^2}{1200V(\Omega)K^2}$  is satisfied if for any fixed constant  $C_{12}$ ,

$$C_{12} \left( \frac{1}{\delta_n} \right)^{d/m_i} \leq n\delta_n^2/K^2, \quad (\text{C.24})$$

since  $n \rightarrow \infty$ , where  $K = \sup_{g \in \mathcal{G}'} \|g\|_\infty$ . By (C.23),

$$K \leq C_{10} \rho_i^{\frac{d-2m_i}{2m_i}} = C_{10} \left( \frac{nt_i}{\mu_i} \right)^{\frac{d-2m_i}{4m_i}},$$

By (C.22), we have  $\delta^2 \geq C_8 \mu_{i+1}/n$ . Therefore, (C.24) is true if for any fixed constant  $C_{13}$ ,

$$n \left( \frac{\mu_{i+1}}{n} \right)^{\frac{2m_i+d}{2m_i}} \left( \frac{\mu_i}{n} \right)^{\frac{d-2m_i}{2m_i}} \geq C_{13} t_i^{\frac{d-2m_i}{2m_i}}. \quad (\text{C.25})$$

We only present the case  $f \in H^{m_0}$ . The case  $f \notin H^{m_0}$  is similar. By plugging  $\mu_i$  and  $\mu_{i+1}$

in (C.25), we obtain

$$\begin{aligned}
& n^{1 - \frac{m_{i+1}}{m_i} \frac{2m_i+d}{2m_{i+1}+d} + \frac{2m_i-d}{2m_i} - \frac{m_0}{m_i} - \frac{2m_i-d}{2m_i}} h_1(n)^{\frac{d-2m_i}{2m_i}} \geq C_{13} \\
& \Leftrightarrow n^{-\frac{(m_{i+1}-m_i)(2m_i+d)}{(2m_{i+1}+d)m_i} + \frac{(2m_i-d)(m_i-m_0)}{(2m_i+d)m_i}} h_1(n)^{\frac{d-2m_i}{2m_i}} \geq C_{13} \\
& \Leftrightarrow \left( -\frac{(m_{i+1}-m_i)(2m_i+d)}{(2m_{i+1}+d)m_i} + \frac{(2m_i-d)(m_i-m_0)}{(2m_i+d)m_i} \right) (\log n) \\
& \quad + \left( \frac{d-2m_i}{2m_i} \right) \log h_1(n) \geq \log C_{13}
\end{aligned}$$

for any fixed constant  $C_{13}$ . By the conditions of Theorem C.1.1, it is true that for some large  $n$ ,  $\epsilon_n \geq C_{14} \log h_1(n) / \log n$  for constant  $C_{14}$  related to  $d$  and  $C_{13}$ . Therefore, by picking  $m_{i+1} - m_i$  such that

$$\frac{(m_{i+1}-m_i)(2m_i+d)}{(2m_{i+1}+d)m_i} \leq \frac{(2m_i-d)(m_i-m_0)}{2m_i(2m_i+d)},$$

the condition of Lemma C.5.2 holds. By Lemma C.5.2, with probability at least  $1 - C_{15} \exp(-C_{16}n^{\eta_0})$ , where  $\eta_0 > 0$  is a constant,  $\|g\|_n^2 \geq \eta_1 \|g\|_2^2$  for some constant  $\eta_1$ , and  $\|g\|_n^2 \leq \eta_2 \|g\|_2^2$  for some constant  $\eta_2$ , for any  $g \in \mathcal{G}'_i$ . Thus, Lemma C.5.2 also can be applied to  $\mathcal{G}_i$ .

### Step 3:

In this step, we give an upper bound of  $\|f - f_m^*\|_2^2 + (C_5 \mu_m / n) \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2$ , where  $C_5$  is a constant. Let  $C_n = (C_5 \mu_m) / n$ .

By Fourier transform, we have

$$\begin{aligned}
& \|f - f_m^*\|_2^2 + C_n \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 \\
&= \int_{\mathbb{R}^d} |\mathcal{F}(f)(\omega) - \mathcal{F}(f_m^*)(\omega)|^2 d\omega + C_n \int_{\mathbb{R}^d} |\mathcal{F}(f_m^*)(\omega)|^2 (1 + |\omega|^2)^m d\omega \\
&= \int_{\mathbb{R}^d} |\mathcal{F}(f)(\omega) - \mathcal{F}(f_m^*)(\omega)|^2 + C_n |\mathcal{F}(f_m^*)(\omega)|^2 (1 + |\omega|^2)^m d\omega \\
&= \int_{\mathbb{R}^d} \frac{C_n(1 + |\omega|^2)^m}{1 + C_n(1 + |\omega|^2)^m} |\mathcal{F}(f)(\omega)|^2 d\omega \\
&= \int_{\Omega_1} \frac{C_n(1 + |\omega|^2)^m}{1 + C_n(1 + |\omega|^2)^m} |\mathcal{F}(f)(\omega)|^2 d\omega + \int_{\Omega_1^c} \frac{C_n(1 + |\omega|^2)^m}{1 + C_n(1 + |\omega|^2)^m} |\mathcal{F}(f)(\omega)|^2 d\omega,
\end{aligned}$$

where  $\Omega_1 = \{\omega : C_n(1 + |\omega|^2)^m < 1\}$ , and the third equality is because of (C.18). Therefore,

$$\begin{aligned}
& \|f - f_m^*\|_2^2 + C_n \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 \tag{C.26} \\
&= \int_{\Omega_1} \frac{C_n(1 + |\omega|^2)^m}{1 + C_n(1 + |\omega|^2)^m} |\mathcal{F}(f)(\omega)|^2 d\omega + \int_{\Omega_1^c} \frac{C_n(1 + |\omega|^2)^m}{1 + C_n(1 + |\omega|^2)^m} |\mathcal{F}(f)(\omega)|^2 d\omega \\
&\leq \int_{\Omega_1} C_n(1 + |\omega|^2)^m |\mathcal{F}(f)(\omega)|^2 d\omega + \int_{\Omega_1^c} |\mathcal{F}(f)(\omega)|^2 d\omega.
\end{aligned}$$

If  $f \in H^{m_0}(\Omega)$  (therefore the extended function is within  $H^{m_0}(\mathbb{R}^d)$ ), we have

$$\begin{aligned}
& \int_{\Omega_1} C_n(1 + |\omega|^2)^m |\mathcal{F}(f)(\omega)|^2 d\omega + \int_{\Omega_1^c} |\mathcal{F}(f)(\omega)|^2 d\omega \tag{C.27} \\
&\leq C_n^{\frac{m_0}{m}} \int_{\Omega_1} (1 + |\omega|^2)^{m_0} |\mathcal{F}(f)(\omega)|^2 d\omega + C_n^{\frac{m_0}{m}} \int_{\Omega_1^c} (1 + |\omega|^2)^{m_0} |\mathcal{F}(f)(\omega)|^2 d\omega \\
&\leq C_n^{\frac{m_0}{m}} \|f\|_{\mathcal{N}_{\Psi_{m_0}}(\Omega)}^2 = C_6 n^{-\frac{2m_0}{2m+d}}.
\end{aligned}$$

where the third inequality is because of  $C_n(1 + |\omega|^2)^m < 1$  for  $\omega \in \Omega_1$  and  $C_n(1 + |\omega|^2)^m > 1$  for  $\omega \in \Omega_1^c$ .

If  $f \notin H^{m_0}(\Omega)$ , by Lemma 3.3.1, there exists an increasing  $h_1(|\omega|)$  such that

$$\begin{aligned} \int \frac{|\mathcal{F}(f)(\omega)|^2}{h_1(|\omega|)} (1 + |\omega|^2)^{m_0} d\omega &< \infty, \\ \int \frac{|\mathcal{F}(f)(\omega)|^2}{h_1(|\omega|)} (1 + |\omega|^2)^{m_0 + \tau_1} d\omega &= \infty, \end{aligned}$$

and

$$\lim_{n \rightarrow \infty} \frac{\log h_1(n)}{\log n} = 0,$$

for any  $\tau_1 > 0$ . By the proof of Lemma 3.3.1, there exists a constant  $C'$  such that  $h_1(|\omega|) \leq C'(1 + |\omega|^2)^{m_{\min}/10}$ . By setting  $\Omega_2 = \{\omega : C_n(1 + |\omega|^2)^m < h_1(|\omega|)^{m/m_0}\}$ , we have

$$\begin{aligned} &\|f - f_m^*\|_2^2 + C_n \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 \tag{C.28} \\ &= \int_{\mathbb{R}^d} \frac{C_n(1 + |\omega|^2)^m}{1 + C_n(1 + |\omega|^2)^m} |\mathcal{F}(f)(\omega)|^2 d\omega \\ &\leq \int_{\Omega_2} C_n(1 + |\omega|^2)^m |\mathcal{F}(f)(\omega)|^2 d\omega + \int_{\Omega_2^c} |\mathcal{F}(f)(\omega)|^2 d\omega \\ &\leq C' C_n^{\frac{m_0}{m}} h_1(n) \int_{\Omega_2} (1 + |\omega|^2)^{m_0} \frac{|\mathcal{F}(f)(\omega)|^2}{h_1(|\omega|)} d\omega + C_n^{\frac{m_0}{m}} \int_{\Omega_2^c} (1 + |\omega|^2)^{m_0} \frac{|\mathcal{F}(f)(\omega)|^2}{h_1(|\omega|)} d\omega \\ &\leq C_6 n^{-\frac{2m_0}{2m+d}} h_1(n), \end{aligned}$$

where the second inequality is because when  $\omega \in \Omega_2$ ,

$$\frac{C_n(1 + |\omega|^2)^m}{h_1(|\omega|)^{m/m_0}} < \left( \frac{C_n(1 + |\omega|^2)^m}{h_1(|\omega|)^{m/m_0}} \right)^{m_0/m}$$

and  $|\omega| < n$ , and

$$\left( \frac{C_n(1 + |\omega|^2)^m}{h_1(|\omega|)^{m/m_0}} \right)^{m_0/m} > 1$$

when  $\omega \notin \Omega_2$ .

Combining (C.26), (C.27) and (C.28), we have

$$\|f - f_m^*\|_2^2 + (C_5\mu_m/n)\|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 \leq \begin{cases} C_6n^{-\frac{2m_0}{2m+d}} & \text{if } f \in H^{m_0}, \\ C_6n^{-\frac{2m_0}{2m+d}}h_1(n) & \text{if } f \notin H^{m_0}. \end{cases} \quad (\text{C.29})$$

**Step 4:**

Note that inequality (C.20) can be also applied to  $g_1 = f - f_m^*$ , where  $f_m^*$  is defined in (C.18). By (3.2) and  $y_i = f(x_i) + e_i$ , we have

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n (f(x_i) - \hat{f}_m(x_i))^2 + \frac{2}{n} \sum_{i=1}^n \epsilon_i (f(x_i) - \hat{f}_m(x_i)) + \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \\ & \leq \frac{1}{n} \sum_{i=1}^n (f(x_i) - f_m^*(x_i))^2 + \frac{2}{n} \sum_{i=1}^n \epsilon_i (f(x_i) - f_m^*(x_i)) + \frac{\mu_m}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2. \end{aligned}$$

By (C.20), with probability at least  $1 - C_3 \exp(-C_4 T^2)$ ,

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n (f(x_i) - \hat{f}_m(x_i))^2 + \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \quad (\text{C.30}) \\ & \leq \frac{1}{n} \sum_{i=1}^n (f(x_i) - f_m^*(x_i))^2 + \frac{2}{n} \sum_{i=1}^n \epsilon_i (\hat{f}_m(x_i) - f_m^*(x_i)) + \frac{\mu_m}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \\ & \leq \frac{1}{n} \sum_{i=1}^n (f(x_i) - f_m^*(x_i))^2 + \frac{\mu_m}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \\ & \quad + 2Tn^{-1/2} \|f - f_m^*\|_n^{1-\frac{d}{2m_i}} \rho_{i+1}^{\frac{d}{2m_i}} + 2Tn^{-1/2} \|f - \hat{f}_m\|_n^{1-\frac{d}{2m_i}} \rho_{i+1}^{\frac{d}{2m_i}}. \end{aligned}$$

In order to show  $2\langle e, f - \hat{f}_m \rangle$  is small, by (C.20), we need to show that  $n^{-1/2} \|f - \hat{f}_m\|_n^{1-\frac{d}{2m_i}} \rho_{i+1}^{\frac{d}{2m_i}}$  is small compared to  $t_i$ . We split it into three cases.

