

Posterior Consistency in Nonparametric Regression Problems under Gaussian Process Priors

Taeryon Choi

July 2005

Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy*

Thesis Committee:

Professor Mark J. Schervish, Chair¹

Professor Larry Wasserman¹

Professor Anthony Brockwell¹

Professor Subhashis Ghosal²

©2005 Taeryon Choi

¹Department of Statistics, Carnegie Mellon University

²Department of Statistics, North Carolina State University

UMI Number: 3190067

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI[®]

UMI Microform 3190067

Copyright 2006 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

CarnegieMellon

COLLEGE OF HUMANITIES & SOCIAL SCIENCES

DISSERTATION

Submitted in Partial Fulfillment of the Requirements


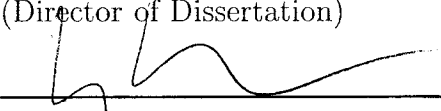
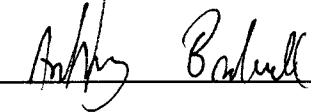
For the Degree of **DOCTOR OF PHILOSOPHY**

Title:

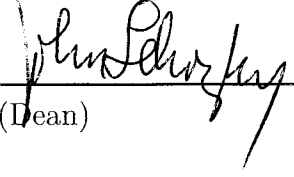
**“POSTERIOR CONSISTENCY IN
NONPARAMETRIC
REGRESSION PROBLEMS
UNDER GAUSSIAN PROCESS
PRIORS”**

Presented by: **TAERYON CHOI**

Accepted by the DEPARTMENT OF STATISTICS

Readers:		<u>25 July 2005</u>
	(Director of Dissertation)	Date
		<u>7/25/05</u>
		(Date)
		<u>7/25/05</u>
		(Date)
	<u>Subhashis Ghosal</u>	<u>07-25-05</u>
		(Date)

Approved by the Committee on Graduate Degrees

	<u>8/15/05</u>
(Dean)	(Date)

Abstract

In the Bayesian approach to statistical inference, the posterior distribution summarizes information regarding unknown parameters after observing data. As the sample size increases, we expect that the posterior distribution will concentrate around the true value of the parameter. This is known as posterior consistency. Posterior consistency is sometimes used as a theoretical justification of the Bayesian method.

In this dissertation, we establish posterior consistency in nonparametric regression problems using Gaussian process priors, known as Gaussian process regression. Gaussian process regression is one of the most popular Bayesian nonparametric regression approaches, and it consists of nonparametrically modeling the unknown regression function as a Gaussian process a priori.

We provide an extension of the posterior consistency theorem of Schwartz (1965) for independent but non-identically distributed observations. Then we apply our theorem to the Gaussian process regression problem by giving conditions on the Gaussian process prior that allow us to verify the conditions of the consistency theorem, namely the prior positivity of neighborhoods of the true parameter and the existence of tests with suitable properties.

The specific nonparametric regression problem that we consider assumes that the data have normal or Laplace (double exponential) noise distributions. The error variance is also assumed to be unknown and needs to be estimated. We consider three different metric topologies on the space of regression functions, and then define joint neighborhoods of

the regression function and the noise variance. The joint neighborhoods are based on the L^1 metric, an in-probability metric, and the Hellinger metric. For each of these joint neighborhoods, we prove almost sure consistency of the posterior distribution by showing that the posterior probability of these joint neighborhoods converges almost surely to 1 under appropriate conditions on the regression function and covariates. Specifically, when the covariate values are fixed in advance, we prove almost sure consistency based on the L^1 metric and the in probability metric assuming the covariate values are spread out sufficiently well. When the covariate values are sampled from a probability distribution, we prove almost sure consistency based on the in-probability metric and the Hellinger metric. Furthermore, in the random covariate case, under an additional uniform boundedness condition, we prove almost sure consistency with the L^1 metric as well.

Finally, we also study how much posterior consistency can be extended into other regression model structures, which deal with hyperparameters of a covariance function and multidimensional covariates.

Acknowledgements

First and foremost, I would like to thank Mark J. Schervish, my advisor, for his many suggestions and constant support during this research. I am also deeply thankful to him for his guidance through the early years of chaos and confusion as a foreign student from a non-English based culture. He was my advisor during all my graduate years, beginning as a temporary academic advisor in my first year at Carnegie Mellon University, then as an ADA advisor, and as a thesis advisor. He always infused me with profound guidance and insights in every discussion, and he kindly spared me his time whenever I knocked on his office door. I cannot imagine a more ideal advisor than he, who has been passionately committed and engaged to this student.

I also would like to thank my other committee members, Larry Wasserman, Anthony Brockwell and Subhashis Ghosal. Larry Wasserman expressed interest in my work and supplied me with helpful discussions and insightful perspectives on my results. Anthony Brockwell has also provided me with helpful discussions, endless encouragement and warmth. Subhashis Ghosal served as an external advisor, whose earlier work was an indispensable pedestal of my research work and who gave critical feedback and helpful suggestions.

I had the pleasure of being in a department full of kind people and would like to express my appreciation to the rest of faculty, staff and students in the department for their warm encouragement and friendly help through my graduate studies. Special thanks go to Woncheol Jang, a former graduate student in the department as well as an fellow alumnus

from the same university and department in Korea. He helped me in countless ways.

I am grateful to my parents and parents-in-law. They provided me and my wife with unconditional love and ceaseless support, for which I can never repay them.

Last but not least, I cannot thank my lovely wife, Jeongeun, enough. I owe to her what I am now and I cannot imagine a life without her. In addition, I am proud of fatherhood and dedicate this dissertation to my growing family.

Contents

1	Introduction	1
1.1	Problem Definition	1
1.2	Bayesian Inference and Posterior Consistency	2
1.3	Bayesian Nonparametric Regression Function Estimation	5
1.3.1	Orthogonal Basis Expansion	7
1.3.2	Smoothing Splines	9
1.3.3	Gaussian Process Method	10
1.4	Thesis Outline	11
2	Gaussian Process Regression	13
2.1	Introduction	13
2.2	Nonparametric Bayesian Regression and Gaussian Process	14
2.3	Aspects of Gaussian Process Regression	16
2.3.1	Covariance Function, $R(\cdot, \cdot)$	16
2.3.2	GP Regression Implementation	19
2.3.3	Existing Consistency Results	20
3	Posterior Consistency of GP Regression	23
3.1	Introduction	23
3.2	Topologies on the Set of Regression Functions	25
3.3	Consistency Theorems for Non i.i.d. Observations	30

3.4	Consistency in Nonparametric Regression	36
3.4.1	Main Theorems	40
3.5	Verification	42
3.5.1	Prior Positivity Condition	43
3.5.2	Construction of Sieves	46
3.5.3	Existence of Tests	52
3.5.4	Proofs of Main Theorems	70
3.5.5	Illustrations of Covariance Functions	73
4	Extension of Posterior Consistency Results	77
4.1	Hyperparameters in Covariance Function	78
4.2	Multi-dimensional Covariates	82
5	Conclusions and Future work	92
5.1	Summary and Contributions	92
5.2	Future Work	94
	Appendix	97

List of Figures

3.1	A comparison between $L^1(Q)$ metric and d_Q metric	29
3.2	Graphical display of Theorem 3.3.1	32

List of Tables

2.1	Examples of isotropic covariance functions	18
3.1	Summary of Consistency results	41

Chapter 1

Introduction

1.1 Problem Definition

This thesis establishes posterior consistency in nonparametric regression problems using Gaussian process (GP) priors, known as GP regression. Nonparametric regression problems refer to statistical methods to estimate the regression function with only a few underlying assumptions. GP regression is one of the most popular approaches to nonparametric Bayesian regression problems, and this approach is to model the regression function as a Gaussian process a priori. In other words, GP regression uses a Gaussian process as a nonparametric prior distribution on an unknown regression function.

GP regression has been investigated in a variety of ways both in statistics and machine learning communities (see Paciorek (2003) or Seeger (2004)). So far, much effort has been given to the methodological development of GP regression. More efficient implementation tools for GP regression using various simulation-based techniques have been suggested and new covariance functions have been defined to handle complicated situations. In addition, GP regression methods have been applied to real data using these covariance functions (Paciorek, 2003).

Compared to these expansive methodological developments and successful empirical results of GP regression, theoretical justification of this method has not been investigated

fully in the literature. Theoretical validation of all Bayes procedures are based on the posterior distribution, and it is expected that the more data that is available, the closer the posterior distribution of each Bayes procedure will be to the true parameter, formalized as the concept of posterior consistency. Our thesis provides a theoretical underpinning of the GP regression model through an asymptotic analysis of the method by establishing posterior consistency of the method.

In this thesis, we study posterior consistency of GP regression problems with various model settings, assuming two conditions, one about the covariates X 's in the regression model and the other about smoothness condition of the regression function $\eta(x)$. Specifically, we establish that posterior distribution of the regression function is almost surely consistent under appropriate topologies of the regression functions.

1.2 Bayesian Inference and Posterior Consistency

The problem of statistical inference is to draw meaningful conclusions about unknown parameters. The parameter, θ is unobservable, and thus the inference must be based on the data or observations. Bayesian inference treats all unknown parameters as random variables and uses conditional probability to express all forms of uncertainty. A Bayesian would consider a probability distribution of the unobservable, θ , which reflects prior opinion or belief about the unknown parameter, θ , and is called prior distribution. Before the data are observed, this prior distribution quantifies our uncertainty about the unobservable. After we observe data, our opinion or belief is updated through a conditional distribution of the parameter given data, and this conditional distribution is called a posterior distribution.

Consequently, in the Bayesian approach to statistical inference, posterior distribution summarizes information regarding unknown parameters, combined with likelihood or probability model and prior distribution. As sample sizes increase or we observe more and more

data, we expect that the posterior distribution will concentrate around the true distribution of parameter, summarized as the concept of posterior consistency.

Roughly speaking, posterior consistency means that the posterior concentrates around the true distribution. A systematic and mathematical definition of posterior consistency can be found in several references such as Ghosh and Ramamoorthi (2003) and Choudhuri et al. (2004) as follows:

Let $\{(\mathcal{X}^{(n)}, \mathcal{A}^{(n)}, P_{\theta}^{(n)} : \theta \in \Theta)\}$ be a sequence of statistical experiments with observations $X^{(n)}$, where the parameter set Θ is an arbitrary topological space and n is a sample size and $X^{(n)}$ is a sequence of X_1, \dots, X_n . Let \mathcal{B} be the Borel σ -field on Θ and Π_n be a probability measure on (Θ, \mathcal{B}) . The posterior distribution is defined to be a version of the regular conditional probability of θ given $X^{(n)}$, and is denoted by $\Pi_n(\cdot | X^{(n)})$. Let $\theta_0 \in \Theta$ be the true parameter. We say that the posterior distribution is consistent at θ_0 (with respect to the given topology on Θ) if for every neighborhood U of θ_0 , $\Pi_n(U^C | X^{(n)}) \rightarrow 0$ in probability under $P_{\theta_0}^{(n)}$ or almost surely with respect to P_{θ_0} , where P_{θ_0} is a probability measure induced by the true parameter value θ_0 .

Posterior consistency can be thought of as a theoretical justification of the Bayesian method. In addition, posterior consistency can be used to define a true model as in Walker and Damien (2000), particularly when we consider a nonparametric prior for infinite dimensional parameter space.

When the parameter space Θ is finite dimensional, posterior consistency can be achieved easily under fairly general conditions. When the problem involves infinite-dimensional parameters, however, the consistency of posterior distributions is a much more challenging problem. There have been many results giving general conditions under which features of posterior distributions are consistent in infinite-dimensional spaces. For examples, see Doob (1949), Schwartz (1965), Barron et al. (1999), Ghosal et al. (1999), Amewou-Atisso et al. (2003), Walker (2003), Choudhuri et al. (2004), Ghosal and Roy (2005) and Walker (2004).

Doob (1949) established an early and fundamental consistency result under weak condition using the martingale approach. Roughly speaking, Doob's theorem means that if there exists a consistent estimator of Θ , then the posterior distribution of Θ will tend to concentrate near the true value of Θ , with probability 1 under the joint distribution of the data and parameter. Doob's theorem guarantees that the posterior will be consistent for all θ except on a null set, a set of probability measure zero. If the prior misses the true value or the null set is huge, then the conclusion to the theorem is not very attractive. Schwartz (1965) obtained an important and general result on consistency. Schwartz' theorem is crucial to study the behavior of posterior distribution in the infinite dimensional parameter space and will be also investigated further for our theoretical results.

Schwartz (1965) proved a theorem that gave conditions for consistency of posterior distributions of parameters from the distributions of independent and identically distributed (*i.i.d.*) random variables. These conditions include the existence of tests with sufficiently small error rates and the prior positivity of certain neighborhoods of true value of parameter.

Barron et al. (1999) provided another sufficient conditions to guarantee the posterior consistency based on the Hellinger neighborhood of the true distribution. The conditions are a support condition which requires prior positivity around the true parameter and a smoothness condition based on entropy-type bound on the roughness of densities.

Ghosal et al. (1999) showed consistency under weaker condition based on metric entropy and reaffirmed the usefulness of the approach based on testing.

As discussed in previous literature reviews, early results on posterior consistency in the infinite dimensional parameter space have focused mainly on density estimation, that is on estimating a density function for a random sample without assuming that the density belongs to a finite-dimensional parametric family. More recently, attention has turned to posterior consistency in nonparametric and semiparametric regression problems. Since the celebrated Schwartz' theorem (Schwartz (1965)) was originally designed for *i.i.d.* random

variables, its variants have been studied under a more general model setting, and they include Amewou-Atisso et al. (2003) and Choudhuri et al. (2004). We provide another extension of Schwartz's theorem to almost sure convergence, tailored into GP regression in Chapter 3 and our extension is based on both Amewou-Atisso et al. (2003) and Choudhuri et al. (2004).

We apply our extension to the GP regression problem and prove our main consistency result in Chapter 3. While proving this, we identify two conditions on the design point for the regression model and the smoothness of the Gaussian process, required for our extension. Two conditions are described in Chapter 3.

1.3 Bayesian Nonparametric Regression Function Estimation

Estimating unknown functions is one of the most important and widely studied topics in modern statistical analysis. Often, this area is called nonparametric inference since we do not have much idea of the unknown functions or unknown quantities. Thus, the underlying idea of nonparametric inference is to make use of data with only a few assumptions about the unknown quantities. Most cases, nonparametric inference or nonparametric statistics deals with a very large index set or parameter space, even infinite dimensional parameter space such as a set of functions which are continuous. There exist diverse statistical approaches to tackle this complicated task, and each different approach has its own pros and cons. One good reference to survey this area is Wasserman (2005).

Nonparametric function estimation includes density estimation, distribution function estimation, regression function estimation, survival function estimation, and others. These topics are concerned mainly with nonparametric inference and still are being investigated actively. Among them, nonparametric regression function estimation is the main topic of the nonparametric statistical methods. Several techniques either from frequentist or from Bayesian perspectives have been proposed and developed. Kernel smoothing, spline

smoothing, orthogonal series estimation and wavelet smoothing belong to a set of classical frequentist methods and they have been well explored. Bayesian version of their counterparts and other novel Bayesian approaches are also have been studied but require further examination in both aspects of theory and method.

We give a brief survey of nonparametric regression problems from a Bayesian perspective. The nonparametric regression model that we consider, is

$$Y_i = \eta(X_i) + \epsilon_i, \quad i = 1, \dots, n, \quad (1.1)$$

where ϵ_i 's are assumed to be i.i.d. with a normal distribution whose mean is zero and the unknown variance is σ^2 . The goal here is to estimate the unknown regression function $\eta(x)$ with as few assumptions as possible, such as $\eta(x)$ is continuously differentiable, and the unknown variance σ^2 . The unknown regression function $\eta(x)$ is viewed as conditional expectation of Y given $X = x$, namely $\eta(x) = E(Y|X = x)$.

Nonparametric Bayesian regression method begins with the prior specification on $\eta(x)$. In other words, Bayesians need to express uncertainty about a function or infinite dimensional parameters. Therefore, Bayesians treat $\eta(x)$ as a random function and choose a prior distribution for $\eta(x)$. Several ways of putting a prior for $\eta(x)$ have been proposed and primary nonparametric Bayesian methods that we will consider include orthogonal basis expansion, spline smoothing and Gaussian process method. We will introduce each method briefly and review their methodological development and theoretical results according to the consistency. Later in this thesis, we will focus on a specific Bayesian approach for nonparametric regression function estimation using Gaussian process, *Gaussian process regression*. We will review that approach in Chapter 2 in detail and study its consistency thoroughly in Chapter 3 and 4.

1.3.1 Orthogonal Basis Expansion

Generally speaking, functions or regression functions can be represented as an infinite series of basis functions under regularity conditions. An orthogonal basis expansion for a regression function $\eta(x)$ is a representation of $\eta(x)$ by an infinite sum as $\eta(x) = \sum_{i=1}^{\infty} \theta_i \varphi_i(x)$ where, $\{\varphi_i(x)\}_{i=1}^{\infty}$ is the orthonormal basis for an L_2 space. For example, $\{\varphi_i\}_{i=1}^{\infty}$ can be a known basis such as Fourier series and then $\{\theta_i\}_{i=1}^{\infty}$ becomes the unknown Fourier coefficients. Then, the goal in this approach is to estimate the unknown Fourier coefficients. In practice, an infinite number of coefficients cannot always be estimated and the number of terms, k with $\eta(x) \approx \sum_{i=1}^k \theta_i \varphi_i(x)$ needs to be estimated or chosen.

In Bayesian analysis, θ_i 's are treated as random variables and a prior distribution for $\eta(x)$ is constructed by specifying a prior for θ_i 's. The prior for θ_i is usually given as a normal prior with a suitably chosen variance τ_i^2 , $\theta_i \sim N(0, \tau_i^2)$ and in some cases, the number of terms, k is also treated as a random quantity that needs to be estimated, and the prior for k is specified. The orthogonal basis approach has been explored in several articles, dealing with asymptotic properties of this method. For instance, Zhao (2000) and Shen and Wasserman (2001) considered this problem using prior distributions, constructed from a sequence of finite dimensional distributions and called "sieve priors".

An advantage of an orthogonal basis expansion is that it converts nonparametric regression, (1.1) into a type of problem known as the *infinitely many normal mean problem*, which turns out to be theoretically easy to handle. In this regard, asymptotic properties of these expansions have been studied by re-expressing the regression model as a problem of estimating the infinitely many parameters $\{\theta_i\}_{i=1}^{\infty}$. This approach is called the *infinitely many normal means problem* and it has been studied extensively by Cox (1993), Freedman (1999), Zhao (2000) and Shen and Wasserman (2001). Briefly, the parameter is $\eta = (\eta_1, \eta_2, \dots)$ and the data are $Y = (Y_1, Y_2, \dots)$ where $Y_j = \eta_j + n^{-1/2} \epsilon_j$, the ϵ_j 's are i.i.d. $N(0,1)$ and n is the sample size as in (1.1). Under this setting, one models the η_j 's as independent

normal random variables with mean 0 and variance τ_j^2 . Freedman (1999) studied the non-linear functional $\|\eta - \hat{\eta}\|^2$, where $\hat{\eta}$ is the Bayes estimator, both from the Bayesian and the frequentist perspectives. His main results imply consistency of the Bayes estimator for all $\{\eta_j\}_{j=1}^\infty \in \ell_2$ if $\sum_{j=1}^\infty \tau_j^2 < \infty$. Zhao (2000) showed a similar consistency result and that the Bayes estimator attains the minimax rate of convergence for certain class of priors. Shen and Wasserman (2001) investigated asymptotic properties of the posterior distribution and obtained convergence rates.

Although Brown and Low (1996) exploited an asymptotic equivalence between the infinitely many normal means problem and nonparametric regression, it is still not clear that the asymptotic properties of the posteriors from one problem correspond directly to those of the other problem. Indeed, Shen and Wasserman (2001) considered both examples from the orthogonal basis expansion approach to nonparametric regression and the infinite normal mean problem, and established the general results of the rate of convergence of posterior distributions and applied the general results to two examples. They also studied the convergence rates for various cases including these two examples and compared their results with those of Zhao (2000).

When the smoothness parameter or “smoothness conditions” of a density or regression function is unknown, an approach known as *adaptive estimation* is used to estimate the regression function. Belitser and Ghosal (2003) considered the general convergence rate of adaptive estimation under the infinite normal mean setting, and Huang (2004) considered convergence rates of posterior distributions of density estimation and regression using sieve-based priors in the adaptive estimation. Huang (2004) dealt with a Bayesian convergence rate theorem in the context of regression, assuming the error, ϵ_i is normally distributed with known variance σ^2 and the true regression function is bounded by a known constant.

1.3.2 Smoothing Splines

Smoothing splines are powerful and flexible tools for nonparametric curve estimation, and they provide computationally efficient methods. Spline smoothing can be motivated by considering the penalized sum of squares. In many cases, the regression function estimator, $\hat{\eta}_n(x)$, to estimate $\eta(x)$, is chosen to minimize the residual sum of squares, $\sum_{i=1}^n (Y_i - \hat{\eta}_n(x_i))^2$. This approach can have some limitations because it merely interpolates the data without considering any data structure. Spline smoothing provides an alternative way to solve this limitation by adding a penalty term as $\sum_{i=1}^n (Y_i - \hat{\eta}_n(x_i))^2 + \lambda J(\eta)$, where the roughness penalty term $J(\eta)$ is usually given as $J(\eta) = \int (\eta''(x))^2 dx$. A more detailed description can be found in Wasserman (2005) or Gu (2002).

Splines are special piecewise polynomials with appropriate continuity conditions at points in the domain of the function, and these points are ordered and called *knots*. There have been many nonparametric regression approaches using spline smoothing from both frequentist and Bayesian interpretations. One fundamental difficulty, however is determining the number and the location of knots, and this issue has led to a new approach, called *free-knot splines*, where the number of knots and their locations are determined from the data (Denison et al. (1998)). DiMatteo et al. (2001) considered a fully Bayesian method for curve-fitting with free-knot splines, called *Bayesian adaptive regression splines* (BARS). Specifically, they modeled η as a polynomial spline of fixed order, while putting a prior on the number of the knots, the locations of the knots and the coefficients of the polynomials. BARS uses a computationally intensive reversible jump MCMC method to perform nonparametric regression and it has shown successful empirical results in many examples.

In terms of asymptotic properties of smoothing splines approach, further investigation needs to be considered. In frequentist approach, there are many well-known results about asymptotic theory concerning the rates of convergence of penalized likelihood estimates (Gu (2002) and Györfi et al. (2002)). General posterior consistency or convergence rates

in Bayesian spline models have been shown using the the same methods as those used for orthonormal basis expansions, as in Huang (2004). However, in free-knot spline models, consistency has not been investigated yet and we hope that asymptotic theory will be studied further in future.