**Case 1:**  $\|f - \hat{f}_m\|_n^2 \leq \|f - f_m^*\|_n^2$ . In this case we have  $\|f - \hat{f}_m\|_n^2 \leq \|f - f_m^*\|_n^2 \leq \eta_2 \|f - f_m^*\|_2^2 + \eta_2 C_n \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 \leq C_6 n^{-\frac{2m_0}{2m+d}} \leq C_{17} n^{-\frac{2m_0}{2m_{i+1}+d}}$  if  $f \in H^{m_0}$ , and  $\|f - \hat{f}_m\|_n^2 \leq C_6 n^{-\frac{2m_0}{2m_{i+1}+d}} h_1(n)$  if  $f \notin H^{m_0}$ . By direct calculation we have  $n^{-1/2} \|f - \hat{f}_m\|_n^{1-\frac{d}{2m_i}} \rho_{i+1}^{\frac{d}{2m_i}} = t_i n^{-\eta_3}$ , where  $\eta_3 > 0$  is a constant.

**Case 2:**  $\|f - \hat{f}_m\|_n^2 \geq \|f - f_m^*\|_n^2$ . By (C.30), we have

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n (f(x_i) - \hat{f}_m(x_i))^2 + \frac{\mu}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \\
& \leq \frac{1}{n} \sum_{i=1}^n (f(x_i) - f_m^*(x_i))^2 + \frac{\mu}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 + 2Tn^{-1/2} \|f - f_m^*\|_n^{1 - \frac{d}{2m_i}} \rho_{i+1}^{\frac{d}{2m_i}} \\
& \quad + 2Tn^{-1/2} \|f - \hat{f}_m\|_n^{1 - \frac{d}{2m_i}} \rho_{i+1}^{\frac{d}{2m_i}} \\
& \leq \frac{1}{n} \sum_{i=1}^n (f(x_i) - f_m^*(x_i))^2 + \frac{\mu}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 + 4Tn^{-1/2} \|f - \hat{f}_m\|_n^{1 - \frac{d}{2m_i}} \rho_{i+1}^{\frac{d}{2m_i}}.
\end{aligned}$$

Under this case we have two cases:

**Case 2.1:**  $\|f - \hat{f}_m\|_n^2 \leq 4Tn^{-1/2} \|f - \hat{f}_m\|_n^{1 - \frac{d}{2m_i}} \rho_{i+1}^{\frac{d}{2m_i}}$ . Therefore,  $\|f - \hat{f}_m\|_n^{1 + \frac{d}{2m_i}} \leq 4Tn^{-1/2} \rho_{i+1}^{\frac{d}{2m_i}}$ . By direct calculation we can show that  $n^{-1/2} \|f - \hat{f}_m\|_n^{1 - \frac{d}{2m_i}} \rho_{i+1}^{\frac{d}{2m_i}} = t_i n^{-\eta_4}$ , where  $\eta_4 > 0$  is a constant.

**Case 2.2:**  $\|f - \hat{f}_m\|_n^2 \geq 4Tn^{-1/2} \|f - \hat{f}_m\|_n^{1 - \frac{d}{2m_i}} \rho_{i+1}^{\frac{d}{2m_i}}$ . In this case we have  $\|f - \hat{f}_m\|_n^2 \leq 4\|f - f_m^*\|_n^2 + \frac{4\mu}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2$ . Therefore, similar to Case 1,  $n^{-1/2} \|f - \hat{f}_m\|_n^{1 - \frac{d}{2m_i}} \rho_{i+1}^{\frac{d}{2m_i}} = t_i n^{-\eta_5}$ , where  $\eta_5 > 0$  is a constant.

Combining these three cases, taking  $\eta_6 = \min\{\eta_3, \eta_4, \eta_5\}$ , and noting  $T = n^{\frac{\eta_6}{2}}$ , we have with probability at least  $1 - C_{18} \exp(-C_{19}n^{\eta_6}) - C_{15} \exp(-C_{16}n^{\eta_0})$ ,  $2\langle e, f - \hat{f}_m \rangle \leq n^{-\eta_6/2} t_m$ , and complete the proof.

## C.6 Proof of Lemma C.3.1

Before the proof, we present the following lemma, whose proof can be found in Appendix C.12.

**Lemma C.6.1.** Define  $f_m^*$  as

$$f_m^* = \operatorname{argmin}_{\hat{f} \in \mathcal{N}_{\Psi_m}(\mathbb{R}^d)} \|f - \hat{f}\|_2^2 + \frac{C_1 \mu_m}{n} \|\hat{f}\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2, \quad (\text{C.31})$$

where  $C_1$  is a constant. For  $m \in [m_{\min}, m_0 - \epsilon_n]$ , we have

$$\|f - f_m^*\|_2^2 + \frac{C_1 \mu_m}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 \geq C_2 n^{\frac{-2m}{2m+d}},$$

where  $C_2$  is a constant depending on  $f$ ,  $\Omega$ ,  $C$  and  $C_1$ .

Now we are ready to prove Lemma C.3.1.

*Proof of Lemma C.3.1:*

Let  $m_{\min} = m_1 < m_2 < \dots < m_p = m_0 - \epsilon_n$  be a partition of  $[m_{\min}, m_0 - \epsilon_n]$ ,  $\mathcal{G}_i = \{g : g = f - \hat{f}_m, \forall m \in [m_i, m_{i+1}], \hat{f}_m \in \mathcal{C}\}$ , and  $\mathcal{G} = \bigcup \mathcal{G}_i$ . Consider class  $\mathcal{F}_i = \{g : g = \hat{f}_m, \forall m \in [m_i, m_{i+1}], \hat{f}_m \in \mathcal{C}\}$ . The bracket entropy number of class  $\mathcal{G}_i$  is the same as it of class  $\mathcal{F}_i$ . Note that  $\mathcal{F}_i \subset H^{m_i}$ , and  $\sup_{g \in \mathcal{F}_i} \|g\|_{H^{m_i}} \leq C_3 / \log n$  for some constant  $C_3$  and for any  $1 \leq i \leq p$ .

Similar to the proof of Lemma C.2.2, we will use Lemma C.5.2. Before using Lemma C.5.2, we need to verify the entropy condition of Lemma C.5.2,  $H_B(\delta_n/V(\Omega), \mathcal{F}_i, \|\cdot\|_\infty) \leq \frac{n\delta_n^2}{1200V(\Omega)K^2}$ , holds.

Since  $\mathcal{F}_i \in H^{m_i}(C_3/\log n)$ , the bracket entropy can be bounded by

$$H_B(\delta_n/V(\Omega), \mathcal{F}_i, \|\cdot\|_\infty) \leq C_4 \left( \frac{1}{\delta_n \log n} \right)^{d/m_i}$$

for some constant  $C_4$ . By interpolation inequality,

$$\|\hat{f}_m\|_\infty \leq C_5 \|\hat{f}_m\|_2^{1-\frac{d}{2m}} \|\hat{f}_m\|_{H_m}^{\frac{d}{2m}} \leq C_5 \|\hat{f}_m\|_{H_m} \leq \frac{C_6}{\log n}.$$

Therefore, the entropy condition is satisfied if for any fixed constant  $C_7$ ,

$$C_7 \left( \frac{1}{\delta_n \log n} \right)^{d/m_i} \leq n \delta_n^2 (\log n)^2, \quad (\text{C.32})$$

where  $\delta_n = C_8 n^{\frac{-m_{i+1}}{2m_{i+1}+d}}$  with  $C_8$  a constant determined later. By (C.32), we have that the entropy condition is satisfied if for any fixed constant  $C_9$ ,

$$C_9 \leq n^{-\frac{(m_{i+1}-m_i)d}{(2m_{i+1}+d)m_i}} (\log n)^{2+\frac{d}{m_i}} \Leftrightarrow \log C_9 \leq -\frac{(m_{i+1}-m_i)d}{(2m_{i+1}+d)m_i} \log n + \frac{2m_i+d}{m_i} \log \log n.$$

If we pick  $m_{i+1} - m_i \leq C_{10} \frac{\log \log n}{\log n}$ , the condition of Lemma C.5.2 is satisfied.

Next we show there exists a constant  $C_8$  such that  $\|f - \hat{f}_m\|_2^2 \geq C_8 n^{\frac{-m_{i+1}}{2m_{i+1}+d}}$  for  $f - \hat{f}_m \in \mathcal{G}_i$ .

By Lemma C.6.1 and extension, and that  $f_m^*$  is the solution to (C.31), we have

$$\begin{aligned} \|f - f_m^*\|_2^2 + \frac{C_1 \mu_m}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 &\leq \|f - \hat{f}_m\|_2^2 + \frac{C_1 \mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 \\ &\leq 1/C_{11} (\|f - \hat{f}_m\|_2^2 + \frac{C_1 \mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2) \\ &\leq 1/C_{11} (\|f - \hat{f}_m\|_2^2 + C_{12} t_m), \end{aligned}$$

which indicates there exists a constant  $C_8$  such that

$$\begin{aligned} \|f - \hat{f}_m\|_2^2 &\geq C_{11} (\|f - f_m^*\|_2^2 + \frac{C_2 \mu}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2) - C_{12} t_m \\ &\geq C_{13} n^{\frac{-2m}{2m+d}} - C_{12} t_m \geq C_8 n^{\frac{-2m}{2m+d}}, \end{aligned} \quad (\text{C.33})$$

since  $t_m = C_0 n^{\frac{-2m}{2m+d}} / \log n$ .

Therefore, the condition of Lemma C.5.2 holds. By applying Lemma C.5.2, there exist constants  $C_{14}$ ,  $C_{15}$ ,  $\eta_0$ ,  $\eta_1$ , and  $\eta_2$ , such that with probability at least  $1 - C_{14} \exp(-C_{15} n^{\eta_0})$ ,  $\eta_1 \|g\|_2 \leq \|g\|_n \leq \eta_2 \|g\|_2$  for  $g \in \mathcal{G}_i$ .