1.3.3 Gaussian Process Method

Gaussian process (GP) modelling has been widely used to solve many statistical problems and their probabilistic characteristics from stochastic process theory have been investigated extensively. There are numerous statistical approaches to deal with GP modelling in statistical literature. In Chapter 2, we will investigate this alternative approach to nonparametric Bayesian regression in detail. This approach is featured by the use Gaussian processes as a prior distribution for the unknown regression function, and is called *Gaussian process regression*, or GP regression. Thus, the approach on which we focus in our thesis is to model η as a Gaussian processes a priori. As we will explain in the next Chapter, Gaussian processes are a natural way of defining prior distributions over spaces of functions, which are the parameter spaces for nonparametric Bayesian regression models.

The first steps in the use of Gaussian processes include O'Hagan (1978), Wahba (1978), Leonard (1978) and others. O'Hagan (1978) and Wahba (1978) suggested the use of Gaussian processes as nonparametric regression priors. Wahba (1978)'s approach was to use a prior for the regression function based on a Brownian motion, which turned out to be the Bayes estimator of a smoothing spline. Leonard (1978) considered a density estimation problem using Gaussian processes and defined a random density with the transformation of Gaussian process. Lenk (1988) and Lenk (1991) extended this idea to nonparametric density estimation with a log normal processes, a logit transformation of Gaussian process. Most recently, Tokdar and Ghosh (2004) made another extension of the work by Lenk (1988) in density estimation using Gaussian process. In addition, essentially the same model has

long been used in spatial statistics under the name of “kriging” or *Gaussian* geostatistical model. A basic description of this model can be found in Diggle et al. (2003)

Although these seminal works have initiated the developments of GP modelling in several ways, GP modelling has not flourished until the mid-nineties, compared to its long history. Recently, GP modelling gained remarkable attention and became popular among machine learning researchers. Specifically, Gaussian processes have been used successfully for regression and classification, particularly in machine learning (Seeger, 2004). Neal (1996) has shown that many Bayesian regression models based on neural networks converge to Gaussian processes in the limit as the number of nodes becomes infinite. This has motivated examination of Gaussian process models for the high-dimensional applications to which neural networks are typically applied (Rasmussen, 1996). Other recent research on the use of Gaussian processes in machine learning includes Gibbs (1997), Girard (2004) and Seeger (2003). As a result, Gaussian process regression models are becoming increasingly popular and a variety of new implementational tools and methods are being developed (e.g. Shi et al. (2005)). Furthermore, the use of Gaussian processes as priors in spatial statistics applications has been also studied actively. Many new covariance functions for Gaussian processes have been suggested to handle different data structures. Along with this Bayesian approach in spatial smoothing, Gaussian process regression or Bayesian kriging approach has shown successful empirical results and become a promising research area. Some examples include Higdon et al. (1998), Fuentes and Smith (2001), Paciorek (2003) and Paciorek and Schervish (2004).

1.4 Thesis Outline

The rest of the thesis is organized as follows.

In Chapter 2, we describe the Gaussian process (GP) regression model in detail. First, we clarify the definition of the Gaussian process and relate it to the nonparametric Bayesian

regression problem. Next, we explore aspects of Gaussian process regression which include covariance functions of the Gaussian process and its implementation. We investigate the property of the covariance function with respect to sample path smoothness for the Gaussian process and outline how to fit Gaussian process regression to the data.

In Chapter 3, we present our main results on posterior consistency of the Gaussian process regression model. First, we describe the model that we are using and then define our metric topology on the set of regression functions to establish almost sure consistency. We state the extension of Schwartz' theorem based on Amewou-Atisso et al. (2003) and Choudhuri et al. (2004). To apply our extension to the problem we consider, we state the assumptions needed to prove consistency in nonparametric regression and verify almost sure consistency. We also give examples of covariance functions for Gaussian processes that both satisfy the smoothness conditions of our theorems and are used in practice.

In Chapter 4, we discuss several issues to be considered after establishing posterior consistency. Among these issues, we address a few problems in detail, including problems of hyperparameters of covariance function, misspecification of error distribution and multidimensional covariates. We provide promising solutions to these questions which apply under appropriate conditions.

Finally, in Chapter 5, we make conclusions and discuss the contributions of the thesis and future work.

Chapter 2

Gaussian Process Regression

2.1 Introduction

We introduced the Gaussian process method in Chapter 1 briefly. In this chapter, we investigate the Gaussian process (GP) regression model in detail. First, we illustrate how the Gaussian process can be used as a prior distribution for the unknown nonparametric regression function by associating the concept of the stochastic process, in particular the Gaussian process with a random function. The covariance function is essential for characterizing Gaussian processes, and we next address the important features of the covariance of functions at large. In addition, we demonstrate how to implement Gaussian process regression, starting from the simple GP regression model to the more complicated one that contains additional parameters which need to be estimated. Finally, we summarize consistency results known in the literature about Gaussian process regression, and point out the limitations of the existing results and motivation for further study in this subject.

2.2 Nonparametric Bayesian Regression and Gaussian Process

Nonparametric Bayesian regression can be simply thought of as an ordinary Bayesian regression with infinite dimensional parameter space of an unknown regression function, $\eta(x)$. Thus, $\eta(x)$ is regarded as the random function and the prior distribution of $\eta(x)$ should be specified. We need to consider specifying probability distributions for a random function $\eta(x)$ and can utilize a stochastic process, in particular the Gaussian process for this purpose.

A stochastic process is a collection of random variables defined on the same probability space (Ω, \mathcal{F}, P) . The details of stochastic processes can be found in standard probability theory books such as Billingsley (1995) and Breiman (1968). If we denote the index of the collection by t from a index set T , the stochastic process can be written as $X = \{X_t(\omega) : t \in T\}$. When we fix t , $X(\omega, t)$ is nothing but a random variable and the stochastic process can be thought of as a family of random variables indexed by $t \in T$. On the other hand, a stochastic process can be interpreted as a single random variable taking values in an infinite dimensional function space by fixing ω . In this case, the random function, $t \rightarrow X_t(\omega)$ is a realization of stochastic process over $t \in T$. Therefore, the stochastic process becomes simply the study of infinite dimensional probability theory and thus it is natural to describe the prior probability distribution for the random regression function $\eta(x)$ as a stochastic process.

There have been several other approaches in other nonparametric Bayesian analyses in the literature. The Dirichlet process and its mixture have been used frequently for density estimation problem and for a prior distribution of distribution function (for example, see Escobar and West (1995), Ghosal et al. (1999) and Choudhuri et al. (2003)). The Polya tree process, the Gamma and the Beta processes as well as the Dirichlet process have been used for a prior distribution in Bayesian survival analysis (Ibrahim et al. (2001)). For nonparametric Bayesian regression problems, the Dirichlet normal mixture approach was

also considered by Müller et al. (1996), following the same approach as in Escobar and West (1995), but Gaussian processes became a popular approach as an alternative to the nonparametric prior distribution of regression function.

A Gaussian process is a stochastic process parameterized by its mean function $\mu : T \rightarrow \mathbb{R}$, $\mu(t) = E(X_t)$, and its covariance function $R : T^2 \rightarrow \mathbb{R}$, $R(s, t) = \text{Cov}(X_s, X_t)$ which we denote $GP(\mu, R)$. According to Kolmogorov's existence theorem, a stochastic process can be characterized by finite-dimensional distributions and to say that $\eta \sim GP(\mu, R)$ means that, for all n and all $t_1, \dots, t_n \in T$, the joint distribution of $(\eta(t_1), \dots, \eta(t_n))$ is an n -variate normal distribution with mean vector $(\mu(t_1), \dots, \mu(t_n))$ and covariance matrix Σ whose (i, j) entry is $R(t_i, t_j)$. A Gaussian process can be viewed as two different points of view, one from the stochastic process and the other from the extension of multivariate normal random variables. From the second point of view, a Gaussian process can be taken as an infinite dimensional generalization of multivariate normal random variables to infinite index sets. In other words, we can say that a stochastic process is Gaussian if every finite dimensional joint distribution is multivariate normal and the covariance function of the Gaussian process is an extension of the covariance matrix of multivariate normal random variables. In this regard, as we put a multivariate normal distribution for a prior distribution on a finite set of parameters, so the natural extension will put a Gaussian process prior for a random function from infinite dimensional parameter spaces. Accordingly, Gaussian processes are a natural way of defining prior distributions over spaces of functions, which are the parameter spaces for nonparametric Bayesian regression models.

One of the most important aspects in Gaussian process is a covariance function $R(\cdot, \cdot)$, which is indispensable to study characteristics of Gaussian processes and which determines the smoothness property of Gaussian processes such as the continuity of sample path or its differentiability. The intrinsic property of the covariance function is that it is a nonnegative

definite function, defined as follows. For all n , every $t_1, \dots, t_n \in T$ and $a_1, \dots, a_n \in R$

$$\sum_{i,j=1}^n a_i a_j R(t_i, t_j) \geq 0$$

In GP regression modeling, there are widely used covariance function choices and we introduce some of them in the next section. In particular, we need to choose a suitable Gaussian process with a specific covariance function as a prior distribution for the regression function $\eta(x)$ to achieve our posterior consistency. Appropriate smoothness assumptions on the covariance function is required for this purpose and it will be summarized briefly in Section 2.3.1.

2.3 Aspects of Gaussian Process Regression

Broadly speaking, a simple GP regression model that we consider can be specified as follows:

$$\begin{aligned} Y_i | \eta &\sim N(\eta(X_i), \sigma^2), \\ \eta(\cdot) &\sim GP(\mu(\cdot), R(\cdot, \cdot)) \end{aligned} \tag{2.1}$$

where σ^2 is the noise variance and $\eta(\cdot)$ is the unknown regression function with a Gaussian process prior distribution, whose mean function is $\mu(\cdot)$ and covariance function is $R(\cdot, \cdot)$. In Section 2.3.2 for model implementation, we mainly focus on this simple GP regression model and mention briefly more complicated model in the end. Before moving to the main theoretical results that we will present in Chapter 3, we will also discuss a few fundamental issues in GP regression, which include the covariance function, model implementation and existing consistency results of GP regression.

2.3.1 Covariance Function, $R(\cdot, \cdot)$

The covariance function is the most important feature of a Gaussian process and it is crucial in GP regression. In estimating the unknown regression function, the covariance function

plays a central role in determining the property of sample paths of the Gaussian process and controlling the smoothness of data. In proving posterior consistency results in Chapter 3, the choices of the covariance functions play an important part in guaranteeing the existence of sample paths of the Gaussian process with a suitable smoothness condition that we will assume. Analytic properties of the sample functions of Gaussian processes have been carefully studied and reviewed in the stochastic process literature (Cramer and Leadbetter (1967), Adler (1990) and Abrahamsen (1997)).

Any form of covariance function can be defined, provided it is a nonnegative definite function as mentioned in the previous section. It is a little hard to have an explicit rule for how to choose a covariance function in modeling, but in many cases, we first need to select the covariance function based on the shape and the smoothness of the sample path to model $\eta(x)$, as a component of a prior distribution. However, the covariance function that we choose before observing the data might contain additional smoothing parameters to be estimated from the model. In such a case, the specific choice of a family of covariance functions is not so critical as long as the family is sufficiently flexible.

There are numerous covariance functions used in practice and we give some examples of commonly used covariance functions. One of the most frequently used types of covariance functions is an isotropic covariance function. An isotropic covariance function is one which depends only on the distance between two values s and t . That is, the isotropic covariance function, $R(s, t)$ is a function of $|s - t|$ as $R(s, t) = g(|s - t|)$, where g is some arbitrary function. Among isotropic covariance functions, popular choices are powered exponential, rational quadratic and Matérn covariance functions as in Table 2.1.

Regarding the isotropic covariance functions, there exists a sufficient condition for the sample paths of these covariance functions to be continuously differentiable.

Theorem 2.3.1. (*Cramer and Leadbetter (1967), P. 171*) Let $\sigma(x, y)$ be an isotropic cor-

Table 2.1: Examples of isotropic covariance functions

Model	Covariance function, $R(s, t)$
Powered exponential	$R(s, t) = \lambda_1^2 \exp(- \beta(s - t) ^p), \quad 0 < p \leq 2$
Rational quadratic	$R(s, t) = \lambda_1^2 \left(1 - \frac{\beta^2(s - t)^2}{1 + \beta^2(s - t)^2} \right)$
Matérn	$R(s, t) = \frac{1}{\Gamma(v)2^{v-1}} (\beta(s - t))^v \mathcal{K}_v(\beta(s - t))$
	$\beta > 0$ and $\mathcal{K}_v(x)$: a modified Bessel function of order v

relation function and $\sigma(x, y) = r(|x - y|)$. Suppose that, for some $a > 3$ and $\lambda > 0$, $r(h)$ has the expansion

$$r(h) = 1 - \frac{\lambda h^2}{2} + O\left\{ \frac{h^2}{|\log |h||^a} \right\}$$

There then exists an equivalent process, possessing continuously differentiable sample path with probability one.

Furthermore, the generalization of the differentiability theorem above can be obtained for the nonstationary or anisotropic covariance function case.

Theorem 2.3.2. (Cramer and Leadbetter (1967), P. 185) Suppose the mean function, $\mu(t)$ has a continuous derivative $\mu'(t)$ in a bounded interval, $0 \leq t \leq T$. Let $r(t, u)$ have a continuous mixed second derivative $r_{11}(t, u) = \partial^2 r / \partial t \partial u$ satisfying, for some constants $C > 0$, $a > 3$, $0 \leq t \leq T$ and all sufficiently small h

$$\Delta_h \Delta_h r_{11}(t, t) \leq \frac{C}{|\log |h||^a},$$

where $\Delta_h \Delta_k r(t, u) = r(t + h, u + k) - r(t + h, u) - r(t, u + k) + r(t, u)$. Then, there exists an equivalent process, possessing a sample derivative which is continuous on $0 \leq t \leq T$, with probability one.

Both theorems guarantee the almost sure existence of a continuous derivative, $\eta'(x)$ of

a random function, $\eta(x)$ and this existence is an essential assumption in proving posterior consistency in GP regression modeling as we will illustrate in the next chapter.

2.3.2 GP Regression Implementation

Under the simplest GP regression model, as in Equation (2.1), the estimation problem becomes a multivariate normal problem, in particular, using conditional distributions, and thus the implementation is straightforward. As our approach is a Bayesian inference, it will be based on the posterior distribution. When the noise variance σ^2 is assumed to be known and the observed data are $(Y_1, x_1), \dots, (Y_n, x_n)$, the posterior distribution of $\{\eta(x_i)\}_{i=1}^n$ can be calculated in the following way. Let η be a vector of values, evaluated at the n points x_1, \dots, x_n , $\eta^t \equiv (\eta(x_1), \eta(x_2), \dots, \eta(x_n))$. Let $R_{n \times n}$ be the covariance matrix calculated by applying the covariance function $R(\cdot, \cdot)$ to all pairs of the n design points as $R_{i,j} = \text{Cov}(\eta(x_i), \eta(x_j)) = R(x_i, x_j)$. That is, given the noise variance, σ^2 , the observed responses (Y_1, \dots, Y_n) have a multivariate normal distribution with the mean vector, η , and the covariance matrix, $\sigma^2 \times I_{n \times n}$, while η has another multivariate normal distribution with mean vector $\mu_n \equiv (\mu(x_1), \dots, \mu(x_n))^t$ and the covariance matrix, $R_{n \times n}$.

$$\begin{aligned} (Y_1, \dots, Y_n | \eta, \sigma^2) &\sim N_n(\eta^t, \sigma^2 I_{n \times n}), \quad \eta^t \equiv (\eta(x_1), \dots, \eta(x_n)) \\ (\eta(x_1), \dots, \eta(x_n)) &\sim N_n(\mu_n, R_{n \times n}), \end{aligned}$$

where $I_{n \times n}$ is an identity matrix. Thus, the posterior distribution of η , $\Pi(\eta | (\{Y_i, X_i\}_{i=1}^n, \sigma^2))$ is proportional to the product of two normal distributions as follows.

$$\begin{aligned} P(\eta | Y, \sigma^2) &= \frac{P(\eta, Y, \sigma^2)}{P(Y, \sigma^2)} = \frac{P(Y | \eta, \sigma^2) P(\eta, \sigma^2)}{\int P(Y | \eta, \sigma^2) P(\eta, \sigma^2) d\eta} \\ &\propto N_n(Y, \sigma^2 I_n) \cdot N_n(\mu_n, K_{n \times n}) \end{aligned}$$

Then, from multivariate normal distribution theory, the conditional posterior distribution $\Pi(\eta|Y, \sigma^2)$ is a multivariate normal distribution with

$$\begin{aligned} E(\eta|Y, \sigma^2) &= \mu_n + R_{n \times n}(R_{n \times n} + \sigma^2 I_n)^{-1}(Y - \mu_n) \\ \text{Var}(\eta|Y, \sigma^2) &= R_{n \times n}(R_{n \times n} + \sigma^2 I_n)^{-1} \sigma^2 I_n. \end{aligned}$$

In addition, the prediction of $\eta(x^*)$ at an unobserved data point, x^* , can be obtained easily with the same principle because we assume $\eta(x)$ is a Gaussian process and thus $(\eta(x_1), \dots, \eta(x_n), \eta(x^*))$ is a $(n+1)$ -variate normal vector. Consequently, the posterior distribution of $\eta(x^*)$ is another normal distribution. When σ^2 is assumed to be unknown, the implementation can be also performed easily based on monte carlo methods, such as Gibbs sampling. We can calculate the conditional posterior distribution of $\Pi(\eta|Y, \sigma^2)$ given σ^2 as above and the conditional posterior of σ^2 , $\Pi(\sigma^2|\eta, Y)$ given η . Thus, Monte Carlo simulation based on Gibbs sampling can be applied in a straightforward manner. In addition, hyperparameters that define the covariance function of the Gaussian process can be sampled using Markov chain methods, and these methods can be found in Neal (1996), Neal (1997) or Paciorek (2003).

2.3.3 Existing Consistency Results

Posterior consistency in nonparametric regression problems with Gaussian process priors has been studied mainly in the orthogonal basis expansion framework mentioned earlier. Brown and Low (1996); Freedman (1999); Zhao (2000) have exploited an asymptotic equivalence between white-noise problems and nonparametric regression to prove that the existence of a consistent estimator in a white-noise problem implies the existence of a corresponding consistent estimator in a nonparametric regression problem. The white-noise problem is one in which an observation process $Y(x)$ is modeled as

$$Y(x) = \int_0^x \eta(t) dt + n^{-1/2} \epsilon(x),$$

where $\epsilon(x)$ is a Brownian motion. Since the white-noise problem can be converted easily to the infinitely many normal means problem (see Wasserman (2005)), they use an infinitely many normal means prior for η by expanding $\eta(x)$ as $\eta(x) = \sum_{i=1}^{\infty} \theta_i \varphi(x)$, where $\varphi(x)$ is an orthonormal basis and $\theta_i = \int \eta(x) \varphi(x) dx$. They show that the posterior mean of η is consistent in the white-noise problem under certain conditions on the prior, and thus argue the consistency of an estimator in nonparametric problems. The corresponding estimator in the nonparametric regression problem might not be the posterior mean of η , however.

In addition, to use a prior distribution in the white-noise problem requires either knowing the eigenvalue decomposition of the desired covariance function (analytically, not numerically) or letting the covariance function be determined by the orthogonal basis. For example, Freedman (1999) reconsidered the example to treat the regression function $\eta(x)$ as stochastic processes, specifically integrated Brownian motion on the unit interval, from Cox (1993). While the example was used for another purpose rather than consistency itself, he also pointed out the difficulty in converting it into discrete (infinitely many normal means) problem due to the difficulty of finding the eigenvalues.

In many applications, such as those described by Neal (1996); Rasmussen (1996); Seeger (2004), one uses a particular form of covariance function (like squared exponential) in order to model desired forms of dependence. In such cases, one would like to be able to verify consistency for the particular prior distribution that one is using. For example, Ghosal and Roy (2005) prove in-probability consistency of posterior distributions in binary regression problems with mild conditions on the Gaussian process prior. They do this by extending a result of Schwartz (1965) to the case of independent, not identically distributed observations. For density estimation using Gaussian process priors, Tokdar and Ghosh (2004) establish posterior consistency. They identify the support of a Gaussian process that contains all continuous functions, for a wide range of Gaussian processes.

The posterior consistency problem of GP regression models has not been seriously stud-

ied in the literature. We follow the approach of Ghosal and Roy (2005) for nonparametric regression problems and provide our results in the next chapter. In the following chapter, we show that the posterior distribution of $\eta(x)$ is consistent under suitable conditions.

Chapter 3

Posterior Consistency of GP Regression

3.1 Introduction

In this chapter, we establish almost sure consistency for posterior distributions in GP regression and identify conditions to achieve it. The consistency of the posterior is proven by using the extension of results by Ghosal and Roy (2005). While extending their results to almost sure consistency, we adapt this extension to nonparametric regression problems with normal and other types of errors and with unknown error variance. The work of Ghosal and Roy (2005) was devised for nonparametric binary regression problems, where the regression function is a probability function and thus is uniformly bounded by one, and they established in probability consistency results. We broaden the scope of their approach to general nonparametric regression problems by providing almost sure consistency results in a GP regression model.

In order to establish posterior consistency, neighborhoods of the true parameter, i.e. neighborhoods of the true regression function must be specified, first. For this purpose,

we consider different topologies of functions, depending on the type of covariate in the regression model, whether it is a fixed design covariate or randomly chosen covariate with a probability distribution, and then we provide consistency results based on these topologies. Specifically, we will show almost sure consistency in a topology that is weaker than the L^1 topology, usually used in the spaces of functions. The need for a weaker topology arises from the fact that our regression functions can be unbounded. With the weaker topology, we can handle both random covariate and nonrandom design points. When the design points are fixed, we will strengthen the result to the case of the L^1 topology. When the covariate are random variable with a probability distribution, we will prove almost sure consistency of the posterior probabilities of Hellinger neighborhoods of the joint density of the response and design point. If we assume that the regression functions are uniformly bounded, we will prove almost sure consistency of L^1 neighborhoods of the true regression function.

To obtain our final results described above, we need to clarify the model that we are using. The model under consideration is a fully Bayesian GP regression model as follows. Let us consider a random response Y corresponding to a single covariate X taking values in a bounded interval $T \subset \mathbb{R}$. We are interested in estimating the regression function, $\eta(x) = E(Y|X = x)$ based on independent observations of (X, Y) . We do not assume a parametric form for the regression function, but rather we assume some smoothness conditions. We model the unknown function η as a random process with a Gaussian process (GP) prior distribution.