By Lemma C.5.1, it can be shown that there exists a constant  $C_{16}$  such that for any

$T > C_{16}$

$$P\left(\sup_{g \in H^{m_i}(\frac{C_1}{\log n})} \frac{|\frac{1}{n} \sum_{i=1}^n e_i g(x_i)|}{\|g\|_n^{1-\frac{d}{2m_i}}} \geq T n^{-1/2} \left(\frac{C_1}{\log n}\right)^{\frac{d}{2m_i}}\right) \leq C_3 \exp(-C_4 T^2).$$

The rest of the proof is similar to the proof of the proof of Theorem C.2.2, where the difference is that we need to show that  $n^{-1/2} \left(\frac{C_1}{\log n}\right)^{\frac{d}{2m_i}} \|f - \hat{f}_m\|_n^{1-\frac{d}{2m_i}}$  is small compared with  $t_{i+1}$ , which can be done by using similar approach shown in Step 4 of the proof of Theorem C.2.2. We do not present the proof here.

### C.7 Proof of Lemma 3.3.1

There are two cases in Lemma 3.3.1. First We prove the case  $g \in H^{m_0}(\mathbb{R}^d)$ .

Since  $g \in H^{m_0}(\mathbb{R}^d)$ , we have

$$\int_{\mathbb{R}^d} |\mathcal{F}(g)(\omega)|^2 (1 + \|\omega\|^2)^{m_0} d\omega < \infty. \quad (\text{C.34})$$

By hyperspherical coordinate transformation, we transform  $\omega$  into a radial coordinate  $r$ , and  $d - 1$  angular coordinates  $\phi_1, \phi_2, \dots, \phi_d$ . Let  $\phi = (\phi_1, \phi_2, \dots, \phi_d)^T$ , and the Jacobian of the transformation be  $J$ . We can change the left-hand side in (C.34) by

$$\int_0^\infty \int_{[0, 2\pi]^{d-1}} |\mathcal{F}(g)(r, \phi)|^2 (1 + r^2)^{m_0} |\det(J)| d\phi dr. \quad (\text{C.35})$$

Let  $g_1(r) = (1 + r^2)^{m_0} \int_{[0, 2\pi]^{d-1}} |\mathcal{F}(g)(r, \phi)|^2 |\det(J)| d\phi$ . Therefore, (C.35) is equal to  $\int_0^\infty g_1(r) dr$ , which is finite. It is enough to find an increasing function  $h_1(r)$  satisfies

$$\int_0^\infty g_1(r) h_1(r) dr = \infty \quad (\text{C.36})$$

and

$$\lim_{r \rightarrow +\infty} \frac{\log h_1(r)}{\log r} = 0. \quad (\text{C.37})$$

This is because

$$\int_{\mathbb{R}^d} |\mathcal{F}(g)(\omega)|^2 h_1(\|\omega\|) (1 + \|\omega\|^2)^{m_0 - \epsilon_1} d\omega < \infty \quad (\text{C.38})$$

for any  $\epsilon_1 > 0$  follows (C.37). To be more specify, suppose (C.38) is not true, which means there exists an  $\epsilon_1 > 0$  such that

$$\int_{\mathbb{R}^d} |\mathcal{F}(g)(\omega)|^2 h_1(\|\omega\|) (1 + \|\omega\|^2)^{m_0 - \epsilon_1} d\omega = \infty. \quad (\text{C.39})$$

By (C.37), there exists a constant  $C_1$  such that for  $r > C$ ,  $\frac{\log h_1(r)}{\log r} < \epsilon_1$ , which is the same as  $h_1(r) < r^{\epsilon_1}$ . Therefore, by (C.39), we have

$$\begin{aligned} \infty &= \int_{\mathbb{R}^d} |\mathcal{F}(g)(\omega)|^2 h_1(\|\omega\|) (1 + \|\omega\|^2)^{m_0 - \epsilon_1} d\omega \\ &< C_2 \int_{\mathbb{R}^d} |\mathcal{F}(g)(\omega)|^2 (1 + \|\omega\|^2)^{m_0 - \epsilon_1/2} d\omega < \infty, \end{aligned}$$

which leads to a contradiction.

We construct  $h_1(r)$  by the following way. Let  $\alpha_i = 2^{-i}$  for  $i \in \mathbb{N}_+$  and  $x_1 = 1$ . Let  $h_1(r) = 1$  for  $0 \leq r < x_1$ . Since  $\int_0^\infty g_1(r) r^\alpha dr = \infty$  for any  $\alpha > 0$ , there exists  $x_2 > x_1^{x_1}$  such that  $\int_{x_1}^{x_2} g_1(r) r^{\alpha_1} dr > 1$ . Let  $h(r) = x_1^{\alpha_0 - \alpha_1} r^{\alpha_1}$  for  $r \in (x_1, x_2]$ . Suppose we have defined  $h(r)$  for  $r \in (x_{i-1}, x_i]$ . There exists an  $x_{i+1} > x_i^{x_i}$  such that  $\int_{x_i}^{x_{i+1}} g_1(r) r^{\alpha_i} dr > 1$ . Take  $h_1(r) = h_1(x_i) x_i^{-\alpha_i} r^{\alpha_i}$  for  $r \in (x_i, x_{i+1}]$ . It can be seen that

$$h_1(r) = \left[ \prod_{j=1}^i x_j^{\alpha_{j-1} - \alpha_j} \right] r^{\alpha_i}$$

and  $h_1(r)$  is an increasing function satisfying (C.36). Next we show  $h_1(r)$  satisfies (C.37).

For any  $\epsilon > 0$ , there exists an integer  $N$  such that for  $x > x_N$ ,

$$\begin{aligned} \frac{\log h_1(r)}{\log r} &= \frac{\left(\sum_{i=1}^{N-1} (\alpha_{i-1} - \alpha_i) \log x_i\right) + (\alpha_{N-1} - \alpha_N) \log x_N + \alpha_N \log r}{\log r} \\ &< \frac{\left(\sum_{i=1}^{N-1} (\alpha_{i-1} - \alpha_i) \log x_i\right)}{\log r} + \frac{(\alpha_{N-1} - \alpha_N) \log x_N + \alpha_N \log r}{\log r} \\ &< \frac{1}{x_{N-1}} + \alpha_{N-1} < \epsilon, \end{aligned}$$

since  $x_{N-1} \rightarrow \infty$  and  $\alpha_{N-1} \rightarrow 0$ .

The proof of case  $g \notin H^{m_0}(\mathbb{R}^d)$  is similar. The difference is that since we have  $\int_0^\infty g_1(r)r^{-\alpha}dr < \infty$ , we can pick  $\alpha_i = 2^{-i}$  and pick  $x_i$  sequentially satisfying  $\int_{x_i}^\infty g_1(r)r^{-\alpha_i}dr < 2^{-i}$ . Taking  $h_2(r) = h_2(x_i)x_i^{-\alpha_i}r^{\alpha_i}$  for  $r \in (x_i, x_{i+1}]$ , we finish the proof.

### C.8 Proof of Lemma C.5.3

It suffices to show that the probability of  $\|\hat{f}_m\|_2^2 \geq C_3$  can be bounded by some small number, where  $C_3$  is some constant which will be specified later. Choose function  $\hat{f} \equiv 0 \in \mathcal{N}_{\Psi_m}(\Omega)$ . Combing (3.2) with (3.1), we have

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n (f(x_i) - \hat{f}_m(x_i))^2 + \frac{2}{n} \sum_{i=1}^n e_i (f(x_i) - \hat{f}_m(x_i)) + \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \quad (\text{C.40}) \\ &\leq \frac{1}{n} \sum_{i=1}^n f(x_i)^2 + \frac{2}{n} \sum_{i=1}^n e_i f(x_i) \leq \frac{2}{n} \sum_{i=1}^n f(x_i)^2 + \frac{1}{n} \sum_{i=1}^n e_i^2, \end{aligned}$$

where the last inequality is because of Cauchy-Schwartz inequality. Using Cauchy-Schwartz inequality again, we have

$$\frac{2}{n} \sum_{i=1}^n e_i (f(x_i) - \hat{f}_m(x_i)) \geq -\frac{1}{2n} \sum_{i=1}^n (f(x_i) - \hat{f}_m(x_i))^2 - \frac{8}{n} \sum_{i=1}^n e_i^2, \quad (\text{C.41})$$

since  $\sum_{i=1}^n e_i^2 > 0$  with probability one. From (C.40) and (C.41), it can be seen that

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n (f(x_i) - \hat{f}_m(x_i))^2 + \frac{2}{n} \sum_{i=1}^n e_i (f(x_i) - \hat{f}_m(x_i)) + \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \\ & \geq \frac{1}{2n} \sum_{i=1}^n (f(x_i) - \hat{f}_m(x_i))^2 + \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 - \frac{8}{n} \sum_{i=1}^n e_i^2. \end{aligned} \quad (\text{C.42})$$

By (C.40) and (C.42), we have

$$\frac{1}{2n} \sum_{i=1}^n (f(x_i) - \hat{f}_m(x_i))^2 + \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq \frac{9}{n} \sum_{i=1}^n e_i^2 + \frac{2}{n} \sum_{i=1}^n f(x_i)^2. \quad (\text{C.43})$$

By Cauchy-Schwartz inequality, it can be shown that

$$\frac{1}{n} \sum_{i=1}^n (f(x_i) - \hat{f}_m(x_i))^2 \geq \frac{1}{n} \sum_{i=1}^n \left( \frac{3}{4} \hat{f}_m(x_i)^2 - 3f(x_i)^2 \right). \quad (\text{C.44})$$

Combining (C.43) and (C.44), we obtain

$$\frac{4}{3} \left( \frac{18}{n} \sum_{i=1}^n e_i^2 + \frac{7}{n} \sum_{i=1}^n f(x_i)^2 \right) \geq \frac{1}{n} \sum_{i=1}^n \hat{f}_m(x_i)^2.$$

By Hoeffding's inequality, we have

$$\begin{aligned} & P \left( \frac{1}{n} \sum_{i=1}^n f(x_i)^2 - \|f\|_2^2 \geq \|f\|_2^2 \right) \\ & \leq \exp \left( - \frac{2n^2 \|f_0\|_2^4}{4n \|f\|_\infty^4} \right) = \exp \left( - \frac{n \|f\|_2^4}{2 \|f\|_\infty^4} \right), \end{aligned}$$

By Bernstein inequality,

$$P \left( \frac{1}{n} \sum_{i=1}^n e_i^2 - \sigma_1^2 \geq 2\sqrt{\mathbb{E}(e_i^2 - \sigma_1^2)^2} \right) \leq \exp(-n).$$

Noting that  $e_i$ 's are i.i.d. normal distributed, we have with probability at least  $1 - \exp(-n)$ ,

$\frac{1}{n} \sum_{i=1}^n e_i^2 \leq C_4$ , where  $C_4 = 2\sqrt{\mathbb{E}(e_i^2 - \sigma_1^2)^2} + \sigma_1^2$  is a constant.

Therefore, with probability at least  $1 - \exp\left(-\frac{n\|f\|_2^4}{2\|f\|_\infty^4}\right) - \exp(-n)$ , we have

$$\frac{1}{n} \sum_{i=1}^n \hat{f}_m(x_i)^2 \leq C_5, \quad (\text{C.45})$$

where  $C_5 = 4(18C_4 + 14\|f\|_2^2)/3$ .

Recall that  $\mathcal{C} = \{\hat{f}_m : \forall m \in [m_0 + \epsilon_n, m_{\max}], \frac{\mu_m}{n} \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}^2 \leq t_m\}$ . Since  $\hat{f}_m \in \mathcal{C}$ ,  $\|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)} \leq \rho_m$ , where  $\rho_m = (nt_m/\mu_m)^{1/2}$ . Since  $\|\cdot\|_{\mathcal{N}_{\Psi_m}(\Omega)}$  and  $\|\cdot\|_{H^m}$  are equivalent, by interpolation theorem,  $\|\hat{f}_m\|_\infty \leq C_6 \|\hat{f}_m\|_{\mathcal{N}_{\Psi_m}(\Omega)}$ , where  $C_6$  is a constant.

Let  $\mathcal{G} = \{g : g = \hat{f}_m, \|\hat{f}_m\|_2^2 \geq C_3, \hat{f}_m \in \mathcal{C}, m \in [m_0 + \epsilon_n, m_{\max}]\}$ . Let  $m_0 + \epsilon_n = m_1 < m_2 < \dots < m_p = m_{\max}$  be a partition, and let  $\mathcal{G}_i = \{g : g = \hat{f}_m, \|\hat{f}_m\|_2^2 \geq C_3, \hat{f}_m \in \mathcal{C}, \forall m \in [m_i, m_{i+1}]\}$ . Obviously  $\bigcup_{i=1}^{p-1} \mathcal{G}_i = \mathcal{G}$ . We will use Lemma C.5.2 to link the empirical norm shown in (C.45) and the  $L_2$  norm. In order to use Lemma C.5.2, we need to check the entropy condition of Lemma C.5.2,  $H_B(\delta_n/V(\Omega), \mathcal{F}_i, \|\cdot\|_\infty) \leq \frac{n\delta_n^2}{1200V(\Omega)K^2}$ , holds.

Consider class  $\mathcal{F}_i = \{g : g = \hat{f}_m / ((C_6 + 1)\rho_m), \hat{f}_m \in \mathcal{G}_i\}$ . Thus,  $\|g\|_\infty < 1$  for all  $g \in \mathcal{F}_i$ . Since  $\mathcal{F}_i \subset H^{m_i}$ , the bracket entropy  $H_B(\delta_n/V(\Omega), \mathcal{F}_i, \|\cdot\|_\infty)$  satisfies

$$H_B(\delta_n/V(\Omega), \mathcal{F}_i, \|\cdot\|_\infty) \leq C_7 \left(\frac{1}{\delta_n}\right)^{d/m_i}.$$

The entropy condition of Lemma C.5.2 is satisfied if for any fixed constant  $C_8$ ,

$$C_8 \left(\frac{1}{\delta_n}\right)^{d/m_i} \leq n\delta_n^2. \quad (\text{C.46})$$

Direct calculation shows that if  $m_{i+1} - m_i = 1/(C_9 \log n)$ , where  $C_9$  is a constant which is related to  $m_0$  and  $m_{\max}$ , (C.46) is satisfied, and thus the entropy condition of Lemma C.5.2 is satisfied. Therefore, by Lemma C.5.2, there exists a constant  $\eta$ , which only relates

to  $V(\Omega)$ , such that

$$P\left(\inf_{g \in \mathcal{F}_i} \frac{\|g\|_n^2}{\|g\|_2^2} < \eta\right) \leq C_{10} \exp(-C_{11}\mu_{m_i}),$$

which indicates

$$P\left(\inf_{g \in \mathcal{G}} \frac{\|g\|_n^2}{\|g\|_2^2} < \eta\right) \leq C_{12} \exp(-C_{13}\mu_{m_0}), \quad (\text{C.47})$$

by taking the summation of the probabilities.

Combine (C.47) with (C.45), and choose  $C_3 = C_5/\eta + 1$ ,

$$P\left(g \in \mathcal{G}, \inf_{g \in \mathcal{G}} \frac{\|g\|_n^2}{\|g\|_2^2} \geq \eta, g \in \mathcal{C}\right) \leq \exp\left(-\frac{n\|f\|_2^4}{2\|f\|_\infty^4}\right) + \exp(-n).$$

By (C.47), it can be shown that

$$P\left(g \in \mathcal{G}, \inf_{g \in \mathcal{G}} \frac{\|g\|_n^2}{\|g\|_2^2} < \eta, g \in \mathcal{C}\right) \leq C_{12} \exp(-C_{13}\mu_{m_0}).$$

Therefore, we have

$$\begin{aligned} P(g \in \mathcal{G}, g \in \mathcal{C}) &= P\left(g \in \mathcal{G}, \inf_{g \in \mathcal{G}} \frac{\|g\|_n^2}{\|g\|_2^2} \geq \eta, g \in \mathcal{C}\right) + P\left(g \in \mathcal{G}, \inf_{g \in \mathcal{G}} \frac{\|g\|_n^2}{\|g\|_2^2} < \eta, g \in \mathcal{C}\right) \\ &\leq C_{12} \exp(-C_{13}\mu_{m_0}) + \exp\left(-\frac{n\|f\|_2^4}{2\|f\|_\infty^4}\right) + \exp(-n), \end{aligned}$$

which indicates with probability at least  $1 - C_{12} \exp(-C_{13}\mu_{m_0}) - \exp\left(-\frac{n\|f\|_2^4}{2\|f\|_\infty^4}\right) - \exp(-n)$ ,

$$\|\hat{f}_m\|_2^2 \leq C_3.$$

By noting that  $1 - C_{12} \exp(-C_{13}\mu_{m_0}) - \exp\left(-\frac{n\|f\|_2^4}{2\|f\|_\infty^4}\right) - \exp(-n) \geq 1 - C_1 \exp(-C_2\mu_{m_0})$

for some constant  $C_1$  and  $C_2$ , we finish the proof.

### C.9 Proof of Lemma C.5.4

Suppose  $f \in H^{m_0}$ . By (C.15), we have

$$\|f - f_{m_0}^*\|_2^2 + \frac{C_1 \mu_{m_0}}{n} \|f_{m_0}^*\|_{\mathcal{N}_{\Psi_{m_0}}(\mathbb{R}^d)}^2 \leq \frac{C_1 \mu_{m_0}}{n} \|f\|_{\mathcal{N}_{\Psi_{m_0}}(\mathbb{R}^d)}^2 \leq C_2 n^{-\frac{2m}{2m+d}}.$$

If  $f \notin H^{m_0}$ , let  $C_n = C_1 \mu_{m_0}/n$ . By Fourier transform and (C.15), similar to the proof of Lemma C.2.2 (see Appendix C.5), we have

$$\begin{aligned} & \|f - f_{m_0}^*\|_2^2 + C_n \|f_{m_0}^*\|_{\mathcal{N}_{\Psi_{m_0}}(\mathbb{R}^d)}^2 \\ &= \int_{\mathbb{R}^d} |\mathcal{F}(f)(\omega) - \mathcal{F}(f_{m_0}^*)(\omega)|^2 d\omega + C_n \int_{\mathbb{R}^d} |\mathcal{F}(f_{m_0}^*)(\omega)|^2 (1 + |\omega|^2)^m d\omega \\ &= \int_{\mathbb{R}^d} |\mathcal{F}(f)(\omega) - \mathcal{F}(f_{m_0}^*)(\omega)|^2 + C_n |\mathcal{F}(f_{m_0}^*)(\omega)|^2 (1 + |\omega|^2)^m d\omega \\ &= \int_{\mathbb{R}^d} \frac{C_n (1 + |\omega|^2)^{m_0}}{1 + C_n (1 + |\omega|^2)^{m_0}} |\mathcal{F}(f)(\omega)|^2 d\omega \\ &= \int_{\Omega_1} \frac{C_n (1 + |\omega|^2)^{m_0}}{1 + C_n (1 + |\omega|^2)^{m_0}} |\mathcal{F}(f)(\omega)|^2 d\omega + \int_{\Omega_1^c} \frac{C_n (1 + |\omega|^2)^{m_0}}{1 + C_n (1 + |\omega|^2)^{m_0}} |\mathcal{F}(f)(\omega)|^2 d\omega \\ &\leq \int_{\Omega_1} C_n (1 + |\omega|^2)^{m_0} |\mathcal{F}(f)(\omega)|^2 d\omega + \int_{\Omega_1^c} |\mathcal{F}(f)(\omega)|^2 d\omega \\ &\leq C_3 C_n h_2(n) \int_{\Omega_1} (1 + |\omega|^2)^{m_0} \frac{|\mathcal{F}(f)(\omega)|^2}{h_2(|\omega|)} d\omega + C_3 C_n \int_{\Omega_1^c} (1 + |\omega|^2)^{m_0} \frac{|\mathcal{F}(f)(\omega)|^2}{h_2(|\omega|)} d\omega \\ &\leq C_4 n^{-\frac{2m_0}{2m_0+d}} h_2(n), \end{aligned}$$

where  $h_2(|\omega|)$  is defined in (3.20), and  $\Omega_1 = \{\omega : C_n (1 + |\omega|^2)^{m_0} < h_2(|\omega|)\}$ . Therefore, we finish the proof.

### C.10 Proof of Lemma C.5.2

Before the proof, we present the following lemma, which states Bernstein's inequality for a single  $g$ . See, for example, [108].

**Lemma C.10.1.** *Suppose  $X_i \sim Unif(\Omega)$  for  $i = 1, \dots, n$ . Let  $Z_i = (\|g\|_2^2/V(\Omega) - g(X_i)^2)/\|g\|_{N_{\Psi_m}(\Omega)}^2$ . Therefore,  $E(Z_i) = 0$ . 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

$$P\left(\frac{\|g\|_2^2/V(\Omega)}{\|g\|_{N_{\Psi_m}(\Omega)}^2} - \frac{\|g\|_n^2}{\|g\|_{N_{\Psi_m}(\Omega)}^2} \geq t\right) \leq \exp\left[-\frac{nt^2/2}{E(Z_1^2) + bt/3}\right].$$

Now we can prove Lemma C.5.2. Take  $g \in \mathcal{G}$ , and suppose that  $s\delta_n \leq \|g\|_2 \leq (s+1)\delta_n$ , where  $s \in \{2, \dots\}$ . Furthermore, let  $-K \leq g_L \leq g \leq g_U \leq K$ , and  $\|g_U - g_L\|_\infty \leq \delta_n/V(\Omega)$ , for functions  $g_L$  and  $g_U$ . For  $0 < C \leq \frac{1}{4V(\Omega)}$ , by Cauchy-Schwartz inequality, we have

$$g_L^2 \leq 2g^2/C + 2C(g - g_L)^2 \leq 2g^2/C + 2C\delta_n^2/V(\Omega)^2,$$

which indicates

$$2\|g\|_n^2 \geq C\|g_L\|_n^2 - 2C^2\delta_n^2/V(\Omega)^2.$$

The inequality  $\|g\|_n^2/\|g\|_2^2 < \eta_1$  implies

$$\begin{aligned} \|g_L\|_n^2 - \|g_L\|_2^2/V(\Omega) &\leq 2\eta_1\|g\|_2^2/C - \|g_L\|_2^2/V(\Omega) + 2C\delta_n^2/V(\Omega)^2 \\ &\leq 2\eta_1(s+1)^2\delta_n^2/C - (s-1)^2\delta_n^2/V(\Omega) + 2C\delta_n^2/V(\Omega)^2 \\ &\leq 2\eta_1(s+1)^2\delta_n^2/C - (s-1)^2\delta_n^2/V(\Omega) + 2C\delta_n^2/V(\Omega)^2. \end{aligned}$$