To be specific, the GP regression model we consider here, is the following.

$$\begin{aligned}
 Y_i &= \eta(X_i) + \epsilon_i, \quad i = 1, \dots, n, \\
 \epsilon_i &\sim N(0, \sigma^2) \text{ or } DE(0, \sigma) \text{ given } \sigma, \\
 \eta(\cdot) &\sim GP(\mu(\cdot), R(\cdot, \cdot; \beta)) \text{ conditional on } \beta, \text{ independent of } \sigma \text{ and } (\epsilon_1, \dots, \epsilon_n), \\
 \sigma &\sim \nu, \text{ and } \beta \sim \kappa
 \end{aligned} \tag{3.1}$$

where ν and κ are probability measures with support \mathbb{R}^+ , and $DE(0, \sigma)$ stands for the double exponential (or Laplace) distribution with median 0 and scale factor σ . The additional parameter β that parameterizes the covariance function is needed in order to ensure that the support of the prior distribution is sufficiently broad. This will be clarified in Lemma 3.5.3. In other words, the conditional distribution of Y given $X = x$ with the unknown parameters η and σ , is normal or double exponential with mean function $\eta(x) \equiv E(Y|X = x)$. The covariate, X , will be either be fixed or sampled from a distribution, Q , independent of ϵ 's, σ and η .

The objective of this chapter is to identify conditions on the GP prior distribution of η and the sequence of predictors $\{X_i\}_{i=1}^\infty$ that guarantee almost sure consistency of the posterior distribution under the model described above.

Suppose that the true response function, $\eta_0(x)$, as a function of the covariate, X , is a continuously differentiable function on a bounded interval T . Without loss of generality, we will assume that $T = [0, 1]$ for the remainder of this chapter. Our work is similar to that of Ghosal and Roy (2005) who give conditions for in-probability consistency of posteriors in general non-identically distributed data problems. We prove an extension of their theorem to the case of almost sure consistency and we identify conditions on the GP prior that allow us to verify the conditions of this extension

3.2 Topologies on the Set of Regression Functions

As we discussed in Chapter 1, posterior consistency involves examining a posterior probability of a set of parameter values as the sample size, n , goes to infinity. The set is any neighborhood of the true parameter θ_0 and it is necessary to define suitable topologies for each parameter of interest. Since our parameter space involves a space of regression functions, there are many choices of topology.

First, we need to be clear on what we mean by the expression “almost sure consistency”. Let \mathcal{F} be the set of Borel measurable functions defined on T . For now, assume that we have chosen a topology on \mathcal{F} . For each neighborhood \mathcal{N} of the true regression function η_0 and each sample size n , we compute the posterior probability

$$p_{n,\mathcal{N}}(Y_1, \dots, Y_n, X_1, \dots, X_n) = \Pr(\{\eta \in \mathcal{N}\} | Y_1, \dots, Y_n, X_1, \dots, X_n), \quad (3.2)$$

as a function of the data. To say that the posterior distribution of η is almost surely consistent means that, for every neighborhood \mathcal{N} , $\lim_{n \rightarrow \infty} p_{n,\mathcal{N}} = 1$ a.s. with respect to the joint distribution of the infinite sequence of data values. Similarly, in-probability consistency means that for all \mathcal{N} , $p_{n,\mathcal{N}}$ converges to 1 in probability.

To make these definitions precise, we must specify the topology on \mathcal{F} . This topology can be chosen independently of whether one wishes to consider almost sure consistency or in-probability consistency of the posterior. We consider three topologies on \mathcal{F} and they are L^1 topology, a topology based on the probability measure of the domain of functions, named “in-probability topology” and a topology based on a Hellinger distance. Popular choices of topology on \mathcal{F} include the L^p topologies related to a probability measure Q on the domain $[0, 1]$ of the regression functions. For $1 \leq p < \infty$, the $L^p(Q)$ distance between two functions η_1 and η_2 is $\|\eta_1 - \eta_2\|_p = [\int_T |\eta_1 - \eta_2|^p dQ]^{1/p}$. For $p = \infty$, the $L^\infty(Q)$ distance is

$$\|\eta_1 - \eta_2\|_\infty = \inf_{A: Q(A)=1} \sup_{x \in A} |\eta_1(x) - \eta_2(x)|.$$

For example, Ghosal and Roy (2005) use the L^1 topology related to Lebesgue measure and prove in-probability consistency in the binary regression setting. We will also consider L^1 topology to show one of our final results.

Another topology on \mathcal{F} is the topology of in-probability convergence related to a probability Q . This topology is weaker than the $L^p(Q)$ topologies. As with the $L^p(Q)$ topologies, we must count as identical all functions that equal each other a.s. $[Q]$. Lemma 3.2.1 gives a metric representation of the topology of in-probability convergence.

Lemma 3.2.1. *Let (T, \mathcal{B}, Q) be a probability space, and let \mathcal{F} be the set of all real-valued measurable functions defined on T . Define*

$$d_Q(\eta_1, \eta_2) = \inf\{\epsilon : Q(\{x : |\eta_1(x) - \eta_2(x)| > \epsilon\}) < \epsilon\}.$$

Then d_Q is a metric on the set of equivalence classes under the relation $\eta_1 \sim \eta_2$ if $\eta_1 = \eta_2$ a.s. $[Q]$. If X has distribution Q and $\{\eta_n\}_{n=1}^\infty$ is a sequence of elements of \mathcal{F} , then $\eta_n(X)$ converges to $\eta(X)$ in probability if and only if $\lim_{n \rightarrow \infty} d_Q(\eta_n, \eta) = 0$.

Proof. Clearly, $d_Q(f, g) = d_Q(g, f)$, $d_Q(f, g) \geq 0$, and $d_Q(f, f) = 0$. If $d_Q(f, g) = 0$ then $f = g$ a.s. $[Q]$. All that remains for the proof that d_Q is a metric is to verify the triangle inequality.

For each $f, g \in \mathcal{F}$, define $B_{f,g} = \{\epsilon : Q(\{x : |f(x) - g(x)| > \epsilon\}) < \epsilon\}$. Then $d_Q(f, g) = \inf B_{f,g}$. We need to verify

$$\inf B_{f,g} \leq \inf B_{f,h} + \inf B_{h,g}. \quad (3.3)$$

We will show that if $\epsilon_1 \in B_{f,h}$ and $\epsilon_2 \in B_{h,g}$ then $\epsilon_1 + \epsilon_2 \in B_{f,g}$, which implies 3.3. Let $\epsilon_1 \in B_{f,h}$ and $\epsilon_2 \in B_{h,g}$. Then

$$\{x : |f(x) - g(x)| > \epsilon_1 + \epsilon_2\} \subseteq \{x : |f(x) - h(x)| > \epsilon_1\} \cup \{x : |h(x) - g(x)| > \epsilon_2\}.$$

It follows that

$$\begin{aligned} Q(\{x : |f(x) - g(x)| > \epsilon_1 + \epsilon_2\}) &\leq Q(\{x : |f(x) - h(x)| > \epsilon_1\}) + Q(\{x : |h(x) - g(x)| > \epsilon_2\}) \\ &\leq \epsilon_1 + \epsilon_2. \end{aligned}$$

Hence $\epsilon_1 + \epsilon_2 \in B_{f,g}$.

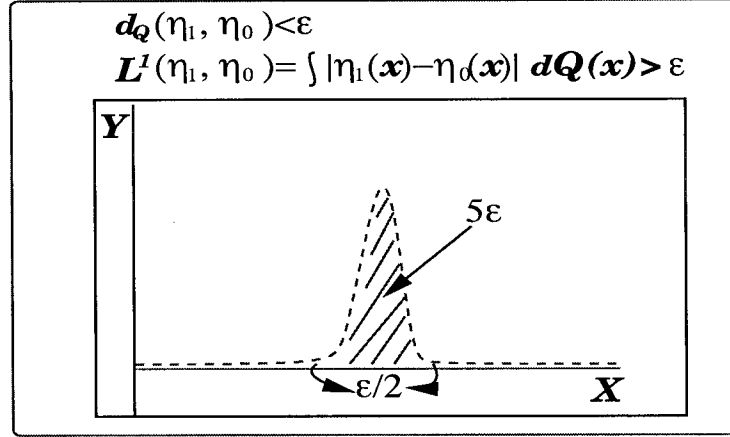
To prove the equivalence of d_Q convergence and convergence in probability, assume that X has distribution Q . First, assume that $\eta_n(X)$ converges to $\eta(X)$ in probability. Then, for every $\epsilon > 0$, $\lim_{n \rightarrow \infty} Q(\{x : |\eta_n(x) - \eta(x)| > \epsilon\}) = 0$. So, for every $\epsilon > 0$

there exists N such that for all $n \geq N$, $Q(\{x : |\eta_n(x) - \eta(x)| > \epsilon\}) < \epsilon$. In other words, for every $\epsilon > 0$, there exists N such that for all $n \geq N$, $d_Q(\eta_n, \eta) \leq \epsilon$. This is what it means to say $\lim_{n \rightarrow \infty} d_Q(\eta_n, \eta) = 0$. Finally, assume that $\lim_{n \rightarrow \infty} d_Q(\eta_n, \eta) = 0$. Then, for every $\epsilon > 0$ there exists N such that for all $n \geq N$, $d_Q(\eta_n, \eta) \leq \epsilon$, which is equivalent to $Q(\{x : |\eta_n(x) - \eta(x)| > \epsilon\}) < \epsilon$. Hence, $\eta_n(X)$ converges to $\eta(X)$ in probability. ■

It is well known that $L^p(Q)$ convergence implies in-probability convergence, so that $L^p(Q)$ neighborhoods must be smaller than d_Q neighborhoods in some sense. It is not difficult to show that for every $\epsilon \in (0, 1)$, the ball of radius ϵ under d_Q contains the ball of radius $\epsilon^{1+1/p}$ in $L^p(Q)$ for all $1 \leq p \leq \infty$. In addition, for every p and every $\delta > \epsilon^{1+1/p}$, there exist functions in the ball of radius δ under $L^p(Q)$ that are not in the ball of radius ϵ under d_Q . When the random variables are all bounded, in-probability convergence implies L^p convergence for all finite p .

Figure 3.2 shows the difference between the $L^1(Q)$ and d_Q metrics graphically. There are two functions, η_0 and η_1 . The solid straight line indicates η_0 , and the dotted curve indicates η_1 . In Figure 3.2, the length of the region on the X axis where η_0 and η_1 are different is less than ϵ while the size of the shaded area is bigger than ϵ . This means that the distance between the two functions in terms of the d_Q metric is less than ϵ while the distance between the two functions in terms of $L^1(Q)$ metric is larger than ϵ . In fact, for each $\delta, \epsilon > 0$ there are examples like the one in Figure 3.2 in which η_1 is inside the d_Q neighborhood of size ϵ around η_0 but η_1 is outside the $L^1(Q)$ neighborhood of size δ around η_0 . This distinction will be important in proving posterior consistency, particularly for constructing a test in Section 3.5.3.

When the values of the predictor X are chosen deterministically (and satisfy a condition relative to Lebesgue measure λ) we prove almost sure consistency of posterior probabilities of $L^1(\lambda)$ neighborhoods of the true regression function. When the values of the predictor X have a distribution Q , we prove almost sure consistency of posterior probabilities of d_Q

Figure 3.1: A comparison between $L^1(Q)$ metric and d_Q metric

neighborhoods of the true regression function. If we make an additional assumption that the regression function is uniformly bounded by a known constant, we can prove almost sure consistency of posterior probabilities of $L^1(Q)$ neighborhoods even when X is random.

Finally, an alternative to topologizing regression functions is to place a topology on the set of distributions. For the case in which the predictor X is random, we shall consider this alternative as well as the topologies mentioned above. In particular, we shall use the Hellinger metric on the collection of joint distributions of (X, Y) . Suppose that X has distribution Q and Y has a density $f(y|x)$ with respect to Lebesgue measure λ given $X = x$. Then $f(y|x)$ is a joint density of (X, Y) with respect to $\xi = Q \times \lambda$. Under the true regression function η_0 with noise scale parameter σ_0 , we denote the conditional density of Y by $f_0(y|x)$. In this case, the Hellinger distance between the two distributions corresponding to (η, σ) and (η_0, σ_0) is the following:

$$d_H(f, f_0) = \int \left[\sqrt{f(y|x)} - \sqrt{f_0(y|x)} \right]^2 d\xi(x, y).$$

It is easy to show that $d_H(f, f_0)$ is unchanged if one chooses a different dominating measure

instead of ξ . For the case just described, we will show that, under conditions similar to those of our other theorems, the posterior probability of each Hellinger neighborhood of f_0 converges almost surely to 1.

3.3 Consistency Theorems for Non i.i.d. Observations

Schwartz (1965) proved a theorem that gave conditions for consistency of posterior distributions of parameters of the distributions of independent and identically distributed (i.i.d.) random variables. These conditions include the existence of tests with sufficiently small error rates and the prior positivity of certain neighborhoods of η_0 . There have been several extensions of the theorem by Barron et al. (1999) and Ghosal et al. (1999) for i.i.d observations. When the observations are independent but not identically distributed (non i.i.d.), there are some challenges and the renowned theorem of Schwartz should be modified for non i.i.d observations. Amewou-Atisso et al. (2003) considered an extension of the theorem of Schwartz, designed for a semiparametric regression problem, where the observations, Y_1, Y_2, \dots 's are non i.i.d with a linear regression model as $Y_i = \alpha + \beta x_i + \epsilon_i$ but the error distribution is assumed to be unknown. Choudhuri et al. (2004) extended the theorem of Schwartz to a triangular array of independent non-identically distributed observations for the case of convergence in-probability, dealing with a curve estimation problem in a stronger topology than Amewou-Atisso et al. (2003).

We provide another extension of Schwartz's theorem to almost sure convergence. Our extension is based on both Amewou-Atisso et al. (2003) and Choudhuri et al. (2004). We present this extension as Theorem 3.3.1 and also verify the conditions for a wide class of GP priors.

Theorem 3.3.1. *Let $\{Z_i\}_{i=1}^\infty$ be independently distributed with densities $\{f_i(\cdot; \theta)\}_{i=1}^\infty$, with respect to a common σ -finite measure, where the parameter θ belongs to an abstract mea-*

surable space Θ . The densities $f_i(\cdot; \theta)$ are assumed to be jointly measurable. Let $\theta_0 \in \Theta$ and let P_{θ_0} stand for the joint distribution of $\{Z_i\}_{i=1}^{\infty}$ when θ_0 is the true value of θ . Let $\{U_n\}_{n=1}^{\infty}$ be a sequence of subsets of Θ . Let θ have prior Π on Θ . Define

$$\begin{aligned}\Lambda_i(\theta_0, \theta) &= \log \frac{f_i(Z_i; \theta_0)}{f_i(Z_i; \theta)}, \\ K_i(\theta_0, \theta) &= E_{\theta_0}(\Lambda_i(\theta_0, \theta)), \\ V_i(\theta_0, \theta) &= \text{Var}_{\theta_0}(\Lambda_i(\theta_0, \theta)).\end{aligned}$$

(A1) *Prior positivity of neighborhoods.*

Suppose that there exists a set B with $\Pi(B) > 0$ such that

- (i) $\sum_{i=1}^{\infty} \frac{V_i(\theta_0, \theta)}{i^2} < \infty, \forall \theta \in B,$
- (ii) For all $\epsilon > 0$, $\Pi(B \cap \{\theta : K_i(\theta_0, \theta) < \epsilon \text{ for all } i\}) > 0.$

(A2) *Existence of tests*

Suppose that there exist test functions $\{\Phi_n\}_{n=1}^{\infty}$, sets $\{\Theta_n\}_{n=1}^{\infty}$ and constants $C_1, C_2, c_1, c_2 > 0$ such that

- (i) $\sum_{n=1}^{\infty} E_{\theta_0} \Phi_n < \infty,$
- (ii) $\sup_{\theta \in U_n^C \cap \Theta_n} E_{\theta}(1 - \Phi_n) \leq C_1 e^{-c_1 n},$
- (iii) $\Pi(\Theta_n^C) \leq C_2 e^{-c_2 n}.$

Then

$$\Pi(\theta \in U_n^C | Z_1, \dots, Z_n) \rightarrow 0 \quad \text{a.s.}[P_{\theta_0}]. \quad (3.4)$$

Figure 3.2 is helpful to understand some of the conditions in Theorem 3.3.1. The first condition (A1) assumes that there are sets with positive prior probabilities, which could be regarded as neighborhoods of the true parameter θ_0 . In Figure 3.2, the set B is the neighborhood of the true parameter θ_0 which has a positive prior probability and

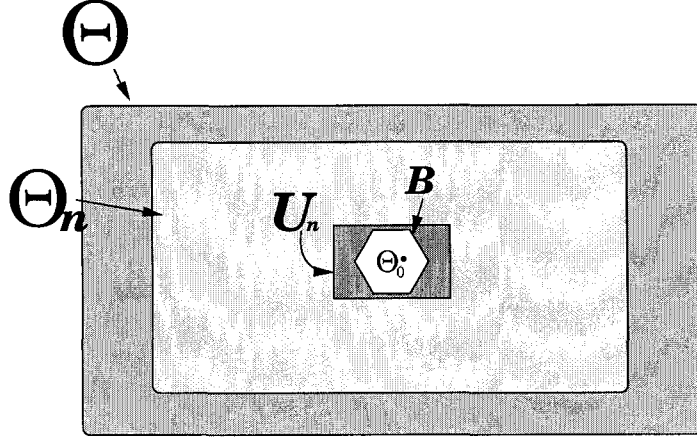


Figure 3.2: Graphical display of Theorem 3.3.1

specifically, it intersects Kullback-Leibler neighborhoods of the true parameter θ_0 . The second condition (A2) assumes the existence of certain tests of the hypothesis $\theta = \theta_0$. We construct a test which makes θ_0 be separated from $\Theta_n \cap U_n^c$ as much as possible. Θ_n is a sieve which grows to the parameter space Θ as sample size increases. We need to consider the sieve Θ_n because the parameter under consideration is infinite dimensional. We assume that tests with vanishingly small type I error probability exist. We also assume that these tests have exponentially small type II error probability on part of the complement of a set U_n containing θ_0 , namely $\Theta_n \cap U_n^c$. Finally, Θ_n^C is assumed to have exponentially small prior probability.

Proof. The posterior probability (3.4) can be written as

$$\begin{aligned}
 \Pi(\theta \in U_n^C | Z_1, \dots, Z_n) &= \frac{\int_{U_n^C \cap \Theta_n} \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta) + \int_{U_n^C \cap \Theta_n^C} \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta)}{\int_{\Theta} \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta)} \\
 &\leq \Phi_n + \frac{(1 - \Phi_n) \int_{U_n^C \cap \Theta_n} \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta) + \int_{U_n^C \cap \Theta_n^C} \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta)}{\int_{\Theta} \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta)} \\
 &= \Phi_n + \frac{I_{1n}(Z_1, \dots, Z_n) + I_{2n}(Z_1, \dots, Z_n)}{I_{3n}(Z_1, \dots, Z_n)}. \tag{3.5}
 \end{aligned}$$

The remainder of the proof consists of proving the following results:

$$\Phi_n \rightarrow 0 \quad \text{a.s.}[P_{\theta_0}], \quad (3.6)$$

$$e^{\beta_1 n} I_{1n}(Z_1, \dots, Z_n) \rightarrow 0 \quad \text{a.s.}[P_{\theta_0}] \text{ for some } \beta_1 > 0, \quad (3.7)$$

$$e^{\beta_2 n} I_{2n}(Z_1, \dots, Z_n) \rightarrow 0 \quad \text{a.s.}[P_{\theta_0}] \text{ for some } \beta_2 > 0, \quad (3.8)$$

$$e^{\beta n} I_{3n}(Z_1, \dots, Z_n) \rightarrow \infty \quad \text{a.s.}[P_{\theta_0}] \text{ for all } \beta > 0. \quad (3.9)$$

Letting $\beta > \max\{\beta_1, \beta_2\}$ will imply (3.4).

Before giving the details of the proof, we offer the following summary. The first term on the right hand side of (3.5) goes to 0 with probability 1, by the first Borel-Canteli lemma from (A2) (i). We show that the two terms in the numerator, I_{1n} and I_{2n} are exponentially small. That is, for some $\beta_1 > 0$ and $\beta_2 > 0$, $e^{\beta_1 n} I_{1n}$ and $e^{\beta_2 n} I_{2n}$ go to 0 from (A2) (ii) and (iii) with P_{θ_0} probability 1. Finally, we show $e^{\beta n} I_{3n} \rightarrow \infty$, with P_{θ_0} probability 1 for all $\beta > 0$, using Kolmogorov's strong law of large numbers for independent but not identically distributed random variables under the condition (A1).

First, we prove (3.6). By the Markov inequality, for every $\epsilon > 0$, $P_{\theta_0}(|\Phi_n| > \epsilon) \leq E_{\theta_0}(|\Phi_n|)/\epsilon$. By (i) of (A2), we have $\sum_{n=1}^{\infty} P_{\theta_0}(|\Phi_n| > \epsilon) < \infty$. By the first Borel-Cantelli lemma, $P_{\theta_0}(|\Phi_n| > \epsilon \text{ i.o.}) = 0$. Since this is true for every $\epsilon > 0$, we have (3.6).

The remainder of the proof is borrowed and modified from the proofs in Amewou-Atisso et al. (2003) and Choudhuri et al. (2004). (3.7) and (3.8) follow from the proof of Choudhuri et al. (2004, Theorem 2), and (3.9) follows from Amewou-Atisso et al. (2003, Lemma A.1) and the proof of Amewou-Atisso et al. (2003, Theorem 2.1).

Here are the remaining details. We continue to prove (3.7). For every nonnegative function ψ ,

$$E_{\theta_0} \left[\psi(Z_1, \dots, Z_n) \int_C \prod_{i=1}^n \frac{f(Z_i, \theta)}{f(Z_i, \theta_0)} d\Pi(\theta) \right] = \int_C E_{\theta}(\psi) d\Pi(\theta), \quad (3.10)$$

by Fubini's theorem. Let $\psi = 1 - \Phi_n$ and get

$$\begin{aligned} \mathbb{E}_{\theta_0} I_{1n}(Z_1, \dots, Z_n) &= \mathbb{E}_{\theta_0} \left[(1 - \Phi_n) \int_{\Theta_n \cap U_n^C} \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta) \right] \\ &= \int_{\Theta_n \cap U_n^C} \mathbb{E}_{\theta} [(1 - \Phi_n)] \\ &\leq \sup_{\theta \in \Theta_n \cap U_n^C} \mathbb{E}_{\theta} (1 - \Phi_n) \\ &\leq C_1 e^{-c_1 n}, \end{aligned}$$

where the final inequality follows from condition (ii) of (A2). Thus,

$$P_{\theta_0} \left\{ (1 - \Phi_n) \int_{\Theta_n \cap U_n^C} \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta) \geq e^{-c_1 \frac{n}{2}} \right\} \leq C_1 e^{c_1 \frac{n}{2}} e^{-c_1 n} = C_1 e^{-c_1 \frac{n}{2}}$$

An application of the first Borel-Cantelli Lemma yields

$$(1 - \Phi_n) \int_{\Theta_n \cap U_n^C} \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta) \leq e^{-c_1 \frac{n}{2}}$$

all but finitely often with P_{θ_0} probability 1. Therefore,

$$e^{c_1 \frac{n}{4}} I_{1n} \rightarrow 0 \quad \text{a.s.}[P_{\theta_0}]$$

Next, we prove (3.8). Applying (3.10) and condition (iii) of (A2), we get

$$\begin{aligned} \mathbb{E}_{\theta_0} I_{2n}(Z_1, \dots, Z_n) &= \mathbb{E}_{\theta_0} \left[\int_{U_n^C \cap \Theta_n^C} \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta) \right] \\ &\leq \Pi(\Theta_n^C) \\ &\leq C_2 e^{-c_2 n}, \end{aligned}$$

Again, the first Borel-Cantelli Lemma implies

$$e^{c_2 \frac{n}{4}} I_{2n} \rightarrow 0 \quad \text{a.s.}[P_{\theta_0}].$$

Next, we prove (3.9). Define $\log_+(x) = \max\{0, \log(x)\}$ and $\log_-(x) = -\min\{0, \log(x)\}$.

Also, define

$$\begin{aligned} W_i &= \log_+ \frac{f_i(Z_i, \theta_0)}{f_i(Z_i, \theta)}, \\ K_i^+(\theta_0, \theta) &= \int f_i(z, \theta_0) \log_+ \frac{f_i(z, \theta_0)}{f_i(z, \theta)} dz, \\ K_i^-(\theta_0, \theta) &= \int f_i(z, \theta_0) \log_- \frac{f_i(z, \theta_0)}{f_i(z, \theta)} dz. \end{aligned}$$

Then

$$\begin{aligned} \text{Var}_{\theta_0}(W_i) &= \mathbb{E}(W_i^2) - \{K_i^+(\theta_0, \theta)\}^2 \\ &\leq \mathbb{E}(W_i^2) - \{K_i^+(\theta_0, \theta) - K_i^-(\theta_0, \theta)\}^2 \\ &= \mathbb{E}(W_i^2) - \{K_i(\theta_0, \theta)\}^2 \\ &\leq \int f_i(Z_i, \theta_0) \left(\log_+ \frac{f_i(Z_i, \theta_0)}{f_i(Z_i, \theta)} \right)^2 + \int f_i(Z_i, \theta_0) \left(\log_- \frac{f_i(Z_i, \theta_0)}{f_i(Z_i, \theta)} \right)^2 - \{K_i(\theta_0, \theta)\}^2 \\ &= \int f_i(Z_i, \theta_0) \left(\log_+ \frac{f_i(Z_i, \theta_0)}{f_i(Z_i, \theta)} - \log_- \frac{f_i(Z_i, \theta_0)}{f_i(Z_i, \theta)} \right)^2 - \{K_i(\theta_0, \theta)\}^2 \\ &= V_i(\theta_0, \theta), \end{aligned}$$

where the next-to-last equality follows from the fact that $\log_+(x) \log_-(x) = 0$ for all x . It follows that $\sum_{i=1}^{\infty} \text{Var}_{\theta_0}(W_i)/i^2 < \infty$ for all $\theta \in B$, the set defined in condition (A1).

According to Kolmogorov's strong law of large numbers for independent non-identically distributed random variables (e.g. Breiman (1968, Theorem 3.27)),

$$\frac{1}{n} \sum_{i=1}^n (W_i - K_i^+(\theta_0, \theta)) \rightarrow 0, \quad \text{a.s.}[P_{\theta_0}]. \quad (3.11)$$

For each $\theta \in B$, with P_{θ_0} probability 1

$$\begin{aligned}
\liminf_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n \log \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} \right) &\geq - \limsup_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n \log_+ \frac{f_i(Z_i, \theta_0)}{f_i(Z_i, \theta)} \right) \\
&= - \limsup_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n K_i^+(\theta_0, \theta) \right) \\
&\geq - \limsup_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n K_i(\theta_0, \theta) + \frac{1}{n} \sum_{i=1}^n \sqrt{K_i(\theta_0, \theta)/2} \right) \\
&\geq - \limsup_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n K_i(\theta_0, \theta) + \sqrt{\frac{1}{n} \sum_{i=1}^n K_i(\theta_0, \theta)/2} \right),
\end{aligned}$$

where the second line follows from (3.11), the third line follows from Amewou-Atisso et al. (2003, Lemma A.1), and the fourth follows from Jensen's inequality.

Let $\beta > 0$, and choose ϵ so that $\epsilon + \sqrt{\epsilon/2} \leq \beta/8$. Let $C = B \cap \{\theta : K_i(\theta_0, \theta) < \epsilon \text{ for all } i\}$.

For $\theta \in C$, $n^{-1} \sum_{i=1}^n K_i(\theta_0, \theta) < \epsilon$, so for each $\theta \in C$,

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} \geq -(\epsilon + \sqrt{\epsilon/2}).$$

Now,

$$I_{3n} \geq \int_C \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta),$$

it follows from Fatou's lemma that

$$e^{n\beta/4} I_{3n} \rightarrow \infty, \text{ a.s. } [P_{\theta_0}], \text{ for all } \beta > 0.$$

■

3.4 Consistency in Nonparametric Regression

In this section, we apply Theorem 3.3.1 to cases of GP regression models (3.1), in which the prior Π is the product of a GP distribution for the regression function given β , a prior ν for σ , and a prior κ for β as described in Section 3.1. We must make assumptions about the

prior Π that we will use in the model, in particular about the smoothness of the GP prior as well as about the rate at which the design points x_i , $i = 1, \dots, n$, fill out the interval $[0, 1]$. For the latter, we consider two versions of the assumption on design points, one for random covariates and one for nonrandom (fixed) covariates.

Assumption RD. The random design points (covariates) $\{X_n\}_{n=1}^\infty$ are independent and identically distributed with probability distribution Q on $[0, 1]$.

Assumption NRD. Let $0 = x_0 < x_1 \leq x_2 \leq \dots \leq x_n < x_{n+1} = 1$ be the design points on $[0, 1]$ and let $S_i = x_{i+1} - x_i$, $i = 0, \dots, n$ denote the spacings between them. There is a constant $0 < K_1 < 1$ such that the $\max_{0 \leq i \leq n} S_i < 1/(K_1 n)$.

Note that assumption NRD is obviously satisfied by the equally spaced design. We also need to specify conditions on the prior distributions for η , β and σ . The conditions on the GP prior for η have two main parts. One is prior positivity, and the other is the smoothness of sample paths. We assume that the covariance function has the form $R(x, x'; \beta) = R_0(\beta|x - x'|)$, where R_0 is a positive multiple of a nowhere zero density function of one real variable with appropriate smoothness. In this respect, we follow the approach of Tokdar and Ghosh (2004) and Ghosal and Roy (2005).

As an example, consider the powered exponential covariance function as in Table 2.1.

$$R(x, x'; \beta) = R_0(\beta|x - x'|) = \lambda_1^2 \exp(-|\beta(x - x')|^p),$$

where $R_0(w) = \lambda_1^2 \exp(-w^p)$ and $0 < p \leq 2$.

We assume that the support of β is R^+ and thus, β determines primarily the shape of covariance function. It can be observed that as β becomes extreme (close to infinity), the shape of covariance function becomes flat. Here, we state our assumption, Assumption P about the prior distributions.

Assumption P.

P1. For all $n \geq 1$, all $\beta > 0$ and all $x_1, \dots, x_n \in [0, 1]$, the n -variate covariance matrix,

$((\Sigma_{i,j}))$ with $\Sigma_{i,j} = R(x_i, x_j; \beta)$, is non-singular.

P2. The covariance function, $R(x, x'; \beta)$ has the form $R_0(\beta|x - x'|)$, where $R_0(x)$ is a positive multiple of a nowhere zero density function on \mathbb{R} and four times continuously differentiable on \mathbb{R} .

P3. The mean function $\mu(x)$ of the Gaussian process $\eta(x)$ is continuously differentiable in $[0, 1]$.

P4. ν assigns positive probability to every neighborhood of the true scale parameter σ_0 , i.e. for every $\epsilon > 0$, $\nu \left\{ \left| \frac{\sigma}{\sigma_0} - 1 \right| < \epsilon \right\} > 0$.

P5. There exists $0 < \delta < 1/2$ and $b_1, b_2 > 0$ such that

$$\kappa \left\{ \beta > n^\delta \right\} = \Pr \left\{ \beta > n^\delta \right\} < b_1 \exp(-b_2 n), \quad \forall n \geq 1$$

A direct consequence of **P1** is that the variance of $\eta(x)$, $R(x, x; \beta)$ is always positive. Otherwise, the covariance matrix generated by the covariance function cannot be always non-singular. **P2**, **P3** and **P5** together are sufficient conditions that the support of the GP prior contains every continuously differentiable function. We will explain this fact in detail in Section 3.5.5, pursuing the same approach as the works of Tokdar and Ghosh (2004) and Ghosal and Roy (2005). In addition, the important consequence of **P2** and **P3** is that there exists a constant $K_2 > 0$ such that

$$r_0(h) = 1 - \frac{K_2}{2} h^2 + O(h^4)$$

where $r_0(h) = R_0(|h|)/R_0(0)$, and $\text{Cov}(\eta'(x), \eta'(x')) = -\beta^2 R_0''(\beta|x - x'|)$. That is, these two conditions guarantee the existence of continuous sample derivative $\eta'(x)$ with probability 1 (See 2.3.1 in Chapter 2) and that $\eta'(x)$ is also a Gaussian process, which will be investigated in Lemma 3.5.5. Many covariance functions of Gaussian processes, which are widely used in the literature, satisfy Assumptions **P1** – **P3**. We give some examples of these covariance

functions that satisfy these assumptions and discuss some possible extensions, particularly for **P2** and **P3**, in Section 3.5.5. **P5** will be used also to ensure that the prior assigns the exponentially small probability on Θ_n^c .

To apply Theorem 3.3.1 to the nonparametric regression problem of (3.1), we use the following notation. The parameter θ in Theorem 3.3.1 is (η, σ) with $\theta_0 = (\eta_0, \sigma_0)$. We do not need to include β in the parameter θ for the following reason. The data need to be conditionally independent with known distribution given θ . In our regression model, the data are conditionally independent with known distribution given (η, σ) . For this reason we can integrate β out of the parameter for the purpose of applying Theorem 3.3.1. The density $f_i(\cdot; \theta)$ is the normal density with mean $\eta(x_i)$ and variance σ^2 or the double exponential density with location parameter $\eta(x_i)$ and scale parameter σ . The parameter space Θ is a product space of a function space Θ_1 and \mathbb{R}^+ . Let θ have prior Π , a product measure, $\Pi_1 \times \nu$, where Π_1 is a Gaussian process prior for η and ν is a prior for σ . A sieve, Θ_n is constructed to facilitate finding uniformly consistent tests, which will be described in Section 3.5.2. Finally, we define the sets that play the roles of \mathcal{N} in (3.2) (or more precisely U_n in Theorem 3.3.1) and contain θ_0 in terms of the various topologies that we will use, one for d_Q , one for L^1 , and one for Hellinger. In our theorems, these sets are the same for all n .

$$\begin{aligned} U_\epsilon &= \left\{ (\eta, \sigma) : d_Q(\eta, \eta_0) < \epsilon, \left| \frac{\sigma}{\sigma_0} - 1 \right| < \epsilon \right\}, \\ W_\epsilon &= \left\{ (\eta, \sigma) : \|\eta - \eta_0\|_1 < \epsilon, \left| \frac{\sigma}{\sigma_0} - 1 \right| < \epsilon \right\}, \\ H_\epsilon &= \{(\eta, \sigma) : d_H(f, f_0) < \epsilon\}. \end{aligned} \tag{3.12}$$

In the definition of H_ϵ , we make use of the fact that a parameter (η, σ) is equivalent to a joint density f for (X, Y) with respect to $\xi = Q \times \lambda$, namely $f(x, y) = \phi([y - \eta(x)]/\sigma)/\sigma$, where ϕ is the standard normal density or $f(x, y) = \psi([y - \eta(x)]/\sigma)/\sigma$, where ψ is the double exponential density with $DE(0, 1)$ or the standard Laplace density.

A point is in U_ϵ so long as σ is close to σ_0 and η differs greatly from η_0 only on a set

of small Q measure. It doesn't matter by how much η differs from η_0 on that set of small Q measure. Although U_ϵ is not necessarily open in any familiar topology, it does contain familiar open subsets. For each $1 \leq p \leq \infty$, U_ϵ contains $W_{\epsilon^{1+1/p}}$. For the cases in which the noise terms have normal or Laplace distributions, we will also show that for each ϵ there is a δ such that H_ϵ contains U_δ .

In addition, the marginal neighborhoods of η or σ contain the joint neighborhoods that we consider. Thus, if we prove that the posterior probability of joint neighborhoods converge almost surely to 1, then it obviously follows that the posterior probability of marginal neighborhoods converge almost surely to 1.

3.4.1 Main Theorems

The main theorems that we prove are the following, in which the data $\{Y_n\}_{n=1}^\infty$ are assumed to be conditionally independent with normal distributions or Laplace distributions given η , σ and the covariates. We assume that the true regression function, $\eta_0(x)$, is continuously differentiable. As in (3.12), we deal with three different neighborhoods of the parameter, θ under two different types of covariate, one from a random design and the other from a fixed design, under the Assumption RD or Assumption NRD. Since the Hellinger neighborhood, H_ϵ , is only defined for the random covariate, we provide five consistency results altogether. These are summarized in Table 3.1 and stated in detail afterwards.

Table 3.1: Summary of Consistency results

Neighborhood	Fixed design (NRD)	Random (RD)
W_ϵ	Theorem 3.4.1	Theorem 3.4.4
U_ϵ	Corollary 3.4.1	Theorem 3.4.2
H_ϵ	—	Theorem 3.4.3

Theorem 3.4.1. *Suppose that the values of the covariate in $[0, 1]$ arise according to a design satisfying Assumption NRD. Assume that the prior satisfies Assumption P. Let P_0 denote the joint conditional distribution of $\{Y_n\}_{n=1}^\infty$ assuming that (η_0, σ_0) is the true parameter. Assume that the function η_0 is continuously differentiable. Then for every $\epsilon > 0$,*

$$\Pi \{ W_\epsilon^C \mid Y_1, \dots, Y_n, x_1, \dots, x_n \} \rightarrow 0 \quad \text{a.s. } [P_0].$$

Corollary 3.4.1. *Under the same conditions as Theorem 3.4.1, for every $\epsilon > 0$, the posterior probability of each d_Q neighborhood, U_ϵ , converges to 1 a.s. $[P_0]$.*

Theorem 3.4.2. *Suppose that the values of the covariate in $[0, 1]$ arise according to a design satisfying Assumption RD. Assume that the prior satisfies Assumption P. Let P_0 denote the joint distribution of $\{(X_n, Y_n)\}_{n=1}^\infty$ assuming that (η_0, σ_0) is the true parameter. Assume that the function η_0 is continuously differentiable. Then for every $\epsilon > 0$,*

$$\Pi \{ U_\epsilon^C \mid (X_1, Y_1) \dots (X_n, Y_n) \} \rightarrow 0 \quad \text{a.s. } [P_0].$$

Theorem 3.4.3. *Suppose that the values of the covariate in $[0, 1]$ arise according to a design satisfying Assumption RD. Assume that the prior satisfies Assumption P. Let P_0 denote the joint distribution of $\{(X_n, Y_n)\}_{n=1}^\infty$ assuming that (η_0, σ_0) is the true parameter. Assume that the function η_0 is continuously differentiable. Then for every $\epsilon > 0$,*

$$\Pi \{ H_\epsilon^C \mid (X_1, Y_1), \dots, (X_n, Y_n) \} \rightarrow 0 \quad \text{a.s. } [P_0].$$

Finally, when we deal with random covariates, we can prove consistency of posterior probabilities of L^1 neighborhoods for the case in which the support of the prior distribution contains only uniformly bounded regression functions. Similar assumptions about the uniform boundness of regression functions can be also found in Shen and Wasserman (2001), Huang (2004) and Ghosal and Roy (2005).

Assumption B. Let Π'_1 and ν be a Gaussian process and a prior on σ satisfying Assumption P. Let $V_0 > 0$ be a constant. $\Omega = \{\eta : \|\eta\|_\infty < V_0\}$ with $V_0 > \|\eta_0\|_\infty$. Assume that $\Pi_1(\cdot) = \Pi'_1(\cdot \cap \Omega)/\Pi'_1(\Omega)$ with $\Pi'_1(\Omega) > 0$.

Under Assumption B, L^1 consistency is easily verified from the d_Q consistency or Hellinger consistency. We will provide detailed proofs in Section 3.5.4.

Theorem 3.4.4. *Suppose that the values of the covariate in $[0, 1]$ arise according to a design satisfying Assumption RD. Assume that the prior satisfies Assumption B. Let P_0 denote the joint distribution of $\{(X_n, Y_n)\}_{n=1}^\infty$ assuming that (η_0, σ_0) is the true parameter. Assume that the function η_0 is continuously differentiable. Then for every $\epsilon > 0$,*

$$\Pi \{ W_\epsilon^C | (X_1, Y_1), \dots, (X_n, Y_n) \} \rightarrow 0 \quad \text{a.s.}[P_0],$$

3.5 Verification

This section contains the proofs of the main consistency results. In the previous section, we stated several theorems with different conditions on the univariate covariate (random and nonrandom designs) and different topologies (L^1 , d_Q , and Hellinger). The proofs of these results all rely on Theorem 3.3.1, and thereby have many steps in common. Section 3.5.1 contains the proof of condition (A1) of Theorem 3.3.1, which is virtually the same for all of the main theorems. Section 3.5.2 shows how we construct the sieve that is used in condition (A2). We also verify subcondition (iii) in that section. In Section 3.5.3, we show how to construct uniformly consistent tests. This is done by piecing together finitely many tests, one

for each element of a covering of the sieve by L^∞ balls. This section contains two separate results concerning the spacing of design points in the random and nonrandom covariate cases (Lemmas 3.5.12 and 3.5.14). Section 3.5.4 explains why regression functions that are close in d_Q metric lead to joint distributions of (X, Y) that are close in Hellinger distance. This proves Theorem 3.4.3. Finally, we verify that Assumption B leads to consistency of posterior probabilities of L^1 neighborhoods in Section 3.5.4.

3.5.1 Prior Positivity Condition

In this subsection, we state and prove those results that allow us to verify condition (A1) of Theorem 3.3.1. Define $B = \left\{ (\eta, \sigma) : \|\eta - \eta_0\|_\infty < \epsilon, \left| \frac{\sigma}{\sigma_0} - 1 \right| < \epsilon \right\}$. Lemma 3.5.1 states it is sufficient to show that $\Pi(B) > 0$.

Lemma 3.5.1. *For each $\delta > 0$, define*

$$B_\delta = \left\{ (\eta, \sigma) : \|\eta - \eta_0\|_\infty < \delta, \left| \frac{\sigma}{\sigma_0} - 1 \right| < \delta \right\}.$$

Then

(i) *For every $\epsilon > 0$, there exists $\delta > 0$ such that $\forall \theta \in B_\delta, K_i(\theta_0, \theta) < \epsilon$ for all i*

$$(ii) \sum_{i=1}^{\infty} \frac{V_i(\theta_0, \theta)}{i^2} < \infty, \forall \theta \in B_\delta$$

Proof. We break the proof into two parts, which deal with normal and Laplace distributions separately. We also consider the nonrandom and random designs separately. and treat the nonrandom design case first.

- If $Y_i \sim N(\eta_0(x_i), \sigma_0^2)$

1. Nonrandom design:

$$\begin{aligned}
K_i(\theta_0; \theta) &= \mathbb{E}_{\theta_0}(\Lambda(\theta_0; \theta)) = \mathbb{E}_{\theta_0} \log \frac{f_i(Z_i; \theta_0)}{f_i(Z_i; \theta)} \\
&= \frac{1}{2} \log \frac{\sigma^2}{\sigma_0^2} + \mathbb{E}_{\theta_0} \left[-\frac{1}{2} \frac{(Y_i - \eta_0(x_i))^2}{\sigma_0^2} \right] - \mathbb{E}_{\theta_0} \left[-\frac{1}{2} \frac{(Y_i - \eta(x_i))^2}{\sigma^2} \right] \\
&= \frac{1}{2} \log \frac{\sigma^2}{\sigma_0^2} - \frac{1}{2} \left(1 - \frac{\sigma_0^2}{\sigma^2} \right) + \frac{1}{2} \frac{[\eta_0(x_i) - \eta(x_i)]^2}{\sigma^2}.
\end{aligned}$$

Take $a = \|\eta_0 - \eta\|_\infty$ and $b = \frac{\sigma}{\sigma_0}$.

$$K_i(\theta_0; \theta) \leq g_1(a, b),$$

where $g_1(a, b) = \log b - \frac{1}{2} \left(1 - \frac{1}{b^2} \right) + \frac{1}{2\sigma_0^2} \frac{a^2}{b^2}$ and $g_1(a, b)$ is clearly continuous at $a = 0$ and $b = 1$. Hence, for each $\epsilon > 0$, there exists δ such that for all $\theta \in B_\delta$, $K_i(\theta_0, \theta) \leq g_1(a, b) < \epsilon$ for all i .