By choosing appropriate  $C$  and  $\eta_1$  (the choice only depends on  $V(\Omega)$ ), we have

$$\|g_L\|_n^2 - \|g_L\|_2^2/V(\Omega) \leq -\frac{1}{2}(s-1)^2\delta_n^2/V(\Omega). \quad (\text{C.48})$$

Note that

$$\left| \|g_L\|_n^2 - \|g_L\|_2^2/V(\Omega) \right| \leq K^2 \quad (\text{C.49})$$

and

$$E(g_L^2 - \|g_L\|_2^2/V(\Omega))^2 \leq 4K^2\|g_L\|_2^2/V(\Omega) \leq 4K^2(s+2)^2\delta_n^2/V(\Omega). \quad (\text{C.50})$$

Combining (C.48), (C.49) and (C.50) with Lemma C.10.1, we have

$$\begin{aligned} & P\left(\|g_L\|_2^2/V(\Omega) - \|g_L\|_n^2 \geq \frac{1}{2V(\Omega)}(s-1)^2\delta_n^2\right) \\ & \leq \exp\left[-\frac{n\frac{1}{8V(\Omega)^2}(s-1)^4\delta_n^4}{4K^2(s+2)^2\delta_n^2/V(\Omega) + K^2\frac{1}{6V(\Omega)}(s-1)^2\delta_n^2}\right] \\ & \leq \exp\left[-\frac{n\frac{1}{8V(\Omega)}(s-1)^4\delta_n^4}{4K^2(s+2)^2\delta_n^2 + K^2\frac{1}{6}(s-1)^2\delta_n^2}\right] \\ & \leq \exp\left[-\frac{1}{8V(\Omega)}\frac{n(s-1)^2\delta_n^2}{36K^2 + K^2\frac{1}{6}}\right] \\ & \leq \exp\left[-\frac{1}{8V(\Omega)}\frac{n(s-1)^2\delta_n^2}{37K^2}\right] \\ & \leq \exp\left[-\frac{n(s-1)^2\delta_n^2}{296V(\Omega)K^2}\right] \end{aligned} \quad (\text{C.51})$$

Therefore, by taking all  $g \in \mathcal{G}$ , we have

$$P\left(\inf_{\|g\|_2 \geq 2\delta_n} \frac{\|g\|_n^2}{\|g\|_2^2} < \eta_1\right) \leq \sum_{s=2}^{\infty} \exp\left[H_B(\delta_n/V(\Omega), \mathcal{G}', \|\cdot\|_{\infty}) - \frac{n(s-1)^2\delta_n^2}{300V(\Omega)K^2}\right].$$

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

$$\begin{aligned} P\left(\inf_{\|g\|_2 \geq 2\delta_n} \frac{\|g\|_n^2}{\|g\|_2^2} < \eta_1\right) &\leq \sum_{s=2}^{\infty} \exp\left[-\frac{n(s-1)^2\delta_n^2}{1200V(\Omega)K^2}\right] \\ &\leq C_1 \exp(-C_2 n\delta_n^2/K^2) \end{aligned}$$

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

For  $C_0 \leq \frac{1}{4V(\Omega)}$ , we have

$$g^2 \leq 2g_R^2/C_0 + 2C_0(g - g_R)^2 \leq 2g_R^2/C_0 + 2C_0\delta_n^2/V(\Omega)^2,$$

which indicates

$$\|g\|_n^2 \leq 2\|g_R\|_n^2/C_0 + 2C_0\delta_n^2/V(\Omega)^2.$$

The inequality  $\|g\|_n^2/\|g\|_2^2 > \eta_2$  implies

$$\begin{aligned} \|g_R\|_n^2 - \|g_R\|_2^2/V(\Omega) &\geq \frac{1}{2}\eta_2 C_0 s^2 \delta_n^2 - \|g_R\|_2^2/V(\Omega) - C_4^2 \delta_n^2/V(\Omega)^2 \\ &\geq \frac{1}{2}\eta_2 C_0 s^2 \delta_n^2 - (s-1)^2 \delta_n^2/V(\Omega) - C_0^2 \delta_n^2/V(\Omega)^2. \end{aligned}$$

By choosing appropriate  $C_0$  and  $\eta_2$ , we have

$$\|g_R\|_n^2 - \|g_R\|_2^2/V(\Omega) \geq \frac{1}{4}(s-1)^2 \delta_n^2/V(\Omega). \quad (\text{C.52})$$

Note that

$$\left| \|g_R\|_n^2 - \|g_R\|_2^2/V(\Omega) \right| \leq K^2 \quad (\text{C.53})$$

and

$$E(g_R^2 - \|g_R\|_2^2/V(\Omega))^2 \leq 4K^2\|g_R\|_2^2/V(\Omega) \leq 4K^2(s+2)^2\delta_n^2/V(\Omega). \quad (\text{C.54})$$

By combining (C.52), (C.53) and (C.54) with Lemma C.10.1, similar to (C.51), we have

$$P\left(\|g_R\|_2^2/V(\Omega) - \|g_R\|_n^2 \geq \frac{1}{4V(\Omega)}(s-1)^2\delta_n^2\right) \leq \exp\left[-\frac{n(s-1)^2\delta_n^2}{600V(\Omega)K^2}\right].$$

Therefore, by taking all  $g \in \mathcal{G}$ , we have

$$P\left(\inf_{\|g\|_2 \geq 2\delta_n} \frac{\|g\|_n^2}{\|g\|_2^2} < \eta_2\right) \leq \sum_{s=2}^{\infty} \exp\left[H_B(\delta_n/V(\Omega), \mathcal{G}', \|\cdot\|_{\infty}) - \frac{n(s-1)^2\delta_n^2}{600V(\Omega)K^2}\right].$$

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

$$\begin{aligned} P\left(\inf_{\|g\|_2 \geq 2\delta_n} \frac{\|g\|_n^2}{\|g\|_2^2} < \eta_2\right) &\leq \sum_{s=2}^{\infty} \exp\left[-\frac{n(s-1)^2\delta_n^2}{1200V(\Omega)K^2}\right] \\ &\leq C_3 \exp(-C_4\delta_n^2/K^2) \end{aligned}$$

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

### C.11 Proof of Lemma C.5.4

By Fourier transform, we have

$$\begin{aligned}
& \|f - f_m^*\|_2^2 + \frac{C_1\mu_m}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 \\
&= \int_{\mathbb{R}^d} |\mathcal{F}(f)(\omega) - \mathcal{F}(f_m^*)(\omega)|^2 d\omega + \frac{C_2\mu_m}{n} \int |\mathcal{F}(f_m^*)(\omega)|^2 (1 + |\omega|^2)^m d\omega \\
&= \int_{\mathbb{R}^d} |\mathcal{F}(f)(\omega) - \mathcal{F}(f_m^*)(\omega)|^2 + \frac{C_2\mu_m}{n} |\mathcal{F}(f_m^*)(\omega)|^2 (1 + |\omega|^2)^m d\omega \\
&\geq C_3 \int_{\mathbb{R}^d} \frac{\frac{C_2\mu_m}{n} |\omega|^{2m} |\mathcal{F}(f)(\omega)|^2}{1 + \frac{C_2\mu_m}{n} |\omega|^{2m}} d\omega \\
&\geq C_4 \int_{|\omega|^2 < (\frac{n}{C_2\mu_m})^{1/m}} \frac{C_2\mu_m}{n} |\omega|^{2m} |\mathcal{F}(f)(\omega)|^2 d\omega,
\end{aligned}$$

where the first equality is because  $\|\cdot\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}$  is equivalent to  $\|\cdot\|_{H^m(\mathbb{R}^d)}$ , and the first inequality is by (C.18). We prove the results of Lemma C.5.4 by contradiction.

First, let  $f \notin H^{m_0}(\mathbb{R}^d)$ . Suppose there exists some  $s_1 > 0$  such that

$$\liminf_{n \rightarrow \infty} \frac{\mu_m}{nt_m} \int_{|\omega|^2 < (\frac{n}{C_2\mu_m})^{1/m}} |\omega|^{2m} |\mathcal{F}(f)(\omega)|^2 d\omega = s_1, \quad (\text{C.55})$$

where  $t_m$  is as defined in Lemma C.5.3. Therefore, we can pick a sub-sequence such that (for simplification, we still use  $n$  as the subscribe of this sequence)

$$\lim_{n \rightarrow \infty} \frac{\mu_m}{nt_m} \int_{|\omega|^2 < (\frac{n}{C_2\mu_m})^{1/m}} |\omega|^{2m} |\mathcal{F}(f)(\omega)|^2 d\omega = s_1.$$

Therefore, by summation by parts and (C.55), we have

$$\begin{aligned}
& \int_{|\omega|^2 < (\frac{n}{C_2\mu_m})^{1/m}} |\mathcal{F}(f)(\omega)|^2 |\omega|^{2m_0} d\omega \\
&= \int_{|\omega|^2 < (\frac{n}{C_5\mu_m})^{1/m}} |\mathcal{F}(f)(\omega)|^2 |\omega|^{2m} |\omega|^{2(m_0-m)} d\omega \\
&\leq C_5 \left(\frac{n}{\mu_m}\right)^{\frac{m_0-m}{m}} \int_{|\omega|^2 < (\frac{n}{C_2\mu_m})^{1/m}} |\mathcal{F}(f)(\omega)|^2 |\omega|^{2m} d\omega \\
&+ C_6 \frac{\log(n)^2}{n^{\frac{2(m_0-m)}{2m+d}}} \int_{|\omega|^2 < (\frac{n}{C_2\mu_m})^{1/m}} |\mathcal{F}(f)(\omega)|^2 |\omega|^{2m} d\omega \times \int \frac{1}{|\omega| \log(|\omega|)^2} d\omega \\
&< C_7,
\end{aligned}$$

where  $C_7$  is a constant related to  $s_1$  that does not change when  $n$  changes. Since  $f \notin H^{m_0}(\mathbb{R}^d)$ , we have

$$\infty = \int_{\mathbb{R}^d} |\mathcal{F}(f)(\omega)|^2 (1 + |\omega|^2)^{m_0} d\omega < C_8 \int_{\mathbb{R}^d} |\mathcal{F}(f)(\omega)|^2 |\omega|^{2m_0} d\omega < C_8 C_7,$$

which leads to a contradiction. The proof for the case that  $f \in H^{m_0}(\mathbb{R}^d)$  is similar.