Let $Z = [Y_i - \eta_0(x_i)]/\sigma_0$, which has standard normal distribution. Then

$$\begin{aligned}
V_i(\theta_0; \theta) &= \text{Var}_{\theta_0}(\Lambda(\theta_0; \theta)) \\
&= \text{Var}_{\theta_0} \left(-\left[\frac{(Y_i - \eta_0(x_i))^2}{2\sigma_0^2} \right] + \frac{1}{2} \left[\frac{\sigma_0}{\sigma} \frac{Y_i - \eta_0(x_i) + \eta_0(x_i) - \eta(x)}{\sigma_0} \right]^2 \right) \\
&= \text{Var} \left(\left[-\frac{1}{2} + \frac{1}{2} \frac{\sigma_0^2}{\sigma^2} \right] Z^2 + \frac{\sigma_0^2}{\sigma^2} [\eta(x_i) - \eta_0(x_i)] Z \right) \\
&= 2 \cdot \left[-\frac{1}{2} + \frac{1}{2} \frac{\sigma_0^2}{\sigma^2} \right]^2 + \left[\frac{\sigma_0^2}{\sigma^2} [\eta(x_i) - \eta_0(x_i)] \right]^2 \\
&< \infty, \text{ uniformly in } i.
\end{aligned}$$

- If $Y_i \sim DE(\eta_0(x_i), \sigma_0)$, similar calculation verifies

$$\begin{aligned}
K_{i,n}(\theta_0; \theta) &= \log \frac{\sigma}{\sigma_0} + \mathbb{E}_{\theta_0} \left[-\frac{|y - \eta_0(x_i)|}{\sigma_0} \right] - \mathbb{E}_{\theta_0} \left[-\frac{|y - \eta(x_i)|}{\sigma} \right] \\
&\leq \log \frac{\sigma}{\sigma_0} + \left| \frac{\sigma_0}{\sigma} \right| \left| 1 - \frac{\sigma}{\sigma_0} \right| + \left| \frac{\sigma_0}{\sigma} \right| \left(\frac{\|\eta_0(x_i) - \eta(x_i)\|_\infty}{\sigma_0} \right) \\
&\leq g_2(a, b),
\end{aligned}$$

where $g_2(a, b) = \log b + \left| \frac{1}{b} \right| |1 - b| + \left| \frac{1}{b} \right| \frac{a}{\sigma_0}$, which is also continuous at $a = 0$ and $b = 1$. Hence, for each $\epsilon > 0$, there exists δ such that for all $\theta \in B_\epsilon$, $K_i(\theta_0, \theta) \leq g_2(a, b) < \epsilon$

for all i .

$$\begin{aligned}
V_{i,n}(\theta_0; \theta) &= \text{Var}_{\theta_0} \left(-\left| \frac{y - \eta_0(x_i)}{\sigma_0} \right| + \left| \frac{y - \eta(x_i)}{\sigma} \right| \right) \\
&\leq \mathbb{E}_{\theta_0} \left(\left| \frac{y - \eta_0(x_i)}{\sigma_0} \right|^2 \right) + \mathbb{E}_{\theta_0} \left(\left| \frac{y - \eta(x_i)}{\sigma} \right|^2 \right) \\
&\leq 2 + \frac{\sigma_0^2}{\sigma^2} \left(1 + \frac{|\eta_0(x_i) - \eta(x_i)|^2}{\sigma_0^2} + 2 \frac{|\eta_0(x_i) - \eta(x_i)|}{\sigma_0^2} \right) \\
&< \infty, \text{ uniformly in } i.
\end{aligned}$$

It follows that $\sum_{i=1}^{\infty} \frac{V_i(\theta_0; \theta)}{i^2} < \infty$. Because these results hold uniformly in x_i , they hold for both the random and nonrandom design cases. ■

Lemma 3.5.2. *Assume that $\eta(x)$ is a Gaussian process with the mean function and the covariance function that satisfy Assumption P. Let*

$$\mathcal{A}_R = \left\{ \eta(x) = \sum_{i=1}^k a_i R(x, x_i^*; \beta) \text{ for some } \beta \in \mathbb{R}^+, a_1, \dots, a_k \in \mathbb{R}, x_i^* \in \mathbb{Q} \cap [0, 1] \right\},$$

where \mathbb{Q} is the set of rationals in \mathbb{R} . Let $\bar{\mathcal{A}}_R$ be the closure of \mathcal{A}_R in the supremum metric.

Under Assumption P, if $\eta_0 \in \bar{\mathcal{A}}_R$, then

$$P \left(\sup_{x \in [0,1]} |\eta(x) - \eta_0(x)| < \epsilon \right) > 0$$

Proof. The proof follows from Tokdar and Ghosh(2004, Theorem 3.4) or Ghosal and Roy(2005, Theorem 4) (See Appendices). ■

Lemma 3.5.3. *Let $\epsilon > 0$ and define*

$$B = \left\{ (\eta, \sigma) : \|\eta - \eta_0\|_{\infty} < \epsilon, \left| \frac{\sigma}{\sigma_0} - 1 \right| < \epsilon \right\}.$$

Then, $\Pi(B) > 0$ under Assumption P.

Proof. Under Assumption P, we know that $\nu \left\{ \left| \frac{\sigma}{\sigma_0} - 1 \right| < \epsilon \right\} > 0$. Thus, to verify $\Pi(B) > 0$, it suffices to show that $\Pi_1(\eta : \|\eta - \eta_0\|_\infty < \epsilon) > 0$, and this was ensured by Lemma 3.5.2. The only concern to apply Lemma 3.5.2 is to check whether the true regression function $\eta_0(\cdot)$ is in $\bar{\mathcal{A}}_R$, and it can be shown that the every continuously differentiable function is contained in $\bar{\mathcal{A}}_R$ under Assumption P. The assertion that the true function η_0 is in $\bar{\mathcal{A}}_R$ will be mentioned in detail in Section 3.5.5, following the approach by Tokdar and Ghosh (2004).

A simple corollary of Lemma 3.5.3 is that, in Assumption B, $\Pi'_1(\Omega) > 0$, so that Π_1 is well-defined. Also, it is clear that Π_1 in Assumption B also satisfies the conclusion of Lemma 3.5.3. ■

3.5.2 Construction of Sieves

To verify (A2) of Theorem 3.3.1, we first construct a sieve and then construct a test for each element of the sieve.

Let $M_n = O(n^{\alpha_1})$, where $\frac{2\delta + 1}{2} < \alpha_1 < 1$ for some $0 < \delta < 1/2$ and define $\Theta_n = \Theta_{1n} \times \mathbb{R}^+$, where

$$\Theta_{1n} = \{\eta(\cdot) : \|\eta\|_\infty < M_n, \|\eta'\|_\infty < M_n\}. \quad (3.13)$$

The n th test is constructed by combining a collection of tests, one for each of finitely many elements of Θ_n . Those finitely many elements come from a covering of Θ_{1n} by small balls. The following lemma is straightforward from Theorem 2.7.1 of van der Vaart and Wellner (1996) or Lemma 2.3 of van de Geer (2000).

Lemma 3.5.4. *There exists a constant K' such that the ϵ -covering number $N(\epsilon, \Theta_{1n}, \|\cdot\|_\infty)$ of Θ_{1n} in the supremum norm satisfies*

$$\log N(\epsilon, \Theta_{1n}, \|\cdot\|_\infty) \leq \frac{K' M_n}{\epsilon}.$$

Proof. The proof follows from Theorem 2.7.1. of van der Vaart and Wellner (1996) or from Lemma 2.3 of van de Geer (2000). We choose the former approach.

According to Theorem 2.7.1. of van der Vaart and Wellner (1996), let \mathcal{X} be a bounded convex subset of \mathbb{R} with nonempty interior and let $C_1^1(\mathcal{X})$ be the set of all continuous functions $f : \mathcal{X} \mapsto \mathbb{R}$ with $\|f\|_1 \equiv \sup_x |f(x)| + \sup_{x,y} \frac{|f(x) - f(y)|}{|x - y|} \leq 1$. Let $\mathcal{X}_1 = \{x : \|x - \mathcal{X}\| < 1\}$. Then, there exists a constant K such that

$$\log N(\epsilon, C_1^1(\mathcal{X}), \|\cdot\|_\infty) \leq K\lambda(\mathcal{X}_1) \left(\frac{1}{\epsilon}\right),$$

for every $\epsilon > 0$.

To complete the proof, we replace $f(x)$ with $\eta(x)/2M_n$ when $\eta(x) \in \Theta_n$, then

$$\begin{aligned} \|f\|_1 &= \sup_x |f(x)| + \sup_{x,y} \frac{|f(x) - f(y)|}{|x - y|} \\ &= \sup_x \left| \frac{\eta(x)}{2M_n} \right| + \sup_{x,y} \frac{|\eta(x) - \eta(y)|}{2M_n|x - y|} \\ &\leq \frac{1}{2} + \frac{\sup_x |\eta'(x)|}{2M_n} \leq 1 \end{aligned}$$

In addition, the ϵ -covering number for η is identical to the $\epsilon/2M_n$ -covering number for f and \mathcal{X} is equal to the bounded interval, $[0, 1]$.

Therefore,

$$\log N(\epsilon, \Theta_n, \|\cdot\|_\infty) \leq K' \left(\frac{M_n}{\epsilon}\right),$$

with a constant $K' > 0$. ■

For the proof of subcondition (iii) of (A2), we make use of the assumed smoothness in (P3) of Assumption P. Lemma 3.5.5 below shows that, under Assumption P, the sample paths of the Gaussian process are almost surely continuously differentiable and the first derivative process is also Gaussian. Furthermore, the probability of being outside of the sieve becomes exponentially small.

Lemma 3.5.5. *Given the smoothing parameter, β , let $\eta(x)$ be a mean zero Gaussian process on $[0, 1]$ with a covariance function, $\text{Cov}\{\eta(x), \eta(x')\} = R_0(\beta|x - x'|)$ which satisfies Assumption P. Then $\eta(x)$ has continuously differentiable sample paths and the first derivative process $\eta'(x)$ is also a Gaussian process with a covariance function, $\text{Cov}\{\eta'(x), \eta'(x')\} = -\beta^2 R_0''(\beta|x - x'|)$. In particular, the variance of $\eta'(x)$, $\text{Var}(\eta'(x))$, is given as $-\beta^2 R_0''(0)$.*

Proof. This is the special case ($d = 1$, $w = 1$) of the result of Ghosal and Roy (2005, Theorem 5). ■

Lemma 3.5.6. *Let $R(x, x'; \beta)$ be the covariance function of $\eta(x)$. Define $\sigma_{\beta, R}$, the supremum of the standard deviation of $\eta(x)$, as $\left\{ \sup_{x \in [0, 1]} R(x, x; \beta) \right\}^{1/2}$. Supposed that given β , $\eta(x)$ is a separable, mean-zero Gaussian process such that for some $K > \sigma_{\beta, R}$, and some $0 < \epsilon_0 < \sigma_{\beta, R}$,*

$$N(\epsilon, [0, 1], |\cdot|) \leq \frac{K}{\epsilon}, \quad 0 < \epsilon < \epsilon_0$$

Then, there exists a constant $D_{\beta, R}$ such that, for all $M \geq 2\sigma_{\beta, R}^2/\epsilon_0$,

$$P\left(\sup_{x \in [0, 1]} \eta(x) \geq M \middle| \beta\right) \leq \left(\frac{D_{\beta, R} K M}{\sigma_{\beta, R}^2}\right) \bar{\Phi}\left(\frac{M}{\sigma_{\beta, R}}\right) \quad (3.14)$$

In addition, if $M > 16D_{\beta, R}K$, then

$$P\left(\sup_{x \in [0, 1]} |\eta(x)| \geq M \middle| \beta\right) \leq \exp\left(-\frac{1}{4} \frac{M^2}{\sigma_{\beta, R}^2}\right). \quad (3.15)$$

Proof. First, (3.14) is a straightforward result from Proposition A.2.7 of van der Vaart and Wellner (1996). In order to obtain the inequality of (3.15), we apply Mill's ratio to (3.14) as follows: Note that for all $M \geq 2\sigma_{\beta, R}^2/\epsilon_0$, M is also bigger than $2\sigma_{\beta, R}$ as below,

$$\frac{1}{\epsilon_0} \geq \frac{1}{\sigma_{\beta, R}} \Rightarrow 2\sigma_{\beta, R}^2/\epsilon_0 \geq 2\sigma_{\beta, R} \Rightarrow M > 2\sigma_{\beta, R}.$$

Thus, if $M > 16D_{\beta,R}K$,

$$\begin{aligned}
\Pr \left(\sup_{x \in [0,1]} |\eta(x)| > M \middle| \beta \right) &\leq 2 \Pr \left(\sup_{x \in [0,1]} \eta(x) > M \middle| \beta \right) \leq 2 \left(\frac{D_{\beta,R}KM}{\sigma_{\beta,R}^2} \right) \bar{\Phi}(M/\sigma_{\beta,R}) \\
&\leq 2 \left(\frac{D_{\beta,R}KM}{\sigma_{\beta,R}^2} \right) \frac{\sigma_{\beta,R}}{M} \exp \left(-\frac{1}{2} \frac{M^2}{\sigma_{\beta,R}^2} \right), \quad \text{by Mill's ratio} \\
&\leq \left(\frac{M}{8\sigma_{\beta,R}} \right) \exp \left(-\frac{1}{2} \frac{M^2}{\sigma_{\beta,R}^2} \right), \quad \text{by the assumption} \\
&= \exp \left(\log \frac{M}{8\sigma_{\beta,R}} \right) \exp \left(-\frac{1}{2} \frac{M^2}{\sigma_{\beta,R}^2} \right), \quad x = \exp(\log x) \\
&\leq \exp \left(\frac{M}{8\sigma_{\beta,R}} - \frac{1}{2} \frac{M^2}{\sigma_{\beta,R}^2} \right), \\
&\leq \exp \left(-\frac{1}{4} \frac{M^2}{\sigma_{\beta,R}^2} \right),
\end{aligned}$$

since

$$\frac{M}{\sigma_{\beta,R}} \geq 1/2 \quad \text{and} \quad 1/8x - 1/2x^2 \leq -1/4x^2 \quad \forall x \leq 0 \text{ or } x \geq 1/2$$

■

Lemma 3.5.7. *Given $0 < \delta < 1/2$, let $\beta^2 \leq n^{2\delta}$ and $M_n = O(n^{\alpha_1})$ for $\frac{2\delta+1}{2} < \alpha_1 < 1$.*

Then, under the same condition as in Lemma 3.5.5, there exist constants d_1 and d_2 , independent of β , such that

$$\Pr \left\{ \sup_{0 \leq x \leq 1} |\eta(x)| > M_n \middle| \beta \right\} \leq \exp(-d_1 M_n^2) \quad (3.16)$$

$$\Pr \left\{ \sup_{0 \leq x \leq 1} |\eta'(x)| > M_n \middle| \beta \right\} \leq \exp \left(-d_2 \frac{M_n^2}{n^{\alpha_2}} \right) \quad (3.17)$$

Proof. Note that the mean function $\mu(x)$ can be fixed for all $\beta \in B$, and thus it can be bounded because of the continuity on the compact set. Thus, without loss of generality we can assume the process to have zero mean. Otherwise

$$\begin{aligned}
\Pr \{ \eta : \|\eta\|_\infty > M \mid \beta \} &\leq \Pr \{ \eta : \|\eta(\cdot) - \mu(\cdot)\|_\infty > M - \|\mu\|_\infty \mid \beta \} \\
&\leq \Pr \{ \eta : \|\eta(\cdot) - \mu(\cdot)\|_\infty > M/2 \mid \beta \}.
\end{aligned}$$

To obtain the exponentially small bounds in (3.16) and (3.17), we apply Lemma 3.5.6 to the Gaussian processes, $\eta(x)$ and $\eta'(x)$ and thus we need to consider two covering numbers, $N(\epsilon, [0, 1], \rho_1)$, where ρ_1 and ρ_2 are “natural semimetrics” defined by $\rho_1(s, t) = [\mathbb{E}(\eta(s) - \eta(t))^2]^{1/2}$ and $\rho_2(s, t) = [\mathbb{E}(\eta'(s) - \eta'(t))^2]^{1/2}$.

Note that the squared intrinsic semimetric for $\eta(x)$ is given by

$$\begin{aligned} \mathbb{E}(\eta(x) - \eta(x'))^2 &= R(x, x; \beta) - 2R(x, x'; \beta) + R(x', x'; \beta) \\ &= 2 \{R_0(0) - R_0(\beta|x - x'|)\} \\ &\leq 2\beta^2|x - x'|^2 \sup_{x \in [0, \beta]} |R_0''(x)|, \end{aligned}$$

which leads to

$$[\mathbb{E}(\eta(x) - \eta(x'))^2]^{1/2} \leq \beta|x - x'|A_1 \quad (3.18)$$

for some positive constant A_1 . In addition, it can be also shown easily that there exists a positive constant A_2 such that

$$\begin{aligned} [\mathbb{E} \{ \eta'(x) - \eta'(x') \}^2]^{1/2} &= \left[\lim_{h \rightarrow 0} \mathbb{E} \{ \eta(x+h) - \eta(x) - \eta(x'+h) + \eta(x') \}^2 / h^2 \right]^{1/2} \\ &\leq \beta^2 A_2 |x - x'|. \end{aligned} \quad (3.19)$$

Therefore, from (3.18) and (3.19), we have that there exist positive constants A_1 and A_2 such that

$$\rho_1(s, t) \leq \beta|s - t|A_1 \quad \text{and} \quad \rho_2(s, t) \leq \beta^2|s - t|A_2$$

Hence, there exist K_1 and K_2 such that

$$N(\epsilon, [0, 1], \rho_1) \leq \left(\frac{K_1 \beta}{\epsilon} \right) \quad \text{and} \quad N(\epsilon, [0, 1], \rho_2) \leq \left(\frac{K_2 \beta^2}{\epsilon} \right)$$

In addition, it follows that for sufficiently large n , $M_n > \max\{2\sigma_{\beta, R}, 16D_{\beta, R}\beta K_1\}$ and $M_n > \max\{2\sigma_{\beta, R^*}(X), 16D_{\beta, R^*}\beta^2 K_2\}$ since $M_n = O(n^{\alpha_1})$ for $1/2 < \alpha_1 < 1$, $\beta^2 \leq n^{\alpha_2} < n^{\alpha_1}$ and other quantities are finite constants by Assumption P, where R^* stands for the covariance

function of $\eta'(x)$. Accordingly, by (3.15) in Lemman 3.5.6, we have

$$\Pr \left\{ \sup_{x \in [0,1]} |\eta(x)| > M_n \middle| \beta \right\} \leq \exp \left(-\frac{M_n^2}{4\sigma_{\beta,R}^2} \right)$$

and

$$\Pr \left\{ \sup_{x \in [0,1]} |\eta(x)'| > M_n \middle| \beta \right\} \leq \exp \left(-\frac{M_n^2}{4\sigma_{\beta,R^*}^2} \right)$$

In addition, we have $\sigma_{\beta,R}^2 = \sup_{x \in [0,1]} R(x, x; \beta) = R_0(0)$ and $\sigma_{\beta,R^*}^2 = -\beta^2 R_0''(0)$. Hence,

$$\begin{aligned} \Pr \left\{ \sup_{x \in [0,1]} |\eta(x)| > M_n \middle| \beta \right\} &\leq \exp \left(-\frac{M_n^2}{4R_0(0)} \right) \\ \Pr \left\{ \sup_{x \in [0,1]} |\eta'(x)| > M_n \middle| \beta \right\} &\leq \exp \left(\frac{M_n^2}{4\beta^2 R_0''(0)} \right) \leq \exp \left(\frac{M_n^2}{4n^{2\delta} R_0''(0)} \right). \end{aligned}$$

Then, if we choose

$$d_1 = \frac{1}{4R_0(0)} \quad \text{and} \quad d_2 = -\frac{1}{4R_0''(0)},$$

then d_1 and d_2 are positive constants, which are independent of β . ■

Lemma 3.5.8. *Suppose that Assumption P5 holds. Then, there exist positive constants d_3 and d_4 such that*

$$\Pi(\Theta_n^C) \leq d_3 \exp(-d_4 n). \quad (3.20)$$

Proof. Note that

$$\begin{aligned} \Pi(\Theta_n^C | \beta) &= \Pi \{ (\Theta_{1n}^C \times R^+) | \beta \} = \Pi_1(\Theta_{1n}^C | \beta) \times \nu(R^+) \\ &= \Pi_1(\Theta_{1n}^C | \beta) \\ &\leq \Pr \left\{ \sup_{x \in [0,1]} |\eta(x)| > M_n \middle| \beta \right\} + \Pr \left\{ \sup_{x \in [0,1]} |\eta'(x)| > M_n \middle| \beta \right\} \end{aligned}$$

Therefore, we consider the marginal probabilities, $\Pr \left\{ \sup_{x \in [0,1]} |\eta(x)| > M_n \right\}$ and $\Pr \left\{ \sup_{x \in [0,1]} |\eta'(x)| > M_n \right\}$. Note that the bound for the conditional probability of $\eta(x)$ and $\eta'(x)$ given β is independent

of β if $\beta^2 \leq n^{2\delta}$, as in (3.16) and (3.17). Since the marginal probability can be obtained by integrating β out of the conditional probability, it follows that

$$\begin{aligned} \Pr \left\{ \sup_{x \in [0,1]} |\eta(x)| > M_n \right\} &= \int_{\mathbb{R}^+} \Pr \left\{ \sup_{x \in [0,1]} |\eta(x)| > M_n \middle| \beta \right\} d\Pi_\beta \\ &\leq \exp(-d_1 M_n^2) \leq \exp(-d_1 n) \end{aligned}$$

and, by Assumption P5 that there exist b_1 and b_2 such that $\Pr(\beta > n^{2\delta}) \leq b_1 \exp(-b_2 n)$, it also follows that

$$\begin{aligned} \Pr \left\{ \sup_{x \in [0,1]} |\eta'(x)| > M_n \right\} &= \int_{\mathbb{R}^+} \Pr \left\{ \sup_{x \in [0,1]} |\eta'(x)| > M_n \middle| \beta \right\} d\Pi_\beta \\ &\leq \int_0^{n^{2\delta}} \Pr \left\{ \sup_{x \in [0,1]} |\eta'(x)| > M_n \middle| \beta \right\} d\Pi_\beta + \Pr(\beta > n^{2\delta}) \\ &\leq \exp\left(-d_2 \frac{M_n^2}{n^{2\delta}}\right) + b_1 \exp(-b_2 n) \\ &\leq \exp\left(-d_2 n^{2\alpha_1 - 2\delta}\right) + b_1 \exp(-b_2 n) \leq d_3 \exp(-d_4 n), \end{aligned}$$

since $2\alpha_1 - 2\delta \geq 1$. Hence, the proof is complete. \blacksquare

3.5.3 Existence of Tests

For each n and each ball in the covering of Θ_{1n} , we find a test with small type I and type II error probabilities. Then we combine the tests and show that they satisfy subconditions (i) and (ii) of (A2). The following relatively straightforward result is useful in the construction.

Proposition 3.5.1. (a) *For every random variable X with unimodal distribution symmetric around 0 and every $c \in \mathbb{R}$,*

$$\Pr(|X| \leq x) \geq \Pr(|X + c| \leq x).$$

This is the special case of Anderson's theorem (Anderson, 1955).

(b) Let X_1, \dots, X_n be i.i.d random variables satisfying (a). Let $b_i \in \mathbb{R}$, $i = 1, \dots, n$.

Then, $\forall c \in \mathbb{R}$,

$$\Pr \left(\sum_{i=1}^n |X_i| \leq c \right) \leq \Pr \left(\sum_{i=1}^n |X_i + b_i| \leq c \right).$$

Proof. (a)

$$\begin{aligned} & \Pr \{ |X| \leq x \} - \Pr \{ |X + c| \leq x \} \\ &= F(x) - F(-x) - F(x - c) + F(-x - c) \\ &= \begin{cases} \{F(x) - F(x - c)\} - \{F(x + c) - F(x)\} \geq 0, & c \geq 0 \\ \{F(x) - F(x + c)\} - \{F(x - c) - F(x)\} \geq 0, & c < 0 \end{cases} \end{aligned}$$

(b) We argue by induction and use the law of total probability.

$$\begin{aligned} \Pr \left\{ \sum_{i=1}^n |X_i| \leq c \right\} &= \mathbb{E} \left(\Pr \left\{ |X_1| + \sum_{i=2}^n |X_i| \leq c \mid |X_2| = x_2, \dots, |X_n| = x_n \right\} \right) \\ &\leq \mathbb{E} \left(\Pr \left\{ |X_1 + b_1| \leq c - \sum_{i=2}^n x_i \mid |X_2| = x_2, \dots, |X_n| = x_n \right\} \right) \\ &= \Pr \left\{ |X_1 + b_1| + \sum_{i=2}^n |X_i| \leq c \right\} \end{aligned}$$

Using the same argument as above but conditioning on $(|X_1 + b_1|, |X_3|, \dots, |X_n|)$, we get

$$\Pr \left\{ \sum_{i=1}^n |X_i| \leq c \right\} \leq \Pr \left\{ |X_1 + b_1| + |X_2 + b_2| + \sum_{i=3}^n |X_i| \leq c \right\}$$

Similarly, by conditioning on $(|X_1 + b_1|, \dots, |X_{i-1} + b_{i-1}|, |X_{i+1}|, \dots, |X_n|)$, $i = 3, \dots, n$, we reach the final result of part (b). ■

The main part of test construction is contained in the following Lemmas 3.5.9–3.5.11. For the random design cases, we first condition on the observed values of the covariate. In Lemmas 3.5.9–3.5.11, understand all probability statements as conditional on the covariate values $X_1 = x_1, \dots, X_n = x_n$ in the random design case.