## C.12 Proof of Lemma C.6.1

Pick any fixed orthogonal basis  $\phi_k$  for space  $L_2(\mathbb{R}^d)$ . Therefore, we have  $f = \sum_{k=1}^{\infty} \langle f, \phi_k \rangle \phi_k$ , where  $\langle f, \phi_k \rangle$  denote the inner product of  $f$  and  $\phi_k$  in  $L_2(\mathbb{R}^d)$ . Let  $a_k = \langle f, \phi_k \rangle$ . By the interpolation theorem, we have

$$\|f - f_m^*\|_2^2 + \frac{C_1\mu_m}{n} \|f_m^*\|_{\mathcal{N}_{\Psi_m}(\mathbb{R}^d)}^2 \geq \|f - f_m^*\|_2^2 + \frac{C_2\mu_m}{n} \|f_m^*\|_2^2.$$

Let  $f_1^*$  be the solution to the optimization problem

$$\min_{\hat{f} \in L_2(\mathbb{R}^d)} \|f - \hat{f}\|_2^2 + \frac{C_2\mu_m}{n} \|\hat{f}\|_2^2. \quad (\text{C.56})$$

Suppose  $f_1^* = \sum_{k=1}^{\infty} b_k \phi_k$ . Let  $C_n = C_2 \mu_m / n$ . Since  $f_1^*$  is the solution to (C.56), direct calculation shows  $b_k = \frac{1}{C_{n+1}} a_k$ . Therefore, we have

$$\begin{aligned} & \|f - f_m^*\|_2^2 + \frac{C_2 \mu_m}{n} \|f_m^*\|_2^2 \geq \|f - f_1^*\|_2^2 + \frac{C_2 \mu_m}{n} \|f_1^*\|_2^2 \\ &= \sum_{k=1}^{\infty} \frac{C_n}{C_n + 1} \langle f, \phi_k \rangle^2 \geq \frac{C_2 \mu_m}{(C_2 + 1)n} \sum_{k=1}^{\infty} \langle f, \phi_k \rangle^2 \\ &= C_3 n^{\frac{-2m}{2m+d}}, \end{aligned}$$

which finishes the proof.

### C.13 Asymptotic bounds of the determinant term

In this section, we provide an asymptotic lower bound and upper bound of  $\det(K_m + \mu_m I_n)$ .

#### C.13.1 Properties of eigenvalues

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

$$\Psi_m(x, y) = \sum_{k=1}^{\infty} \lambda_k^{(m)} \varphi_k^{(m)}(x) \varphi_k^{(m)}(y), \quad (\text{C.57})$$

where the summation is uniformly and absolutely convergent. We use the following notations. 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 use  $a_n \lesssim b_n$  to denote  $a_n \leq C b_n$  for some constant  $C > 0$ . The following lemma states the asymptotic property of eigenvalues.

**Lemma C.13.1.** *Let  $\lambda_k^{(m)}$  be as in (C.57). Then,  $\lambda_k^{(m)} \asymp k^{-2m/d}$ .*

*Proof.* Let  $T$  be the embedding operator of  $\mathcal{N}_{\Psi_m}(\Omega)$  into  $L_2(\Omega)$ , and  $T^*$  be the adjoint of

$T$ . By Proposition 10.28 in [46],

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

By Theorem 5.7 in [109],  $T$  and  $T^*$  have the same singular values. By Theorem 5.10 in [109], 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  $\mu_k$  denotes the singular value of  $T$ . By Theorem in Section 3.3.4 in [110], the embedding operator  $T$  has approximation number satisfying

$$C_3k^{-m/d} \leq a_k \leq C_4k^{-m/d}, \quad \forall k \in \mathbb{N}, \quad (\text{C.58})$$

where  $C_3$  and  $C_4$  are two positive numbers. Since  $m \in [m_{\min}, m_{\max}]$ , we can choose  $C_3$  and  $C_4$  that do not depend on  $m$ . By Theorem 5.7 in [109],  $T^*T\varphi_k = \mu_k^2\varphi_k$ , and  $T^*T\varphi_k = T^*\varphi_k = \lambda_k\varphi_k$ , we have  $\lambda_k = \mu_k^2$ . By (C.58),  $\lambda_k \asymp k^{-2m/d}$  holds.  $\square$

### C.13.2 Lower bound of the determinant term

We need the following lemma, which can be found in [111].

**Lemma C.13.2** (Minkowski determinant inequality). *Let  $A, B \in \mathbb{R}^{n \times n}$  be two symmetric, positive definite matrices. Thus,*

$$(\det(A + B))^{1/n} \geq (\det(A))^{1/n} + (\det(B))^{1/n}.$$

Let  $p = \lfloor n^{\frac{d}{2m+d}} \rfloor$ , where  $\lfloor \cdot \rfloor$  is the floor function. Let  $\Psi_1 = \frac{1}{\sqrt{n}}(\varphi_1^{(m)}(X), \dots, \varphi_p^{(m)}(X))$ , and  $\Psi_2 = \frac{1}{\sqrt{n}}(\varphi_{p+1}^{(m)}(X), \varphi_{p+2}^{(m)}(X), \dots)$ , where  $\varphi_k^{(m)}(X) = (\varphi_k^{(m)}(x_1), \dots, \varphi_k^{(m)}(x_n))^T$  for  $k = 1, 2, \dots$ , and  $\varphi_k^{(m)}$ 's are as in (C.57). Let  $\Lambda_1 = \text{diag}(n\lambda_1^{(m)}, \dots, n\lambda_p^{(m)})$  and  $\Lambda_2 = \text{diag}(n\lambda_{p+1}^{(m)}, \dots)$ , where  $\lambda_k^{(m)}$ 's are as in (C.57). Therefore,  $K_m = \sum_{k=1}^{\infty} \lambda_k^{(m)} \varphi_k^{(m)}(X) \varphi_k^{(m)}(X)^T = \Psi_1 \Lambda_1 \Psi_1^T + \Psi_2 \Lambda_2 \Psi_2^T$ .

By Lemma C.13.2, it follows that  $\det(K_m + \mu_m I_n) \geq \det(\Psi_1 \Lambda_1 \Psi_1^T + \mu_m I_n) + \det(\Psi_2 \Lambda_2 \Psi_2^T)$ , since both  $\Psi_1 \Lambda_1 \Psi_1^T + \mu_m I_n$  and  $\Psi_2 \Lambda_2 \Psi_2^T$  are positive definite and symmetric. Therefore, by basic matrix calculation, we have

$$\begin{aligned}
\det(K_m + \mu_m I_n) &\geq \det(\Psi_1 \Lambda_1 \Psi_1^T + \mu_m I_n) && \text{(C.59)} \\
&= \mu_m^n \det(\mu_m^{-1} \Psi_1 \Lambda_1 \Psi_1^T + I_n) \\
&= \mu_m^n \det(\mu_m^{-1} \Lambda_1 \Psi_1^T \Psi_1 + I_p) \\
&= \mu_m^{n-p} \det(\Lambda_1 \Psi_1^T \Psi_1 + \mu_m I_p) \\
&\geq \mu_m^{n-p} \det(\Lambda_1 \Psi_1^T \Psi_1) \\
&= \mu_m^{n-p} n^p \prod \lambda_i^{(m)} \det(\Psi_1^T \Psi_1).
\end{aligned}$$

Since  $\det(\Psi_1^T \Psi_1) \geq \lambda_{\min}(\Psi_1^T \Psi_1)^p$ , it suffices to give a lower bound of  $\lambda_{\min}(\Psi_1^T \Psi_1)$ . Consider  $u^T \Psi_1^T \Psi_1 u$ , where  $u = (u_1, \dots, u_p)^T \in \mathbb{R}^p$  with  $\|u\|_2 = 1$ . Let  $\mathcal{Q}_m = \{g : g = \sum_{i=1}^p u_i \varphi_i^{(m)}\}$  and let  $\mathcal{Q} = \bigcup_p \mathcal{Q}_p$ . Since  $\varphi_i^{(m)}$ 's are orthonormal,  $\|g\|_2 = 1$ . For any  $g \in \mathcal{Q}_m$ , by Lemma C.13.1,  $\|g\|_{H^m}^2 \leq \frac{1}{\lambda_p^{(m_{\max})}} \asymp p^{2m_{\max}/d}$ . By the interpolation inequality,

$$\|g\|_\infty \leq C \|g\|_{H^m}^{\frac{d}{2m}} = C \left( \sum_{j=1}^p \frac{u_j^2}{\lambda_j^{(m)}} \right)^{\frac{d}{4m}} \leq C (\lambda_p^{(m)})^{-\frac{d}{4m}} \leq C_1 n^{\frac{d}{2(2m+d)}},$$

where  $C$  and  $C_1$  are constants. We use Lemma E.2 to link  $\|g\|_n$  to  $\|g\|_2$ . Therefore, we need to check the conditions of Lemma E.2 hold. Since  $\|g\|_2 = 1$ , it suffices to check the entropy condition. Note that  $\|g\|_{H^m}^2 \leq C_2 n^{\frac{2m_{\max}}{2m_{\max}+d}}$ . Let  $\rho = C_2^{1/2} n^{\frac{m_{\max}}{2m_{\max}+d}}$ . Consider class  $\mathcal{Q}' = \{g : g = \frac{f}{\rho}, f \in \mathcal{Q}\}$ .

Since  $\mathcal{Q}' \subset H^{m_{\min}}$ , there exists a constant  $C_3$  such that

$$H_B(\delta_n/V(\Omega), \mathcal{F}', \|\cdot\|_\infty) \leq C_3 \left( \frac{1}{\delta_n} \right)^{d/m_{\min}}.$$

The entropy condition is satisfied if for any fixed constant  $C_4$ ,

$$C_4 \left( \frac{1}{\delta_n} \right)^{d/m_{\min}} \leq n \delta_n^2 / K^2, \quad (\text{C.60})$$

where  $\delta_n = \frac{1}{\rho}$ ,  $K \leq C_2 \frac{n^{\frac{d}{2(2m_{\min}+d)}}}{\rho}$ . By direct calculation, if  $2m_{\min}d + d^2 - 4m_{\min}^2 < 0$ , the condition (C.60) is satisfied when  $n > C$ , where  $C$  is a constant related to  $C_4$ . Otherwise we can divide  $[m_{\min}, m_{\max}]$  into  $p$  parts, and let  $m_{\min} = m_0 < m_1 < \dots < m_q = m_{\max}$  be this partition satisfying  $m_{i+1} - m_i \leq \frac{m_{\min}(4m_{\min}^2 - d^2)}{4m_{\min}d + 2d^2 - 8m_{\min}^2}$ . Let  $\mathcal{Q}'_i = \bigcup_{m \in [m_i, m_{i+1}]} \mathcal{Q}'_m$ . By applying Lemma E.2 into each  $\mathcal{Q}'_i$  and noting that the number of parts is finite, the entropy condition holds.

By Lemma E.2, we have

$$u^T \Psi_1^T \Psi_1 u = \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p u_j \varphi_j(X_i) \right)^2 = \|g\|_n^2 \geq \eta, \quad (\text{C.61})$$

with probability at least  $1 - C_5 \exp(-C_6 n^{\eta_1})$  for some constant  $\eta$  and  $\eta_1$ . Notice that

$$\det(\Psi_1^T \Psi_1) \geq \lambda_{\min}(\Psi_1^T \Psi_1)^p = \left( \min_u u^T \Psi_1^T \Psi_1 u \right)^p. \quad (\text{C.62})$$

Combining (C.61) and (C.62), we obtain

$$\det(\Psi_1^T \Psi_1) \geq \eta^p \quad (\text{C.63})$$

with probability at least  $1 - C_5 \exp(-C_6 n^{\eta_1})$ . By combining (C.63) with (C.59), the lower bound of  $\det(K_m + \mu_m I_n)$  is  $\mu_m^{n-p} n^p \prod \lambda_i^{(m)} \eta^p$ .

### C.13.3 Upper bound of the determinant term

In order to derive an upper bound, we need the following lemma.