Lemma 3.5.9. *Let η_1 be a continuous function on $[0, 1]$ and define $\eta_{ij} = \eta_i(x_j)$ for $i = 0, 1$ and $j = 1, \dots, n$. Let $\epsilon > 0$, and let $r > 0$. Let $c_n = n^{\tau_1}$ for $\alpha_1/2 < \tau_1 < 1/2$ and $1/2 < \alpha_1 < 1$. Let $b_j = 1$ if $\eta_{1j} \geq \eta_{0j}$ and -1 otherwise. Let $\Psi_{1n}[\eta_1, \epsilon]$ be the indicator of the set A_1 , where A_1 is defined as*

1. If $Y_j \sim N(\eta_{0j}, \sigma_0^2)$

$$A_1 = \left\{ \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{0j}}{\sigma_0} \right) > 2c_n \sqrt{n} \right\},$$

2. If $Y_j \sim DE(\eta_{0j}, \sigma_0)$

$$A_1 = \left\{ \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{0j}}{\sigma_0} \right) > 2c_n \sqrt{n} \right\},$$

Then there exists a constant C_3 such that for all η_1 that satisfy

$$\sum_{j=1}^n |\eta_{1j} - \eta_{0j}| > rn, \quad (3.21)$$

$E_{P_0}(\Psi_{1n}[\eta_1, \epsilon]) < C_3 \exp(-2c_n^2)$. Also, there exist constants C_4 and C_5 such that for all sufficiently large n and all η satisfying $\|\eta - \eta_1\|_\infty < r/4$ and for all $\sigma \leq \sigma_0(1 + \epsilon)$,

$$E_P(1 - \Psi_{1n}[\eta_1, \epsilon]) \leq C_4 \exp(-nC_5),$$

where P is the joint distribution of $\{Y_n\}_{n=1}^\infty$ assuming that $\theta = (\eta, \sigma)$.

Proof. 1. Normal data:

(1) Type I error:

By Mill's ratio, we have

$$\begin{aligned} E_{P_0}(\Psi_{1n}) &= P_0 \left\{ \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{0j}}{\sigma_0} \right) > 2c_n \sqrt{n} \right\} \\ &= P_0 \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{0j}}{\sigma_0} \right) > 2c_n \right\} \\ &= 1 - \Phi(2c_n) \leq \frac{\phi(2c_n)}{2c_n} = \frac{1}{2\sqrt{2\pi}} \frac{\exp(-2c_n^2)}{c_n}. \end{aligned}$$

(2) Type II error:

Assume n is large enough so that $c_n/\sqrt{n} < r/(4\sigma_0)$. Let $\eta_{*j} = \eta(x_j)$ for $j = 1, \dots, n$.

Since $\sigma < (1 + \epsilon)\sigma_0$,

$$\begin{aligned}
E_P(1 - \Psi_{1n}[\eta_1, \epsilon]) &= P \left\{ \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{0j}}{\sigma_0} \right) \leq 2c_n \sqrt{n} \right\} \\
&= P \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{*j}}{\sigma} \right) + \frac{1}{\sqrt{n}} \sum_{j=1}^n b_j \left(\frac{\eta_{*j} - \eta_{1j}}{\sigma} \right) + \frac{1}{\sqrt{n}} \sum_{j=1}^n \left| \frac{\eta_{1j} - \eta_{0j}}{\sigma} \right| \leq 2c_n \frac{\sigma_0}{\sigma} \right\} \\
&\leq P \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{*j}}{\sigma} \right) \leq \frac{r\sqrt{n}}{4\sigma} - \frac{r\sqrt{n}}{\sigma} + 2c_n \frac{\sigma_0}{\sigma} \right\} \\
&\leq P \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{*j}}{\sigma} \right) \leq -\frac{r\sqrt{n}}{4\sigma_0(1 + \epsilon)} \right\} \\
&= \Phi \left(-\frac{r\sqrt{n}}{4\sigma_0(1 + \epsilon)} \right) \leq \frac{4\sigma_0(1 + \epsilon)}{r\sqrt{2\pi n}} \exp \left(-\frac{nr^2}{32\sigma_0^2(1 + \epsilon)^2} \right),
\end{aligned}$$

where the last inequality is by Mill's ratio.

2. Laplace data:

(1) Type I error:

For all $0 < t < 1$, by the Markov inequality,

$$\begin{aligned}
E_{P_0}(\Psi_{1n}[\eta_1, \epsilon]) &= P_0 \left\{ \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{0j}}{\sigma_0} \right) > 2c_n \sqrt{n} \right\} \\
&= P_0 \left\{ t \cdot \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{0j}}{\sigma_0} \right) > t \cdot 2c_n \sqrt{n} \right\} \\
&\leq \exp(-t \cdot 2c_n \sqrt{n}) (1 - t^2)^{-n} \\
&= \exp \left(\left[-2t \frac{c_n}{\sqrt{n}} + \log \left(1 + \frac{t^2}{1 - t^2} \right) \right] \cdot n \right) \\
&\leq \exp \left(\left[-2t \frac{c_n}{\sqrt{n}} + \frac{t^2}{1 - t^2} \right] \cdot n \right)
\end{aligned}$$

Take $t = \frac{2c_n}{\sqrt{n}}$. Then,

$$\begin{aligned} E_{P_0}(\Psi_{1n}) &= \exp \left(\left[-4 \frac{c_n^2}{n} + \frac{\frac{4c_n^2}{n}}{1 - \frac{4c_n^2}{n}} \right] \cdot n \right) \\ &\leq \exp \left(\left[-4 + \frac{1}{\frac{n}{4c_n^2} - 1} \right] \cdot c_n^2 \right) \leq \exp(-2c_n^2) \end{aligned}$$

The last inequality is obvious because $\frac{4c_n^2}{n} < \frac{2}{3}$ for sufficiently large n .

(2) Type II error:

As in the Type II error calculation for the normal case, first, assume that n is large enough so that $c_n/\sqrt{n} < r/(4\sigma_0)$. Let $\eta_{*j} = \eta(x_j)$ for $j = 1, \dots, n$. Since $\sigma \leq (1 + \epsilon)\sigma_0$, then for all $-1 < t < 0$,

$$\begin{aligned} E_P(1 - \Psi_n[\eta_1, \epsilon]) &\leq E_P(1 - \Psi_{1n}) = P \left\{ \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{0j}}{\sigma_0} \right) \leq 2c_n \sqrt{n} \right\} \\ &= P \left\{ \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{*j}}{\sigma_1} \right) + \sum_{j=1}^n b_j \left(\frac{\eta_{*j} - \eta_{1j}}{\sigma} \right) + \sum_{j=1}^n \left| \frac{\eta_{1j} - \eta_{0j}}{\sigma} \right| \leq 2c_n \sqrt{n} \frac{\sigma_0}{\sigma} \right\} \\ &\leq P \left\{ \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{*j}}{\sigma} \right) \leq \frac{rn}{4\sigma} - \frac{rn}{\sigma} + 2c_n \sqrt{n} \frac{\sigma_0}{\sigma} \right\} \\ &\leq P \left\{ \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{*j}}{\sigma} \right) \leq \frac{-rn}{4\sigma_0(1 + \epsilon)} \right\} = P \left\{ t \sum_{j=1}^n b_j \left(\frac{Y_j - \eta_{*j}}{\sigma} \right) \geq t \frac{-rn}{4\sigma_0(1 + \epsilon)} \right\} \\ &\leq \exp \left(-t \frac{-rn}{4\sigma_0(1 + \epsilon)} \right) (1 - t^2)^{-n} \leq \exp \left(n \left[\frac{tr}{4\sigma_0(1 + \epsilon)} + \frac{t^2}{1 - t^2} \right] \right) \\ &= \exp \left(nt \left[\frac{r}{4\sigma_0(1 + \epsilon)} + \frac{t}{1 - t^2} \right] \right) \end{aligned} \tag{3.22}$$

Let $c^* = \frac{r}{4\sigma_0(1 + \epsilon)}$. Then, there exists t^* such that $-1 < \frac{1 - \sqrt{1 + 4(c^*)^2}}{2c^*} < t^* < 0$ and

$$-t^* \left(c + \frac{t^*}{1 - (t^*)^2} \right) = C^* > 0.$$

Thus, we have (3.22) is less than $\exp(-nC^*)$.

■

Lemma 3.5.10. *Let η_1 be a continuous function on $[0, 1]$ and define $\eta_{ij} = \eta_i(x_j)$ for $i = 0, 1$ and $j = 1, \dots, n$. Let Ψ_{2n} be the indicator of the set A_2 , where*

1. *If $Y_j \sim N(\eta_{0j}, \sigma_0^2)$*

$$A_2 = \left\{ \sum_{j=1}^n \frac{(Y_j - \eta_{0j})^2}{\sigma_0^2} > n(1 + \epsilon) \right\},$$

2. *If $Y_j \sim DE(\eta_{0j}, \sigma_0)$*

$$A_2 = \left\{ \sum_{j=1}^n \left| \frac{Y_j - \eta_{0j}}{\sigma_0} \right| > n\sqrt{1 + \epsilon} \right\},$$

Then there exists a constant C_6 such that for all η_1 $E_{P_0}(\Psi_{2n}) < \exp(-nC_6)$. Also, there exist constants C_7 such that for all sufficiently large n and all η satisfying $\|\eta - \eta_1\|_\infty < r/4$ and for all $\sigma > \sigma_0(1 + \epsilon)$,

$$E_P(1 - \Psi_{2n}[\eta_1, \epsilon]) \leq \exp(-nC_7),$$

where P is the joint distribution of $\{Y_n\}_{n=1}^\infty$ assuming that $\theta = (\eta, \sigma)$.

Proof. 1. Normal data:

(1) Type I error:

Let $W \sim \chi_n^2$. Then, for all $0 < t_1 < 1/2$,

$$\begin{aligned} E_{P_0}(\Psi_{2n}) &= P_0 \left(\sum_{j=1}^n \left(\frac{Y_j - \eta_{0j}}{\sigma_0} \right)^2 > n(1 + \epsilon) \right) \\ &= \Pr(W > n(1 + \epsilon)) \\ &\leq \exp(-n(1 + \epsilon)t_1) E(\exp(t_1 W)) \\ &= \exp(-n(1 + \epsilon)t_1) (1 - 2t_1)^{-n/2}. \end{aligned}$$

Take

$$t_1 = \frac{1}{2} \left(1 - \frac{1}{1 + \epsilon} \right).$$

Then,

$$E_{P_0}(\Psi_{2n}) \leq \exp\left(-\frac{n\epsilon}{2} + \frac{n}{2} \log[1 + \epsilon]\right) \leq \exp\left(-n\left[\frac{\epsilon^2}{4} - \frac{\epsilon^3}{6}\right]\right),$$

where the last line follows from the fact that $\log(1 + x) \leq x - x^2/2 + x^3/3$, $x > 0$.

Therefore, $E_{P_0}(\Psi_{2n}) \leq \exp(-nC_6)$.

(2) Type II error:

Let $W \sim \chi_n^2$ and let W' have a noncentral χ^2 distribution with n degrees of freedom and noncentrality parameter $\frac{\sum_{j=1}^n (\eta_{*j} - \eta_{0j})^2}{\sigma^2}$. Then, since $\sigma > (1 + \epsilon)\sigma_0$, for all $t < 0$,

$$\begin{aligned} E_P(1 - \Psi_{2n}) &= P\left\{\sum_{j=1}^n \left(\frac{Y_j - \eta_{0j}}{\sigma_0}\right)^2 \leq n[1 + \epsilon]\right\} \\ &= P\left\{\sum_{j=1}^n \left(\frac{Y_j - \eta_{0j}}{\sigma}\right)^2 \frac{\sigma^2}{\sigma_0^2} \leq n[1 + \epsilon]\right\} = \Pr\left(W' \leq n\frac{\sigma_0^2}{\sigma^2}[1 + \epsilon]\right) \\ &\leq \Pr\left(W \leq n\frac{\sigma_0^2}{\sigma^2}[1 + \epsilon]\right) \leq \Pr\left(W \leq \frac{n}{1 + \epsilon}\right) = \Pr\left\{\exp(Wt) \geq \exp\left(\frac{nt}{1 + \epsilon}\right)\right\} \\ &\leq \exp\left(-\frac{nt}{1 + \epsilon}\right) (1 - 2t)^{-n/2}. \end{aligned}$$

Let $t = -\epsilon/2$ to get

$$E_P(1 - \Psi_n[\eta_1, \epsilon]) \leq \exp\left(\frac{n}{2} \left[\frac{\epsilon}{1 + \epsilon} - \log(1 + \epsilon)\right]\right) \leq \exp\left(-n\frac{\epsilon^2 - \epsilon^3}{4(1 + \epsilon)}\right),$$

where the last inequality follows from the fact that $\log(1 + x) > x - x^2/2$.

2. Laplace data:

(1) Type I error:

Let $V \sim \text{Gamma}(n, 1)$. Then, for all $0 < t_1 < 1$,

$$\begin{aligned} E_{P_0}(\Psi_{2n}) &= P_0\left(\sum_{j=1}^n \left|\frac{Y_j - \eta_{0j}}{\sigma_0}\right| > n\sqrt{1 + \epsilon}\right) \\ &= \Pr(V > n\sqrt{1 + \epsilon}) \leq \exp(-n(\sqrt{1 + \epsilon})t_1) E(\exp(t_1 V)) \\ &= \exp(-n(\sqrt{1 + \epsilon})t_1) (1 - t_1)^{-n}. \end{aligned}$$

Take

$$t_1 = 1 - \frac{1}{\sqrt{1+\epsilon}}.$$

Then,

$$\begin{aligned} \mathbb{E}_{P_0}(\Psi_{2n}) &\leq \exp(-n(\sqrt{1+\epsilon}-1) + n \log[1 + \sqrt{1+\epsilon}-1]) \\ &\leq \exp\left(-n\left[\frac{(\sqrt{1+\epsilon}-1)^2}{2} - \frac{(\sqrt{1+\epsilon}-1)^3}{3}\right]\right), \end{aligned}$$

where the last inequality follows from the fact that $\log(1+x) \leq x - x^2/2 + x^3/3$, $\forall x > 0$.

Therefore, $\mathbb{E}_{P_0}(\Psi_n) \leq \exp(-C_6 n)$.

(2) Type II error:

Let $V \sim \text{Gamma}(n, 1)$. Since $\sigma > (1+\epsilon)\sigma_0$, for all $t < 0$,

$$\begin{aligned} \mathbb{E}_P(1 - \Psi_{2n}) &= P\left\{\sum_{j=1}^n \left|\frac{Y_j - \eta_{0j}}{\sigma_0}\right| \leq n\sqrt{1+\epsilon}\right\} \\ &\leq P\left\{\sum_{j=1}^n \left|\frac{Y_j - \eta_{0j}}{\sigma}\right| \frac{\sigma}{\sigma_0} \leq n\sqrt{1+\epsilon}\right\} \\ &= P\left\{\sum_{j=1}^n \left|\left(\frac{Y_j - \eta_{*j}}{\sigma}\right) + \left(\frac{\eta_{*j} - \eta_{0j}}{\sigma}\right)\right| \leq n\frac{\sigma_0}{\sigma}\sqrt{1+\epsilon}\right\} \\ &\leq P\left\{\sum_{j=1}^n \left|\frac{Y_j - \eta_{*j}}{\sigma}\right| \leq n\frac{\sigma_0}{\sigma}\sqrt{1+\epsilon}\right\} \tag{3.23} \\ &\leq \Pr\left\{V \leq \frac{n}{1+\sqrt{\epsilon^*}}\right\} \leq \Pr\left\{\exp(Vt) \geq \exp\left(\frac{nt}{1+\sqrt{\epsilon^*}}\right)\right\} \\ &\leq \exp\left(-\frac{nt}{1+\sqrt{\epsilon^*}}\right)(1-t)^{-n}, \end{aligned}$$

where $\epsilon^* = (\sqrt{1+\epsilon}-1)^2$.

The inequality (3.23) follows from Proposition 3.5.1. Finally, let $t = -\sqrt{\epsilon^*}$ to get

$$\mathbb{E}_P(1 - \Psi_n[\eta_1, \epsilon]) \leq \exp\left(n\left[\frac{\sqrt{\epsilon^*}}{1+\sqrt{\epsilon^*}} - \log(1+\sqrt{\epsilon^*})\right]\right) \leq \exp\left(-n\frac{\epsilon^* - \epsilon^*\sqrt{\epsilon^*}}{2(1+\sqrt{\epsilon^*})}\right),$$

where the last inequality follows from the fact that $\log(1+x) > x - x^2/2$.

■

Lemma 3.5.11. *Let η_1 be a continuous function on $[0, 1]$, and define $\eta_{ij} = \eta_i(x_j)$ for $i = 0, 1$ and $j = 1, \dots, n$. Let $\epsilon > 0$, and $0 < r < 4\sigma_0\sqrt{\epsilon - \epsilon^2}$. Let $\Psi_{3n}[\eta_1, \epsilon]$ be the indicator of the set A_3 , where*

1. If $Y_j \sim N(\eta_{0j}, \sigma_0^2)$

$$A_3 = \left\{ \sum_{j=1}^n \frac{(Y_j - \eta_{1j})^2}{\sigma_0^2} \leq n(1 - \epsilon^2) \right\},$$

2. If $Y_j \sim DE(\eta_{0j}, \sigma_0)$

$$A_3 = \left\{ \sum_{j=1}^n \left| \frac{Y_j - \eta_{1j}}{\sigma_0} \right| > n\sqrt{1 - \epsilon^2} \right\},$$

Then there exists a constant C_8 such that for all η_1 that satisfy

$$\sum_{j=1}^n \|\eta_{1j} - \eta_{0j}\| < rn$$

$E_{P_0}(\Psi_{3n}[\eta_1, \epsilon]) < \exp(-nC_8)$ Also, there exist a constants C_9 such that for all sufficiently large n and all η and σ satisfying $\|\eta - \eta_1\|_\infty < r/4$ and $\sigma < \sigma_0(1 - \epsilon)$,

$$E_P(1 - \Psi_{3n}[\eta_1, \epsilon]) \leq \exp(-nC_9),$$

where P is the joint distribution of $\{Y_n\}_{n=1}^\infty$ assuming that $\theta = (\eta, \sigma)$.

Proof. 1. Normal data:

(1) Type I error:

Let $W \sim \chi_n^2$ and $W' \sim \chi^2(n, \vartheta)$ with $\vartheta = \sum_{j=1}^n \left(\frac{\eta_{1j} - \eta_{0j}}{\sigma} \right)^2$. For all $t < 0$,

$$\begin{aligned}
 E_{P_0}(\Psi_{3n}) &= P_0 \left\{ \sum_{j=1}^n \left(\frac{Y_j - \eta_{1j}}{\sigma_0} \right)^2 \leq n[1 + \epsilon](1 - \epsilon) \right\} \\
 &\leq P_0 \left\{ \sum_{j=1}^n \left(\frac{Y_j - \eta_{0j}}{\sigma_0} + \frac{\eta_{0j} - \eta_{1j}}{\sigma_0} \right)^2 \leq n[1 + \epsilon](1 - \epsilon) \right\} \\
 &\leq \Pr \{ W' \leq n[1 - \epsilon^2] \} \leq \Pr \{ W \leq n[1 - \epsilon^2] \} \\
 &= \Pr \{ \exp(Wt) \geq \exp(nt[1 - \epsilon^2]) \} \\
 &\leq \exp(-nt[1 - \epsilon^2]) (1 - 2t)^{-n/2} \\
 &= \exp\left(-\frac{nt}{1 + \epsilon^*}\right) (1 - 2t)^{-n/2},
 \end{aligned}$$

where $\epsilon^* = \frac{\epsilon^2}{1 - \epsilon^2}$. Let $t = -\epsilon^*/2$ to get

$$E_{P_0}(\Psi_{3n}) \leq \exp\left(-n \frac{(\epsilon^*)^2 - (\epsilon^*)^3}{4(1 + \epsilon^*)}\right).$$

where the last inequality follows from the fact that $\log(1 + x) > x - x^2/2$.

(2) Type II error:

For all $t > 0$,

$$\begin{aligned}
 E_P(1 - \Psi_{3n}[\eta_1, \epsilon]) &= P \left\{ n[1 - \epsilon^2] \leq \sum_{j=1}^n \left(\frac{Y_j - \eta_{1j}}{\sigma_0} \right)^2 \right\} \\
 &= \Pr \left(n \frac{\sigma_0^2}{\sigma^2} [1 - \epsilon^2] \leq W' \right) = \Pr \left\{ \exp(W't) \geq \exp\left(nt[1 - \epsilon^2] \frac{\sigma_0^2}{\sigma^2}\right) \right\},
 \end{aligned}$$

where W' is a noncentral χ^2 distribution, $\chi^2(n, \vartheta)$, where $\vartheta = \sum_{j=1}^n \left(\frac{\eta_{*j} - \eta_{1j}}{\sigma} \right)^2$. Note that the moment generating function of noncentral $\chi^2(n, \vartheta)$ distribution is given as $E(\exp W't) = M_{W'}(t) = (1 - 2t)^{-n/2} \exp\{-\vartheta/2 [1 - (1 - 2t)^{-1}]\}$ for all $t < 1/2$.