**Lemma C.13.3.** *Let  $A, B \in \mathbb{R}^{n \times n}$  be two symmetric, positive definite matrices. Thus,*

$$\det(I_n + A + B) \leq \det(I_n + A) \det(I_n + B).$$

*Proof.* Let  $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_n > 0$  and  $\beta_1 \geq \beta_2 \geq \dots \geq \beta_n > 0$  be eigenvalues of matrices  $A$  and  $B$ , respectively. Therefore, we have

$$\begin{aligned} \det(I_n + A) \det(I_n + B) &= \prod_{i=1}^n (1 + \alpha_i)(1 + \beta_{n+1-i}) \\ &= \prod_{i=1}^n (1 + \alpha_i + \beta_{n+1-i} + \beta_{n+1-i}\alpha_i) \\ &\geq \prod_{i=1}^n (1 + \alpha_i + \beta_{n+1-i}) \\ &\geq \det(I_n + A + B), \end{aligned}$$

where the last inequality is true because of the Fiedler bound [112]. □

Let  $C_0$  denote the bound uniform bound of  $\Psi_m(\cdot, \cdot)$ . By Lemma C.13.3 and basic matrix calculation, it is true that

$$\begin{aligned} \det(K_m + \mu_m I_n) &\leq \mu_m^n \det(\mu_m^{-1} \Psi_1 \Lambda_1 \Psi_1^T + \mu_m^{-1} \Psi_2 \Lambda_2 \Psi_2^T + I_n) \\ &\leq \mu_m^n \det(\mu_m^{-1} \Psi_1 \Lambda_1 \Psi_1^T + I_n) \det(I_n + \mu_m^{-1} \Psi_2 \Lambda_2 \Psi_2^T) \\ &= \mu_m^{n-p} \det(\Lambda_1 \Psi_1^T \Psi_1 + \mu_m I_n) \det(I_n + \mu_m^{-1} \Psi_2 \Lambda_2 \Psi_2^T) \\ &\leq C^p \mu_m^{n-p} \det(\Lambda_1 \Psi_1^T \Psi_1) \det(I_n + \mu_m^{-1} \Psi_2 \Lambda_2 \Psi_2^T) \\ &\leq C^p \mu_m^{n-p} n^p \prod \lambda_i^{(m)} \det(\Psi_1^T \Psi_1) \det(I_n + \mu_m^{-1} \Psi_2 \Lambda_2 \Psi_2^T) \\ &\leq C^p \mu_m^{n-p} n^p \prod \lambda_i^{(m)} \left( \frac{\text{Tr}(\Psi_1^T \Psi_1)}{p} \right)^p \det(I_n + \mu_m^{-1} \Psi_2 \Lambda_2 \Psi_2^T) \\ &\leq C^p C_0^p \mu_m^{n-m} n^m \prod \lambda_i^{(m)} \det(I_n + \mu_m^{-1} \Psi_2 \Lambda_2 \Psi_2^T). \end{aligned}$$

Note that

$$\begin{aligned}
\det(I_n + \mu_m^{-1} \Psi_2 \Lambda_2 \Psi_2^T) &\leq \left( \frac{1}{n} \text{tr}(I_n + \mu_m^{-1} \Psi_2 \Lambda_2 \Psi_2^T) \right)^n \\
&\lesssim (1 + n^{\frac{-2m}{2m+d}})^n \\
&\leq 2^n.
\end{aligned}$$

Therefore, we have

$$\det(K_m + \mu_m I_n) \lesssim 2^n C^p C_0^p \mu_m^{n-p} n^p \prod \lambda_i^{(m)}.$$

Combine the lower bound and upper bound, we can conclude that  $\log \det(K_m + \mu_m I_n) = n \log n - (1 + o_p(1)) \frac{2mn}{2m+d} \log n$ .

## REFERENCES

- [1] G. Matheron, “Principles of geostatistics,” *Economic Geology*, vol. 58, no. 8, pp. 1246–1266, 1963.
- [2] N. Cressie, *Statistics for spatial data*. John Wiley & Sons, 2015.
- [3] P. J. Diggle, *Statistical analysis of spatial and spatio-temporal point patterns*. CRC Press, 2013.
- [4] K.-T. Fang, R. Li, and A. Sudjianto, *Design and modeling for computer experiments*. CRC Press, 2005.
- [5] T. J. Santner, B. J. Williams, and W. I. Notz, *The design and analysis of computer experiments*. Springer Science & Business Media, 2003.
- [6] C. Rasmussen and C. Williams, *Gaussian processes for machine learning*. MIT Press, 2006, p. 272.
- [7] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.
- [8] J. Staum, “Better simulation metamodeling: The why, what, and how of stochastic kriging,” in *Simulation Conference (WSC), Proceedings of the 2009 Winter*, IEEE, 2009, pp. 119–133.
- [9] B. Haaland and P. Z. Qian, “Accurate emulators for large-scale computer experiments,” *The Annals of Statistics*, vol. 39, no. 6, pp. 2974–3002, 2011.
- [10] B. Haaland, W. Wang, and V. Maheshwari, “A framework for controlling sources of inaccuracy in gaussian process emulation of deterministic computer experiments,” *To appear in SIAM/ASA Journal on Uncertainty Quantification*, *arXiv preprint arXiv:1411.7049*, 2018.
- [11] B. Ankenman, B. L. Nelson, and J. Staum, “Stochastic kriging for simulation metamodeling,” *Operations research*, vol. 58, no. 2, pp. 371–382, 2010.
- [12] S. Mak, C.-L. Sung, X. Wang, S.-T. Yeh, Y.-H. Chang, V. R. Joseph, V. Yang, and C. Wu, “An efficient surrogate model of large eddy simulations for design evaluation and physics extraction,” *ArXiv preprint arXiv:1611.07911*, 2016.

- [13] A. N. Burchell, H. Richardson, S. M. Mahmud, H. Trottier, P. P. Tellier, J. Hanley, F. Coutlée, and E. L. Franco, “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*, vol. 163, no. 6, pp. 534–543, 2006.
- [14] A. E. Moran, M. C. Odden, A. Thanataveerat, K. Y. Tzong, P. W. Rasmussen, D. Guzman, L. Williams, K. Bibbins-Domingo, P. G. Coxson, and L. Goldman, “Cost-effectiveness of hypertension therapy according to 2014 guidelines,” *New England Journal of Medicine*, vol. 372, no. 5, pp. 447–455, 2015.
- [15] X. Chen and Q. Zhou, “Sequential experimental designs for stochastic kriging,” in *Proceedings of the 2014 Winter Simulation Conference*, IEEE Press, 2014, pp. 3821–3832.
- [16] M. Binois, R. B. Gramacy, and M. Ludkovski, “Practical heteroskedastic gaussian process modeling for large simulation experiments,” *ArXiv preprint arXiv:1611.05902*, 2016.
- [17] D. A. Harville, *Matrix algebra from a statistician’s perspective*. Springer, 1997, vol. 1.
- [18] B. Schölkopf, R. Herbrich, and A. J. Smola, “A generalized representer theorem,” in *International Conference on Computational Learning Theory*, Springer, 2001, pp. 416–426.
- [19] J. Shao, *Mathematical statistics*. Springer, 1999.
- [20] F. Aurenhammer, “Voronoi diagrams a survey of a fundamental geometric data structure,” *ACM Computing Surveys (CSUR)*, vol. 23, no. 3, pp. 345–405, 1991.
- [21] P. C. Mahalanobis, “On the generalized distance in statistics,” *Proceedings of the National Institute of Sciences (Calcutta)*, vol. 2, pp. 49–55, 1936.
- [22] S. Ba and V. R. Joseph, “Composite gaussian process models for emulating expensive functions,” *The Annals of Applied Statistics*, pp. 1838–1860, 2012.
- [23] J. Nocedal and S. Wright, *Numerical optimization, series in operations research and financial engineering*. Springer, New York, USA, 2006, 2006.
- [24] B. C. Eaves, “Homotopies for computation of fixed points,” *Mathematical Programming*, vol. 3, no. 1, pp. 1–22, 1972.
- [25] R. S. Varga, *Gershgorin and his circles*. Springer Science & Business Media, 2010, vol. 36.

- [26] M. L. Stein, *Interpolation of spatial data: Some theory for kriging*. Springer Science & Business Media, 1999.
- [27] R. Stocki, “A method to improve design reliability using optimal latin hypercube sampling,” *Computer Assisted Mechanics and Engineering Sciences*, vol. 12, no. 4, p. 393, 2005.
- [28] M. Stein, “Large sample properties of simulations using latin hypercube sampling,” *Technometrics*, vol. 29, no. 2, pp. 143–151, 1987.
- [29] V. R. Joseph, E. Gul, and S. Ba, “Maximum projection designs for computer experiments,” *Biometrika*, vol. 102, no. 2, pp. 371–380, 2015.
- [30] R. Carnell, M. R. Carnell, and S. RUnit, “Package lhs,” *CRAN*. <https://cran.rproject.org/web/packages/lhs/lhs.pdf>, 2016.
- [31] S Ba and V. Joseph, “Maxpro: Maximum projection designs,” *R package version*, pp. 3–1, 2015.
- [32] R. L. Iman and W.-J. Conover, “A distribution-free approach to inducing rank correlation among input variables,” *Communications in Statistics-Simulation and Computation*, vol. 11, no. 3, pp. 311–334, 1982.
- [33] B. Tang, “Orthogonal array-based latin hypercubes,” *Journal of the American statistical association*, vol. 88, no. 424, pp. 1392–1397, 1993.
- [34] A. B. Owen, “Controlling correlations in latin hypercube samples,” *Journal of the American Statistical Association*, vol. 89, no. 428, pp. 1517–1522, 1994.
- [35] J.-S. Park, “Optimal latin-hypercube designs for computer experiments,” *Journal of statistical planning and inference*, vol. 39, no. 1, pp. 95–111, 1994.
- [36] M. D. Morris and T. J. Mitchell, “Exploratory designs for computational experiments,” *Journal of statistical planning and inference*, vol. 43, no. 3, pp. 381–402, 1995.
- [37] B. Tang, “Selecting latin hypercubes using correlation criteria,” *Statistica Sinica*, pp. 965–977, 1998.
- [38] K. Q. Ye, “Orthogonal column latin hypercubes and their application in computer experiments,” *Journal of the American Statistical Association*, vol. 93, no. 444, pp. 1430–1439, 1998.

- [39] K. Q. Ye, W. Li, and A. Sudjianto, “Algorithmic construction of optimal symmetric latin hypercube designs,” *Journal of statistical planning and inference*, vol. 90, no. 1, pp. 145–159, 2000.
- [40] R. Jin, W. Chen, and A. Sudjianto, “An efficient algorithm for constructing optimal design of computer experiments,” *Journal of Statistical Planning and Inference*, vol. 134, no. 1, pp. 268–287, 2005.
- [41] X. Xu, B. Haaland, and P. Z. Qian, “Sudoku-based space-filling designs,” *Biometrika*, vol. 98, no. 3, pp. 711–720, 2011.
- [42] J. Chen and P. Z. Qian, “Latin hypercube designs with controlled correlations and multi-dimensional stratification,” *Biometrika*, vol. 101, no. 2, pp. 319–332, 2014.
- [43] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, “Design and analysis of computer experiments,” *Statistical science*, pp. 409–423, 1989.
- [44] J. Fan and I. Gijbels, *Local polynomial modeling and its applications: Monographs on statistics and applied probability*. CRC Press, 1996, vol. 66.
- [45] G. Wahba, *Spline models for observational data*. SIAM, 1990.
- [46] H. Wendland, *Scattered data approximation*. Cambridge university press, 2004, vol. 17.
- [47] A. W. van der Vaart and J. A. Wellner, *Weak convergence and empirical processes*. Springer, 1996.
- [48] R. J. Adler and J. E. Taylor, *Random fields and geometry*. Springer Science & Business Media, 2009.
- [49] H. Cramér and M. R. Leadbetter, *Stationary and related stochastic processes: Sample function properties and their applications*. Courier Corporation, 1967.
- [50] C. F. J. Wu and M. S. Hamada, *Experiments: Planning, analysis, and optimization*. John Wiley & Sons, 2011, vol. 552.
- [51] S. Banerjee, B. P. Carlin, and A. E. Gelfand, *Hierarchical modeling and analysis for spatial data*. Crc Press, 2014.
- [52] A. D. Bull, “Convergence rates of efficient global optimization algorithms,” *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2879–2904, 2011.

- [53] R. Tuo and C. F. J. Wu, “A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties,” *SIAM/ASA Journal on Uncertainty Quantification*, vol. 4, no. 1, pp. 767–795, 2016.
- [54] ———, “Prediction based on the Kennedy-O’Hagan calibration model: Asymptotic consistency and other properties,” *Statistica Sinica*, to appear, 2017.
- [55] S. Yakowitz and F Szidarovszky, “A comparison of kriging with nonparametric regression methods,” *Journal of Multivariate Analysis*, vol. 16, no. 1, pp. 21–53, 1985.
- [56] M. L. Stein, “Asymptotically efficient prediction of a random field with a misspecified covariance function,” *The Annals of Statistics*, vol. 16, no. 1, pp. 55–63, 1988.
- [57] ———, “Bounds on the efficiency of linear predictions using an incorrect covariance function,” *The Annals of Statistics*, pp. 1116–1138, 1990.
- [58] ———, “Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure,” *The Annals of Statistics*, pp. 850–872, 1990.
- [59] ———, “Efficiency of linear predictors for periodic processes using an incorrect covariance function,” *Journal of Statistical Planning and Inference*, vol. 58, no. 2, pp. 321–331, 1997.
- [60] M. E. Johnson, L. M. Moore, and D. Ylvisaker, “Minimax and maximin distance designs,” *Journal of Statistical Planning and Inference*, vol. 26, no. 2, pp. 131–148, 1990.
- [61] Z. Wu and R. Schaback, “Local error estimates for radial basis function interpolation of scattered data,” *IMA Journal of Numerical Analysis*, vol. 13, no. 1, pp. 13–27, 1993.
- [62] A. W. van der Vaart, J. H. van Zanten, *et al.*, “Reproducing kernel hilbert spaces of gaussian priors,” in *Pushing the limits of contemporary statistics: Contributions in honor of Jayanta K. Ghosh*, Institute of Mathematical Statistics, 2008, pp. 200–222.
- [63] R. Tuo and C. J. Wu, “Efficient calibration for imperfect computer models,” *The Annals of Statistics*, vol. 43, no. 6, pp. 2331–2352, 2015.
- [64] D. Xiu, *Numerical methods for stochastic computations: A spectral method approach*. Princeton University Press, 2010.
- [65] R. Tuo, C. J. Wu, and D. Yu, “Surrogate modeling of computer experiments with different mesh densities,” *Technometrics*, vol. 56, no. 3, pp. 372–380, 2014.

- [66] X. He, R. Tuo, and C. J. Wu, “Optimization of multi-fidelity computer experiments via the eqie criterion,” *Technometrics*, vol. 59, no. 1, pp. 58–68, 2017.
- [67] N. Zhang and D. W. Apley, “Fractional brownian fields for response surface meta-modeling,” *Journal of Quality Technology*, vol. 46, no. 4, pp. 285–301, 2014.
- [68] M. Plumlee and D. W. Apley, “Lifted Brownian kriging models,” *Technometrics*, vol. 59, no. 2, pp. 165–177, 2017.
- [69] R. B. Gramacy and H. K. Lee, “Cases for the nugget in modeling computer experiments,” *Statistics and Computing*, vol. 22, no. 3, pp. 713–722, 2012.
- [70] C.-Y. Peng and C. J. Wu, “On the choice of nugget in kriging modeling for deterministic computer experiments,” *Journal of Computational and Graphical Statistics*, vol. 23, no. 1, pp. 151–168, 2014.
- [71] M. L. Stein, “A simple condition for asymptotic optimality of linear predictions of random fields,” *Statistics & Probability Letters*, vol. 17, no. 5, pp. 399–404, 1993.
- [72] C. Saunders, A. Gammerman, and V. Vovk, “Ridge regression learning algorithm in dual variables.” in *ICML*, vol. 98, 1998, pp. 515–521.
- [73] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1.
- [74] F. Douak, F. Melgani, and N. Benoudjit, “Kernel ridge regression with active learning for wind speed prediction,” *Applied energy*, vol. 103, pp. 328–340, 2013.
- [75] G. de los Campos, D. Gianola, and G. J. Rosa, “Reproducing kernel hilbert spaces regression: A general framework for genetic evaluation,” *Journal of Animal Science*, vol. 87, no. 6, pp. 1883–1887, 2009.
- [76] C. J. Stone, “Optimal global rates of convergence for nonparametric regression,” *The annals of statistics*, pp. 1040–1053, 1982.
- [77] W.-L. Loh *et al.*, “Estimating the smoothness of a gaussian random field from irregularly spaced data via higher-order quadratic variations,” *The Annals of Statistics*, vol. 43, no. 6, pp. 2766–2794, 2015.
- [78] T. R. Thomas, “Rough surfaces,” *Longman Group*, vol. 153, 1982.
- [79] A. Constantine and P. Hall, “Characterizing surface smoothness via estimation of effective fractal dimension,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 97–113, 1994.

- [80] P. Hall and A. Wood, “On the performance of box-counting estimators of fractal dimension,” *Biometrika*, pp. 246–252, 1993.
- [81] J. Istas and G. Lang, “Quadratic variations and estimation of the local hölder index of a gaussian process,” in *Annales de l’Institut Henri Poincare (B) Probability and Statistics*, Elsevier, vol. 33, 1997, pp. 407–436.
- [82] J. T. Kent and A. T. Wood, “Estimating the fractal dimension of a locally self-similar gaussian process by using increments,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 679–699, 1997.
- [83] A. W. van der Vaart and J. H. van Zanten, “Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth,” *The Annals of Statistics*, pp. 2655–2675, 2009.
- [84] C. J. Stone, “An asymptotically optimal window selection rule for kernel density estimates,” *The Annals of Statistics*, pp. 1285–1297, 1984.
- [85] D. L. Donoho and I. M. Johnstone, “Adapting to unknown smoothness via wavelet shrinkage,” *Journal of the american statistical association*, vol. 90, no. 432, pp. 1200–1224, 1995.
- [86] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, “Wavelet shrinkage: Asymptopia?” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 301–369, 1995.
- [87] R. L. Eubank, *Nonparametric regression and spline smoothing*. CRC press, 1999.
- [88] C. Gu, *Smoothing spline anova models*. Springer, 2013.
- [89] M. M. Fischer and Y. Leung, *Geocomputational modelling: Techniques and applications*. Springer Science & Business Media, 2013.
- [90] T. M. Mitchell, “Machine learning and data mining,” *Communications of the ACM*, vol. 42, no. 11, pp. 30–36, 1999.
- [91] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2012.
- [92] S. A. van de Geer, *Empirical processes in m-estimation*. Cambridge university press, 2000, vol. 6.
- [93] E. M. Stein and R. Shakarchi, *Real analysis: Measure theory, integration, and hilbert spaces*. Princeton University Press, 2005.

- [94] R. A. Adams and J. J. Fournier, *Sobolev spaces*. Academic press, 2003, vol. 140.
- [95] R. Tuo and J. C. F. Wu, “A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties,” *SIAM/ASA Journal on Uncertainty Quantification*, vol. 4, no. 1, pp. 767–795, 2016.
- [96] R. A. DeVore and R. C. Sharpley, “Besov spaces on domains in,” *Transactions of the American Mathematical Society*, vol. 335, no. 2, pp. 843–864, 1993.
- [97] P. Serra, T. Krivobokova, *et al.*, “Adaptive empirical bayesian smoothing splines,” *Bayesian Analysis*, vol. 12, no. 1, pp. 219–238, 2017.
- [98] G. Chan and A. T. Wood, “Increment-based estimators of fractal dimension for two-dimensional surface data,” *Statistica Sinica*, pp. 343–376, 2000.
- [99] M. Jirak, A. Meister, M. Reiß, *et al.*, “Adaptive function estimation in nonparametric regression with one-sided errors,” *The Annals of Statistics*, vol. 42, no. 5, pp. 1970–2002, 2014.
- [100] A Goldenshluger and A Nemirovski, “On spatially adaptive estimation of nonparametric regression,” *Mathematical methods of Statistics*, vol. 6, no. 2, pp. 135–170, 1997.
- [101] E. Belitser, P. Serra, *et al.*, “Adaptive priors based on splines with random knots,” *Bayesian Analysis*, vol. 9, no. 4, pp. 859–882, 2014.
- [102] R De Jonge, J. Van Zanten, *et al.*, “Adaptive estimation of multivariate functions using conditionally gaussian tensor-product spline priors,” *Electronic Journal of Statistics*, vol. 6, pp. 1984–2001, 2012.
- [103] I. C. Ipsen and B. Nadler, “Refined perturbation bounds for eigenvalues of hermitian and non-hermitian matrices,” *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 1, pp. 40–53, 2009.
- [104] G. Tripathi, “A matrix extension of the cauchy-schwarz inequality,” *Economics Letters*, vol. 63, no. 1, pp. 1–3, 1999.
- [105] G. H. Golub and C. F. Van Loan, *Matrix computations (3rd ed.)* Baltimore, MD, USA: Johns Hopkins University Press, 1996, ISBN: 0-8018-5414-8.
- [106] F. Girosi, M. Jones, and T. Poggio, “Regularization theory and neural networks architectures,” *Neural Computation*, vol. 7, no. 2, pp. 219–269, 1995.

- [107] S. van de Geer *et al.*, “On the uniform convergence of empirical norms and inner products, with application to causal inference,” *Electronic Journal of Statistics*, vol. 8, no. 1, pp. 543–574, 2014.
- [108] P. Massart, *Concentration inequalities and model selection*. Springer, 2007, vol. 6.
- [109] D. E. Edmunds and W. D. Evans, *Spectral theory and differential operators*. Clarendon Press Oxford, 1987, vol. 15.
- [110] D. E. Edmunds and H. Triebel, *Function spaces, entropy numbers, differential operators*. Cambridge University Press, 2008, vol. 120.
- [111] M. Marcus and H. Minc, *A survey of matrix theory and matrix inequalities*. Courier Corporation, 1992, vol. 14.
- [112] M. Fiedler, “Bounds for the determinant of the sum of hermitian matrices,” *Proceedings of the American Mathematical Society*, pp. 27–31, 1971.