Thus for all $0 < t < 1/2$,

$$\begin{aligned}
& \Pr \left\{ \exp(W't) \geq \exp \left(nt[1 - \epsilon^2] \frac{\sigma_0^2}{\sigma^2} \right) \right\} \\
& \leq M_{W'}(t) \exp \left(-nt[1 - \epsilon^2] \frac{\sigma_0^2}{\sigma^2} \right) \\
& = \exp \left\{ \frac{n}{2} \left[-\log(1 - 2t) - \left(1 - \frac{1}{1 - 2t} \right) \frac{1}{n} \sum_{j=1}^n \left(\frac{\eta_{*j} - \eta_{1j}}{\sigma} \right)^2 - 2t[1 - \epsilon^2] \frac{\sigma_0^2}{\sigma^2} \right] \right\}
\end{aligned}$$

If $\|\eta_1 - \eta\|_\infty < r/4$, then

$$|\eta_{1j} - \eta_{*j}| < r/4, \quad \forall j = 1, 2, \dots$$

Therefore,

$$\begin{aligned}
& \exp \left\{ \frac{n}{2} \left[-\log(1 - 2t) - \left(1 - \frac{1}{1 - 2t} \right) \frac{1}{n} \sum_{j=1}^n \left(\frac{\eta_{*j} - \eta_{1j}}{\sigma} \right)^2 - 2t[1 - \epsilon^2] \frac{\sigma_0^2}{\sigma^2} \right] \right\} \\
& \leq \exp \left\{ \frac{n}{2} \left[\log \left(1 + \frac{2t}{1 - 2t} \right) + \left(\frac{2t}{1 - 2t} \right) \left(\frac{r}{4\sigma} \right)^2 - 2t[1 - \epsilon^2] \frac{\sigma_0^2}{\sigma^2} \right] \right\} \\
& \leq \exp \left\{ \frac{n}{2} \left[\frac{2t}{1 - 2t} + \left(\frac{2t}{1 - 2t} \right) \left(\frac{r}{4\sigma} \right)^2 - 2t[1 - \epsilon^2] \frac{\sigma_0^2}{\sigma^2} \right] \right\} \\
& = \exp \left\{ n \frac{t\sigma_0^2}{\sigma^2} \left[\frac{1}{1 - 2t} \left(\frac{\sigma^2}{\sigma_0^2} + \frac{r^2}{16\sigma_0^2} \right) - [1 - \epsilon^2] \right] \right\} \\
& \leq \exp \left\{ n \frac{t\sigma_0^2}{\sigma^2} \left[\frac{1}{1 - 2t} \left((1 - \epsilon)^2 + \frac{r^2}{16\sigma_0^2} \right) - [1 - \epsilon^2] \right] \right\} \tag{3.24}
\end{aligned}$$

Because $r^2 < 16\sigma_0^2(\epsilon - \epsilon^2)$, (3.24) is less than

$$\exp \left\{ nt \frac{\sigma_0^2}{\sigma^2} \left[\frac{1}{1 - 2t} (1 - \epsilon) - [1 - \epsilon^2] \right] \right\}$$

Now, take $t = t^*$ such that $\frac{1}{1 - 2t^*} = \frac{1 - (1 + \epsilon)\epsilon^2}{1 - \epsilon}$ to get

$$E_P(\Psi_{3n}) \leq \exp \left\{ -nt^* \frac{\sigma_0^2}{\sigma^2} \epsilon^3 \right\} \leq \exp \left\{ -\frac{nt^*}{(1 - \epsilon)^2} \epsilon^3 \right\}.$$

2. Laplace data:

(1) Type I error:

Let $V \sim \text{Gamma}(n, 1)$. Then, for all $t < 0$,

$$\begin{aligned}
\mathbb{E}_{P_0}(\Psi_{3n}) &= P_0 \left(\sum_{j=1}^n \left| \frac{Y_j - \eta_{1j}}{\sigma_0} \right| < n\sqrt{1 - \epsilon^2} \right) \\
&\leq \Pr \left(V < n\sqrt{1 - \epsilon^2} \right), \text{ by Lemma 3.5.1} \\
&\leq \exp \left(-n(\sqrt{1 - \epsilon^2})t \right) \mathbb{E}(\exp(t_2 V)) \\
&= \exp \left(-n(\sqrt{1 - \epsilon^2})t \right) (1 - t)^{-n},
\end{aligned}$$

Take

$$t = 1 - \frac{1}{\sqrt{1 - \epsilon^2}}.$$

Then,

$$\begin{aligned}
\mathbb{E}_{P_0}(\Psi_{2n}) &\leq \exp \left(n(1 - \sqrt{1 - \epsilon^2}) + n \log \left[1 - (1 - \sqrt{1 - \epsilon^2}) \right] \right) \\
&\leq \exp \left(-n \frac{(1 - \sqrt{1 - \epsilon^2})^2}{2} \right),
\end{aligned}$$

where the last inequality follows from the fact that $\log(1 - x) \leq -x - x^2/2$, $\forall x > 0$.

Therefore, $\mathbb{E}_{P_0}(\Psi_{3n}) \leq \exp(-C_9 n)$.

(2) Type II error:

Let V have a $\text{Gamma}(n, 1)$ distribution. For all $t > 0$,

$$\begin{aligned}
\mathbb{E}_P(1 - \Psi_{3n}[\eta_1, \epsilon]) &= P \left\{ n\sqrt{1 - \epsilon^2} \leq \sum_{j=1}^n \left| \frac{Y_j - \eta_{1j}}{\sigma_0} \right| \right\} \\
&= P \left\{ n \frac{\sigma_0}{\sigma} \sqrt{1 - \epsilon^2} \leq \sum_{j=1}^n \left| \frac{Y_j - \eta_{*j}}{\sigma} + \frac{\eta_{*j} - \eta_{1j}}{\sigma} \right| \right\} \\
&\leq \Pr \left(n \frac{\sigma_0}{\sigma} \sqrt{1 - \epsilon^2} - \sum_{j=1}^n \left| \frac{\eta_{*j} - \eta_{1j}}{\sigma} \right| \leq V \right) \\
&= \Pr \left\{ \exp(Vt) \geq \exp \left(nt\sqrt{1 - \epsilon^2} \frac{\sigma_0}{\sigma} - t \sum_{j=1}^n \left| \frac{\eta_{*j} - \eta_{1j}}{\sigma} \right| \right) \right\},
\end{aligned}$$

Thus for all $0 < t < 1$,

$$\begin{aligned} & \Pr \left\{ \exp(Vt) \geq \exp \left(nt\sqrt{1-\epsilon^2}\frac{\sigma_0}{\sigma} - t \sum_{j=1}^n \left| \frac{\eta_{*j} - \eta_{1j}}{\sigma} \right| \right) \right\} \\ & \leq M_V(t) \exp \left(-nt\sqrt{1-\epsilon^2}\frac{\sigma_0}{\sigma} + t \sum_{j=1}^n \left| \frac{\eta_{*j} - \eta_{1j}}{\sigma} \right| \right) \\ & = \exp \left\{ n \left[-\log(1-t) - t\sqrt{1-\epsilon^2}\frac{\sigma_0}{\sigma} + t\frac{1}{n} \sum_{j=1}^n \left| \frac{\eta_{*j} - \eta_{1j}}{\sigma} \right| \right] \right\} \end{aligned}$$

If $\|\eta_1 - \eta\|_\infty < r/4$, then $|\eta_{1j} - \eta_{*j}| < r/4$, $\forall j = 1, 2, \dots$. Therefore,

$$\begin{aligned} & \exp \left\{ n \left[-\log(1-t) - t\sqrt{1-\epsilon^2}\frac{\sigma_0}{\sigma} + t\frac{1}{n} \sum_{j=1}^n \left| \frac{\eta_{*j} - \eta_{1j}}{\sigma} \right| \right] \right\} \\ & \leq \exp \left\{ n \left[\log \left(1 + \frac{t}{1-t} \right) + t\frac{r}{4\sigma} - t\sqrt{1-\epsilon^2}\frac{\sigma_0}{\sigma} \right] \right\} \\ & \leq \exp \left\{ n \left[\frac{t}{1-t} + t\frac{r}{4\sigma} - t\sqrt{1-\epsilon^2}\frac{\sigma_0}{\sigma} \right] \right\} \\ & = \exp \left\{ n\frac{t\sigma_0}{\sigma} \left[\frac{1}{1-t} \left(\frac{\sigma}{\sigma_0} + \frac{r}{4\sigma_0} \right) - \sqrt{1-\epsilon^2} \right] \right\} \\ & \leq \exp \left\{ n\frac{t\sigma_0}{\sigma} \left[\frac{1}{1-t} \left((1-\epsilon) + \frac{r}{4\sigma_0} \right) - \sqrt{1-\epsilon^2} \right] \right\} \end{aligned} \tag{3.25}$$

Because $r < 4\sigma_0(\epsilon - \epsilon^2)$, (3.25) is less than

$$\exp \left\{ nt\frac{\sigma_0}{\sigma} \left[\frac{1}{1-t} (1-\epsilon^2) - \sqrt{1-\epsilon^2} \right] \right\}$$

Now, take $t = t^*$ such that $\frac{1}{1-t^*} = \frac{\sqrt{1-\epsilon^2} - \epsilon^3}{1-\epsilon^2}$ to get

$$E_P(\Psi_{3n}) \leq \exp \left\{ -nt^*\frac{\sigma_0}{\sigma}\epsilon^3 \right\} \leq \exp \left\{ -\frac{nt^*}{1-\epsilon}\epsilon^3 \right\}.$$

■

Verifying (3.21) is done differently for the random and nonrandom design cases.

For the random design case, Lemma 3.5.12 tells us that (3.21) occurs all but finitely often with probability 1. Since there are only finitely many η^j to consider for each n , this suffices to complete the proof of Theorem 3.4.2.

Lemma 3.5.12. *Assume Assumption RD. Let η be a function such that $d_Q(\eta, \eta_0) > \epsilon$. Let $0 < r < \epsilon^2$, and define*

$$A_n = \left\{ \sum_{i=1}^n |\eta(X_i) - \eta_0(X_i)| \geq rn \right\}.$$

Then there exists $C_{11} > 0$ such that $\Pr(A_n^C) \leq \exp(-C_{11}n)$ for all n and A_n occurs all but finitely often with probability 1. The same C_{11} works for all η such that $d_Q(\eta, \eta_0) > \epsilon$.

Proof. Let $B = \{x | \eta(x) - \eta_0(x) > \epsilon\}$, so that $Q(B) > \epsilon$. Let $Z = n - \sum_{i=1}^n I_B(X_i)$, and notice that Z has a binomial distribution with parameters n and $1 - Q(B)$. Let $q = r/\epsilon < \epsilon$, and let Z' have a binomial distribution with parameters n and $1 - \epsilon$ so that Z' stochastically dominates Z . Then

$$\Pr(A_n^C) \leq \Pr(Z > n[1 - q]) \leq \Pr(Z' > n[1 - q]).$$

Write

$$\begin{aligned} \Pr(Z' > n[1 - q]) &= \Pr(\exp(tZ') > \exp(tn[1 - q])), \text{ for all } t > 0, \\ &\leq [\epsilon + [1 - \epsilon] \exp(t)]^n \exp(-tn[1 - q]). \end{aligned}$$

Let

$$\begin{aligned} t &= \log \left(\frac{\epsilon(1 - q)}{q(1 - \epsilon)} \right) > 0, \\ C_{11} &= q \log \left(\frac{q}{\epsilon} \right) + (1 - q) \log \left(\frac{1 - q}{1 - \epsilon} \right) > 0. \end{aligned}$$

Then $\Pr(A_n^C) \leq \exp(-C_{11}n)$, and C_{11} doesn't depend on the particular η . The probability one claim follows from the first Borel-Cantelli lemma. ■

For the nonrandom design case, we verify (3.21) for all η_1 that are far from η_0 in L^1 distance.

Lemma 3.5.13. *Assume Assumption NRD. Let λ be Lebesgue measure. Let K_1 be the constant mentioned in Assumption NRD. Let $V > 0$ be a constant. For each integer n , let*

A_n be the set of all continuous functions γ such that $\forall x_1, x_2 \in [0, 1], |\gamma(x_1) - \gamma(x_2)| \leq (M_n + V)|x_1 - x_2|$. For each function γ and $\epsilon > 0$, define $B_{\epsilon, \gamma} = \{x : |\gamma(x)| > \epsilon\}$. Then for each $\epsilon > 0$ there exists an integer N such that, for all $n \geq N$ and all $\gamma \in A_n$,

$$\sum_{i=1}^n |\gamma(x_i)| \geq (\lambda(B_{\epsilon, \gamma})K_1n - 1)\frac{\epsilon}{2}. \quad (3.26)$$

Proof. Let N be large enough so that $(M_n + V)/(K_1n) < \epsilon/2$ for all $n \geq N$. Because γ is continuous, $B_{\epsilon, \gamma}$ is an open set and it is the union of a countable collection of disjoint open intervals, i.e. $B_{\epsilon, \gamma} = \cup_{i=1}^{\infty} B_i$, where $B_i = (x_{L,i}, x_{R,i})$ is an open interval whose length is $\lambda(B_i) = x_{R,i} - x_{L,i} \geq 0$. Some of the B_i intervals might lie entirely between successive design points. Let $x_0 = 0$ and $x_{n+1} = 1$. Define, for $j = 0, \dots, n$,

$$\begin{aligned} a_j &= \{i : x_j < x_{R,i} < x_{j+1}\}, \\ b_j &= \{i : x_j < x_{L,i} < x_{j+1}\}, \\ \ell_j &= \inf\{x_{L,i} : i \in a_j\}, \\ u_j &= \sup\{x_{R,i} : i \in b_j\}. \end{aligned}$$

Then the open interval $F_j = (\ell_j, u_j)$ contains the same design points as

$$D_j = \bigcup_{i \in a_j \cup b_j} B_i.$$

If $x_j \in F_j$ (for $j \in \{1, \dots, n\}$), then $\ell_j < u_{j-1}$ and (ℓ_{j-1}, u_j) contains the same design points as $D_{j-1} \cup D_j$. By combining all of the overlapping F_j intervals, we obtain finitely many disjoint intervals E_1, \dots, E_m whose total length is at least $\lambda(B_{\gamma, \epsilon})$ (because their union contains $B_{\epsilon, \gamma}$) and that contain the same design points as $B_{\epsilon, \gamma}$. Write each $E_j = (f_j, g_j)$ and $L_j = g_j - f_j$, and assume that the intervals are ordered so that $g_j < f_{j+1}$ for all j . Let $E = \cup_{j=1}^m E_j$. Let $[a]$ denote the integer part of a . Each E_j contains at least $\lfloor L_j K_1 n \rfloor$ design points because the maximum spacing is assumed to be less than or equal to $1/(K_1 n)$. For each j such that $f_j > x_1$, let x_j^* be the largest $x_i \leq f_j$. Then $x_j^* \notin E$,

$|\gamma(x_j^*) - \gamma(f_j)| \leq (M_n + V)|x_j^* - f_j|$ and $|x_j^* - f_j| < 1/(K_1 n)$. Hence,

$$|\gamma(x_j^*)| > \epsilon - \frac{M_n + V}{K_1 n} > \frac{\epsilon}{2}.$$

If we include x_j^* with the design points already in E_j , we have, associated with each j such that $f_j > x_1$, at least $\lceil L_j K_1 n \rceil$ design points x with $|\gamma(x)| > \epsilon/2$, where $\lceil a \rceil$ is the smallest integer greater than or equal to a . There is at most one j such that $f_j \leq x_1$ for which we have at least $\lceil L_j K_1 n \rceil - 1$ design points with $|\gamma(x)| > \epsilon/2$. Since we have not counted any design points more than once, we can add over all j to see that there are at least $\lambda(B_{\gamma, \epsilon})K_1 n - 1$ design points x in $B_{\epsilon/2, \gamma}$ so long as $n \geq N$. Hence we satisfy (3.26). ■

Lemma 3.5.14. *Assume Assumption NRD. For each integer n , let A_n be the set of all continuously differentiable functions η such that $\|\eta\|_\infty < M_n$ and $\|\eta'\|_\infty < M_n$. Then for each $\epsilon > 0$ there exist an integer N and $r > 0$ such that, for all $n \geq N$ and all $\eta \in A_n$ such that $\|\eta - \eta_0\|_1 > \epsilon$, $\sum_{i=1}^n |\eta(x_i) - \eta_0(x_i)| \geq rn$.*

Proof. Let V_0 be an upper bound of η_0 and V be an upper bound on the derivative of η_0 . Let $0 < \delta < \epsilon$ and let N be large enough so that $(M_n + V_0)/n < (1/2)K_1(\epsilon - \delta)$ for all $n \geq N$. Let $r = K_1(\epsilon - \delta)/8$ and $D_i = \{x : (i-1)\delta < |\eta(x) - \eta_0(x)| < i\delta\}$.

Let λ be Lebesgue measure. Then, $\|\eta - \eta_0\|_1$ can be bounded as follows.

$$\sum_i i\delta\lambda(D_i) \geq \|\eta - \eta_0\|_1 > \epsilon. \quad (3.27)$$

Let $\zeta(x) = |\eta(x) - \eta_0(x)|$ and $\zeta_m(x) = \min\{m\delta, \zeta(x)\}$, for $m = 0, \dots, n$. Note that $\zeta_{\lceil (M_n + V_0)/\delta \rceil}(x)$ is the same as $\zeta(x)$.

For $m = 1, \dots, n$, define $B_m \equiv \{x : \zeta_m(x) > (2m-1)\delta/2\}$. Then, for all $x \in B_m$, $\zeta_m(x) - \zeta_{m-1}(x) > \delta/2$. Thus, Lemma 3.5.13 (with $\gamma = \zeta_m - \zeta_{m-1}$ and $\epsilon = \delta/2$) implies

$$\sum_{i=1}^n |\zeta_m(x_i) - \zeta_{m-1}(x_i)| \geq (\lambda(B_m)K_1 n - 1) \frac{\delta}{4}.$$

Now, write

$$\begin{aligned}
\sum_{i=1}^n |\eta(x_i) - \eta_0(x_i)| &= \sum_{i=1}^n \zeta_{\lceil (M_n + V_0)/\delta \rceil}(x_i) \\
&= \sum_{i=1}^n \sum_{m=1}^{\lceil (M_n + V_0)/\delta \rceil} \{\zeta_m(x_i) - \zeta_{m-1}(x_i)\} \\
&= \sum_{m=1}^{\lceil (M_n + V_0)/\delta \rceil} \sum_{i=1}^n \{\zeta_m(x_i) - \zeta_{m-1}(x_i)\}, \\
&\geq \sum_{m=1}^{\lceil (M_n + V_0)/\delta \rceil} [\lambda(B_m)K_1n - 1] \frac{\delta}{4}. \tag{3.28}
\end{aligned}$$

Also, for $m = 1, \dots, n$,

$$B_m \supset \bigcup_{i=m+1}^{\lceil (M_n + V_0)/\delta \rceil} D_i.$$

It follows that

$$\begin{aligned}
\sum_{m=1}^{\lceil (M_n + V_0)/\delta \rceil} \delta \lambda(B_m) &\geq \sum_{i=2}^{\lceil (M_n + V_0)/\delta \rceil} (i-1) \delta \lambda(D_i) \\
&\geq \left\{ \sum_{i=2}^{\lceil (M_n + V_0)/\delta \rceil} i \delta \lambda(D_i) \right\} - \delta \\
&\geq \epsilon - \delta,
\end{aligned}$$

where the last inequality follows from (3.27). Combining this with (3.28) gives

$$\sum_{i=1}^n |\eta(x_i) - \eta_0(x_i)| \geq \frac{K_1 n}{4} (\epsilon - \delta) - \frac{M_n + V_0}{4} \geq n \frac{K_1 (\epsilon - \delta)}{8},$$

for all $n \geq N$. ■

Lemma 3.5.15. *There exist test functions $\{\Phi_n\}$ and constants $C_{12} > 0$, that satisfy*

- (i) $\sum_{n=1}^{\infty} \mathbb{E}_{P_0} \Phi_n < \infty$,
- (ii) $\sup_{\theta \in W_\epsilon^C \cap \Theta_n} \mathbb{E}_P(1 - \Phi_n) \leq \exp(-C_{12}n)$,

Proof. First, define $\Psi_n[\eta_1, \epsilon]$ as

$$\Psi_n[\eta_1, \epsilon] = 1 - (1 - \Psi_{1n}[\eta_1, \epsilon])(1 - \Psi_{2n}[\epsilon])(1 - \Psi_{3n}[\eta_1, \epsilon]).$$

From Lemmas 3.5.9–3.5.14, it is clear that $E_{P_0} \Psi_n < \exp(-2c_n^2)$ and $E_P(1 - \Psi_n) \leq \exp(-C_{12}n)$ for each $\eta_1 \in W_\epsilon^C$.

To create a test that doesn't depend on a specific choice of η_1 in Lemma 3.5.9, we make use of the covering of the sieve. Let r be the same number that appears in Lemmas 3.5.9–3.5.11. Let $t = \min\{\epsilon/2, r/4\}$. Let N_t be the t covering number of Θ_{1n} in the supremum (L^∞) norm.

Recall that given $0 < \delta < 1/2$, $\frac{2\delta+1}{2} < \alpha_1 < 1$ and $\frac{\alpha}{2} < \tau_1 < \frac{1}{2}$. Thus, with $M_n = O(n^{\alpha_1})$ and $c_n = n^{\tau_1}$, we have $\log(N_t) = o(c_n^2)$ from Lemma 3.5.4. Let $\eta^1, \dots, \eta^{N_t} \in \Theta_{1n}$ be such that for each $\eta \in \Theta_{1n}$ there exists j such that $\|\eta - \eta^j\|_\infty < t$. If $\|\eta - \eta_0\|_1 > \epsilon$, then $\|\eta^j - \eta_0\|_1 > \epsilon/2$. Similarly, if $d_Q(\eta, \eta_0) > \epsilon$, then $d_Q(\eta^j, \eta_0) > \epsilon/2$.

Finally, define

$$\Phi_n = \max_{1 \leq j \leq N_t} \Psi_n[\eta^j, \epsilon/2].$$

Since we verified that there exists r such that (3.21) holds for every such η^j in Lemma 3.5.14, then

$$\begin{aligned} E_{P_0} \Phi_n &\leq \sum_{j=1}^{N_t} E_{P_0} \Psi_n[\eta^j, \epsilon/2] \\ &\leq C_3 N_t \exp(-2c_n^2) \\ &= C_3 \exp(\log[N_t] - 2c_n^2) \\ &\leq C_3 \exp(-c_n^2). \end{aligned}$$

For $\theta = (\eta, \sigma) \in U_\epsilon^C \cap \Theta_n$ or in $W_\epsilon^C \cap \Theta_n$, the type II error probability of Ψ_n is no larger than the minimum of the individual type II error probabilities of the $\Psi_n[\eta^j, \epsilon/2]$ tests. Hence, we have a uniformly consistent test Φ_n , which has exponentially small type II error evaluated at θ . Hence, we have

- (i) $\sum_{n=1}^{\infty} \mathbb{E}_{P_0} \Phi_n < \infty,$
- (ii) $\sup_{\theta \in W_{\epsilon}^C \cap \Theta_n} \mathbb{E}_P(1 - \Phi_n) \leq \exp(-C_{12}n).$

■

3.5.4 Proofs of Main Theorems

1. Proof of Theorem 3.4.1

We apply Theorem 3.3.1 to the case that the neighborhood is W_{ϵ} . The proof of condition (A1) of Theorem 3.3.1 was made in Section 3.5.1, and the same proof applies to all of the main Theorems, regardless of the topologies that we consider. The proof of condition (A2) of Theorem 3.3.1 was made in Section 3.5.3. The tests are constructed and shown to have exponentially small Type I and Type II errors as in Lemma 3.5.15, and the same tests are used for all of the main Theorems. The crucial point to construct those tests, particularly for $\Psi_{1n}[\eta_1, \epsilon]$, was whether or not the covariate values are spread out sufficiently well so that there are enough covariate values where η and η_0 are separated. This notion is formulated as (3.21) and it depends on the closeness between η and η_0 , in terms of the different topologies. Lemma 3.5.14 establishes (3.21) under the L^1 topology for regression functions. By verifying two conditions (A1) and (A2) of Theorem 3.3.1, the proof of Theorem 3.4.1 is done.

2. Proof of Corollary 3.4.1

It is obvious from the fact that L^1 convergence implies in-probability convergence and Theorem 3.4.1.

3. Proof of Theorem 3.4.2

The proof is the same as the proof of Theorem 3.4.1 except for the step of showing the existence of tests with random covariates, i.e. assumption RD and an in-probability

topology. We use the same test functions as in Lemma 3.5.9–3.5.11 and only need to recompute the type II error probability of $\Psi_{1n}[\eta_1, \epsilon]$ used when $d_Q(\eta_1, \eta_0) > \epsilon$ and $\sigma \leq \sigma_0(1 + \epsilon)$. As mentioned earlier, Lemma 3.5.12 is sufficient to get the exponentially small error bound as follows. Let $\theta \in U_\epsilon^* = \{(\eta, \sigma) : d_Q(\eta_1, \eta_0) > \epsilon, \sigma \leq \sigma_0(1 + \epsilon)\}$ and $A_n = \{\sum_{i=1}^n |\eta(X_i) - \eta_0(X_i)| \geq rn\}$. By Lemma 3.5.12,

$$\begin{aligned} & \mathbb{E}_P(1 - \Psi_n[\eta_1, \epsilon]) \\ &= \int_{A_n \cap U_\epsilon^*} \mathbb{E}_P(1 - \Psi_n[\eta_1, \epsilon] | X_1, \dots, X_n) dQ + \int_{A_n^c \cap U_\epsilon^*} \mathbb{E}_P(1 - \Psi_n[\eta_1, \epsilon] | X_1, \dots, X_n) dQ \\ &\leq \int_{A_n} \mathbb{E}(1 - \Psi_n[\eta_1, \epsilon] | X_1, \dots, X_n) dQ + \Pr(A_n^c \cap U_\epsilon^*) \end{aligned} \quad (3.29)$$

The first term in (3.29) is exponentially small because X_1, \dots, X_n in A_n satisfy the condition (3.21), and the second term is also exponentially small from Lemma 3.5.12. Therefore, the theorem is proven.

4. Proof of Theorem 3.4.3

First, we calculate the Hellinger distance between two density functions, $d_H(f, f_0)$, where f is the joint density of (X, Y) when η and σ are arbitrary, and f_0 is the density when $\eta = \eta_0$ and $\sigma = \sigma_0$. Let ξ be the product of Q and Lebesgue measure λ . Then the joint densities of X and Y defined above with respect to ξ are given by

$$f(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{[y - \eta(x)]^2}{2\sigma^2}\right\} \text{ and } f_0(y|x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{[y - \eta_0(x)]^2}{2\sigma_0^2}\right\}$$

for normal data, or

$$f(y|x) = \frac{1}{2\sigma} \exp\left\{-\frac{|y - \eta(x)|}{\sigma}\right\} \text{ and } f_0(y|x) = \frac{1}{2\sigma_0} \exp\left\{-\frac{|y - \eta_0(x)|}{\sigma_0}\right\}$$

for Laplace data. To simplify the calculation, we consider the quantity $h(f, f_0)$ defined as

$$h(f, f_0) = \frac{1}{2} d_H^2(f, f_0) = 1 - \int \sqrt{ff_0} d\mu$$

and $h(f, f_0)$ is calculated as follows.

(a) $Y_i|X_i \stackrel{\text{ind}}{\sim} N(\eta(X_i), \sigma^2)$

$$\begin{aligned}
h(f, f_0) &= 1 - \frac{1}{\sqrt{2\pi\sigma\sigma_0}} \int \int \exp \left\{ -\frac{1}{4\sigma^2} [y - \eta(x)]^2 - \frac{1}{4\sigma_0^2} [y - \eta_0(x)]^2 \right\} dy dQ \\
&= 1 - \int \int \exp \left\{ -\left(\frac{1}{4\sigma^2} + \frac{1}{4\sigma_0^2} \right) \left[y - \left(\frac{\eta(x)}{4\sigma^2} + \frac{\eta_0(x)}{4\sigma_0^2} \right) \right]^2 \right\} \\
&\quad \frac{1}{\sqrt{2\pi\sigma\sigma_0}} \times \exp \left\{ -\frac{\eta(x)^2}{4\sigma^2} - \frac{\eta_0(x)^2}{4\sigma_0^2} + \left(\frac{\eta(x)}{4\sigma^2} + \frac{\eta_0(x)}{4\sigma_0^2} \right)^2 \right\} dy dQ \\
&= 1 - \int \sqrt{\frac{2\sigma\sigma_0}{\sigma^2 + \sigma_0^2}} \exp \left\{ -\frac{1}{16\sigma^2\sigma_0^2} [\eta(x) - \eta_0(x)]^2 \right\} dQ \quad (3.30)
\end{aligned}$$

The integral in (3.30) is of the form $\int c_1 \exp(-c_2[\eta(x) - \eta_0(x)]^2) dQ(x)$, where c_1 can be made arbitrarily close to 1 by choosing $|\sigma/\sigma_0 - 1|$ small enough and c_2 is bounded when σ is close to σ_0 . In addition, since $c_1 \exp(-c_2[\eta(x) - \eta_0(x)]^2)$ is bounded by c_1 , we have

$$\begin{aligned}
&\int 1 - c_1 \exp(-c_2[\eta(x) - \eta_0(x)]^2) dQ(x) \\
&= \int_{|\eta - \eta_0| \leq \delta} 1 - c_1 \exp(-c_2[\eta(x) - \eta_0(x)]^2) dQ(x) \\
&\quad + \int_{|\eta - \eta_0| > \delta} 1 - c_1 \exp(-c_2[\eta(x) - \eta_0(x)]^2) dQ(x) \\
&\leq \{1 - c_1 \exp(-c_2\delta^2)\} + (1 + c_1)Q(\{x : |\eta(x) - \eta_0(x)| > \delta\})
\end{aligned}$$

Therefore, it follows that for each ϵ there exists a δ such that (3.30) will be less than ϵ whenever $|\sigma/\sigma_0 - 1| < \delta$ and $d_Q(\eta, \eta_0) < \delta$.

(b) $Y_i|X_i \stackrel{\text{ind}}{\sim} DE(\eta(X_i), \sigma)$

$$\begin{aligned}
h(f, f_0) &= 1 - \frac{1}{\sqrt{4\sigma\sigma_0}} \int \int \exp \left\{ -\frac{1}{2\sigma} |y - \eta(x)| - \frac{1}{2\sigma_0} |y - \eta_0(x)| \right\} dy dQ \\
&\leq 1 - \int \int \exp \left\{ -\left(\frac{1}{2\sigma} + \frac{1}{2\sigma_0} \right) \left| y - \left(\frac{\eta(x) + \eta_0(x)}{2} \right) \right| \right\} \\
&\quad \frac{1}{\sqrt{4\sigma\sigma_0}} \times \exp \left\{ -\left(\frac{1}{4\sigma} + \frac{1}{4\sigma_0} \right) |\eta(x) - \eta_0(x)| \right\} dy dQ \\
&\leq 1 - \int \left[\frac{1}{\sqrt{4\sigma\sigma_0}} \times \exp \left\{ -\left(\frac{1}{4\sigma} + \frac{1}{4\sigma_0} \right) |\eta(x) - \eta_0(x)| \right\} \times \left(\frac{1}{4\sigma} + \frac{1}{4\sigma_0} \right)^{-1} \right] dQ \quad (3.31)
\end{aligned}$$

The integral in (3.31) is of the form $\int c_1 \exp(-c_2|\eta(x) - \eta_0(x)|)dQ(x)$, where c_1 can be made arbitrarily close to 1 by choosing $|\sigma/\sigma_0 - 1|$ small enough and c_2 is bounded when σ is close to σ_0 . Similarly to (a), it follows that for each ϵ there exists a δ such that (3.31) will be less than ϵ whenever $|\sigma/\sigma_0 - 1| < \delta$ and $d_Q(\eta, \eta_0) < \delta$.

Consequently, the result of Theorem 3.4.2 implies Theorem 3.4.3.

5. Proof of Theorem 3.4.4

For bounded functions, convergence in probability is equivalent to L^p convergence for all finite p . In particular, for every $\epsilon > 0$ and every finite p , there exists an ϵ' such that $U_{\epsilon'} \subseteq W_\epsilon$. Hence, Theorem 3.4.2 implies the conclusion to Theorem 3.4.4 as long as the GP prior defined in Assumption B also satisfy all the conditions required in Theorem 3.4.2.

If a GP satisfies the smoothness conditions that follow from Assumption P, then the conditional process given a set of bounded functions with positive probability also satisfies the smoothness conditions. We have already verified the prior positivity condition (A1). For subpart (iii) of (A2), we note that, if A and d are the constants guaranteed by Lemmas 3.5.7 and 3.5.8 for Π'_1 , then

$$\Pi_1 \left\{ \sup_{0 \leq s \leq 1} |\eta'(s)| > M \right\} \leq A \exp(-dM^2)/\Pi'_1(\Omega).$$

3.5.5 Illustrations of Covariance Functions

As we studied in the previous sections, conditions about covariance functions in Assumption P play a major part in our posterior consistency results, in particular when we deal with the prior positivity condition and the exponentially small probability outside of the sieve, Θ_n^C . Specifically, in order to ensure the prior positivity for the GP prior, we assumed that the covariance function is isotropic and a positive multiple of nowhere zero continuous density.

In addition, appropriate smoothness conditions on the covariance function for the Gaussian process were required for guaranteeing the exponentially small bound of the large deviation probability for the Gaussian process, as well as the existence of continuously differentiable sample paths. In this section, we verify these conditions for specific covariance functions that are frequently used in practice.

Regarding the prior positivity condition, Lemmas 3.5.1–3.5.3 tells us that the prior positivity condition for the GP regression model becomes the problem of the lower tail probability for the Gaussian process as follows:

$$\Pr \left\{ \sup_{x \in [0,1]} |\eta(x) - \eta_0(x)| < \epsilon \right\} > 0 \quad (3.32)$$

As mentioned in Lemmas 3.5.2 and 3.5.3, if the closure, $\bar{\mathcal{A}}_R$, contains $\eta_0(x)$, then (3.32) is ensured. From Tokdar and Ghosh (2004, Result2), it was identified that the isotropic covariance function with a nonzero continuous density contains every continuous function. We provide a simple extension of their result to an integrable function, $R_0(x)$, which contains the case of a positive multiple of nonzero continuous density.

Proposition 3.5.2. *Suppose that the support of β , is the positive real line, \mathbb{R}^+ and $R_0(x)$ is positive, continuous and integrable on \mathbb{R} and define $R(x, x'; \beta) = R_0(\beta|x - x'|)$. Then the closure of \mathcal{A}_R , $\bar{\mathcal{A}}_R$ is $C[0, 1]$, where*

$$\mathcal{A}_R = \left\{ f(x) = \sum_{i=1}^k a_i R(x, x_i^*; \beta) \text{ for some } \beta \in B, k \in \mathbb{N}, a_i \in \mathbb{R}, x_i^* \in \mathbb{Q} \cap [0, 1] \right\}$$

and $C[0, 1]$ is the set of all continuous functions in $[0, 1]$.

Proof. Since $R_0(x)$ is integrable, consider $R'_0(x) = \frac{R_0(x)}{\int R_0(x)dx}$. Then $R'_0(x)$ is a nowhere zero continuous density. Thus, the proof is straightforward from Tokdar and Ghosh (2004, Result2). ■

On the other hand, the smoothness conditions on the covariance function in the Gaussian process are also essential to achieve posterior consistency. In particular, the support of

Gaussian processes is mainly determined by the appropriate smoothness conditions, which make it sure that the prior assigns positive probability to every Kullback-Leibler neighborhood of true parameters, and it puts a negligible probability on the outside of the sieve, Θ_n . As described in Section 3.4, the consequence of assumption P is that there exists a constant K_2 such that

$$r_0(h) = 1 - \frac{K_2}{2}h^2 + O(h^4), \quad (3.33)$$

where $r_0(h) = R_0(|h|)/R_0(0)$. This condition guarantees the existence of continuous sample derivative $\eta'(\cdot)$ with probability 1. For detailed descriptions of analytic properties of covariance functions and the smoothness of sample paths of Gaussian processes, good references are Cramer and Leadbetter (1967), Adler (1981), Adler (1990) and Abrahamsen (1997).

Many covariance functions of Gaussian processes, which are widely used in the literature mentioned earlier, satisfy Assumption P. We give examples of these isotropic covariance functions and illustrate that they satisfy (3.33) and that each of them is a positive multiple of nonzero density function.

1. squared-exponential covariance function

$$r_0(h) = \exp(-h^2) = 1 - h^2 + O(h^4), \text{ as } h \rightarrow 0$$

and $\exp(-h^2)$ is a positive multiple of normal density.

2. Cauchy covariance function

$$r_0(h) = \frac{1}{1+h^2} = 1 - h^2 + O(h^4), \text{ as } h \rightarrow 0$$

and $\frac{1}{1+h^2}$ is a positive multiple of Cauchy density.

3. Matérn covariance function with $v > 2$

$$R_0(h) = \frac{1}{\Gamma(v)2^{v-1}}(\beta h)^v \mathcal{K}_v(\beta h),$$

where $\beta > 0$ and $\mathcal{K}_\nu(x)$ is a modified Bessel function of order ν .

It is known from Abrahamsen (1997, p. 43) that for $n < v$ then

$$\left. \frac{d^{2n-1}R(h)}{dh^{2n-1}} \right|_{h=0} = 0$$

and

$$-\infty < \left. \frac{d^{2n}R(h)}{dh^{2n}} \right|_{h=0} < 0$$

Consequently, it is straightforward that if $v > 2$, then there exists a constant $\xi > 0$ such that

$$R(h) = R(0) - \xi h^2 + O(h^4), \text{ as } h \rightarrow 0$$

In addition, the Matérn covariance function can be shown to be integrable. From the definition of Matérn covariance function, it is known that when the power, v , is $m + \frac{1}{2}$ with m , a nonnegative integer, the Matérn covariance function is of the form $e^{-\alpha|t|}$ times a polynomial in $|t|$ of degree m . (See Stein (1999, p. 31)). Therefore, if v is in the form of $m + \frac{1}{2}$, the Matérn covariance function is integrable. We think the integrability still holds for every v and would like to leave the proof of this conjecture to future work.

In summary, there are many covariance functions which satisfy Assumption P and these covariance functions are widely used for GP regression model in practice.

Chapter 4

Extension of Posterior Consistency Results

In the previous chapter, we established posterior consistency results for the regression function under various model structures. In this chapter, we study how much posterior consistency can be extended under more general model structures and raise some open issues.

There are several open issues that are worth further consideration. First, we need to think about the case in which the covariance function of the Gaussian process has (finitely many) parameters that need to be estimated. Second we need to deal with the case of multidimensional covariates. In this chapter, we treat these two issues in detail.

In addition to these issues, there are still remaining areas to be considered in the GP regression setting. They include the rate of convergence of the posterior, nonlinear autoregressive model using Gaussian process priors and others. We will discuss these issues briefly in the next chapter.

4.1 Hyperparameters in Covariance Function

In the implementation of GP regression, there are often cases in which the covariance function of the Gaussian process has (finitely many) parameters that need to be estimated. For instance, as mentioned in the description of Assumption P in Section 3.4, the smoothing parameter, β , is an important hyperparameter of the covariance function that is closely related to the prior positivity condition and that clearly must be estimated. That is, we assumed that the regression function η has a distribution $GP(\mu(\cdot), R(\cdot, \cdot; \beta))$ conditional on β and we provided consistency results by taking into account the prior distribution of β .

However, there are still other hyperparameters to be estimated in the covariance function. For example, the squared exponential covariance function can have an additional parameter, λ , as in the form of

$$R_\lambda(s, t) = \lambda \exp\left(-\frac{\beta^2(s-t)^2}{2}\right). \quad (4.1)$$

This additional parameter, λ , can be treated as a scaling parameter of the covariance function, and it can regulate the variability of the Gaussian process. For the Matérn covariance function in Table 2.1, the order ν is an extra parameter that needs to be estimated.

It is typically the case in applications that the various parameters that govern the smoothness of the GP prior are not sufficiently well understood to be chosen with certainty. In such cases, we choose a prior distribution Π_Λ for the additional parameter Λ , and treat Λ as another random quantity to be estimated.

Accordingly, the covariance function that we considered in the previous chapter, is the covariance function with a fixed λ , and thus posterior consistency results in Section 3.4.1 can be regarded as results when the hyperparameter, λ , is given. The natural question to be answered next is how the posterior consistency will be affected if we treat λ as random.

It is true that every result that holds with probability 1 conditional on λ for all $\lambda \in \Lambda$, holds with probability 1 marginally. However, the posterior distribution of η that is

computed when λ is treated as a parameter is not the same as the conditional posterior given $\Lambda = \lambda$, but rather it is the mixture of those posteriors with respect to the posterior distribution of Λ . Additional work will be required to deal with this case. In other words, we know that given $\Lambda = \lambda$, the posterior distribution of η is consistent, i.e.

$$\Pi \{ U_n^C \mid Y_1, \dots, Y_n, x_1, \dots, x_n, \lambda \} \rightarrow 0 \quad \text{a.s.}[P_0]. \quad (4.2)$$

The goal in this section is to show that under suitable conditions,

$$\Pi \{ U_n^C \mid Y_1, \dots, Y_n, x_1, \dots, x_n \} \rightarrow 0 \quad \text{a.s.}[P_0]. \quad (4.3)$$

When Λ is treated as a discrete random quantity with finite support, the posterior distribution is a finite mixture of conditional posteriors and the consistency question becomes a trivial question. However, if we choose a continuous prior for Λ , then the situation becomes complicated. When the squared exponential covariance function or the Matérn covariance function is used, the additional hyperparameter λ must be estimated in the model and thus a prior distribution Π_Λ for λ , has to be specified. Since the support of Λ can be \mathbb{R}^+ , in this case, it is possible that there is a sequence of λ such that the conditions from the previous chapter to guarantee the posterior consistency deteriorate as we move along the sequence.

For example, there are various constants that we prove to exist that make certain quantities exponentially bounded (say by $\exp(-cn)$). What if, as we move λ along a sequence, the corresponding value of c goes to 0, even if it never actually reaches 0 anywhere in the parameter space? If the posterior of λ happens to concentrate near the end of that sequence, there is no telling what happens to consistency.

To prevent this kind of behavior, we consider additional assumptions about the support of Λ and the smoothness of the covariance function as a function of λ . Specifically, the assumption about the support of Λ is that it is a compact set in \mathbb{R}^+ . We assume that the covariance function is smooth enough so that the supremum of the variance of the Gaussian process is continuous with respect to the hyperparameter, λ . Compactness of the support in

the Λ space, is designed to prevent the problem described above by making every λ close to one of finitely many λ 's so that the same argument as in the finite support case should work. Of course, the type of closeness that we need requires continuity as a function of λ , which is where the smoothness condition on the covariance function arises. In fact, what really needs to be continuous in λ is the various exponential bounds that arise out of the covariance function. The compactness condition would be a little strong and could have been weakened as in Ghosal and Roy (2005) with additional conditions when the hyperparameter λ has the form of (4.1). However, there still exist other types of hyperparameters of covariance function, such as the order parameter, v , in Matérn covariance function. We try to generalize the condition for additional hyperparameters. The following theorem formalizes these ideas and states the assumptions.

Theorem 4.1.1. *Let A_Λ be the compact support of Λ and define two functions of λ , $\rho_1(\lambda) = \sup_{x \in [0,1]} R_0(0; \lambda)$ and $\rho_2(\lambda) = \sup_{x \in [0,1]} -\beta^2 R_0''(0; \lambda)$. Suppose that $\rho_1(\lambda)$ and $\rho_2(\lambda)$ are continuous functions of λ , $\forall \lambda \in A_\Lambda$. Then, the posterior probability*

$$\sup_{\lambda \in A_\Lambda} \Pi \{ U_n^C | (\mathbf{Y}, \mathbf{x}), \lambda \} \rightarrow 0 \quad \text{a.s.}[P_0], \quad (4.4)$$

where $(\mathbf{Y}, \mathbf{x}) \equiv (Y_1, \dots, Y_n, x_1, \dots, x_n)$.

In addition, the marginal posterior (4.3) converges to 0 almost surely, as a consequence of (4.4). That is,

$$\Pi \{ U_n^C | Y_1, \dots, Y_n, x_1, \dots, x_n \} \rightarrow 0 \quad \text{a.s.}[P_0].$$

Proof. Since the posterior probability (4.4) can be expressed as in (3.5)

$$\begin{aligned} \Pi(U_n^C | Z_1, \dots, Z_n) &\leq \Phi_n + \frac{(1 - \Phi_n) \int_{U_n^C \cap \Theta_n} \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta) + \int_{U_n^C \cap \Theta_n^C} \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta)}{\int_{\Theta} \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta)} \\ &= \Phi_n + \frac{I_{1n}(Z_1, \dots, Z_n) + I_{2n}(Z_1, \dots, Z_n)}{I_{3n}(Z_1, \dots, Z_n)}, \end{aligned}$$

we have the same inequality for $\sup_{\lambda} \Pi(U_{\epsilon}^C | (\mathbf{Y}, \mathbf{x}), \lambda)$ as follows.

$$\begin{aligned} \sup_{\lambda \in A_{\Lambda}} \Pi(U_n^C | (\mathbf{Y}, \mathbf{x}), \lambda) &\leq \sup_{\lambda \in A_{\Lambda}} \Phi_n + \sup_{\lambda \in A_{\Lambda}} \frac{I_{1n}((\mathbf{Y}, \mathbf{x}), \lambda)}{I_{3n}((\mathbf{Y}, \mathbf{x}), \lambda)} + \sup_{\lambda \in A_{\Lambda}} \frac{I_{2n}((\mathbf{Y}, \mathbf{x}), \lambda)}{I_{3n}((\mathbf{Y}, \mathbf{x}), \lambda)} \end{aligned} \quad (4.5)$$

As we have seen in the proof of Theorem 3.3.1, asymptotic bounds for Φ_n and I_{1n} do not involve the covariance function and the specific choice of λ . Although the prior positivity condition is required in order to obtain the correct asymptotic bound for I_{3n} , and it entails the covariance function of the Gaussian process, the right asymptotic bound is obtained under the continuity of $\rho_2(\lambda)$ and the compactness of A_{Λ} as in the hypotheses of the theorem. In addition, the actual bound for I_{3n} does not involve λ . Correspondingly, the only concern of terms in (4.5) lies in the quantity of $\sup_{\lambda \in A_{\Lambda}} I_{2n}((\mathbf{Y}, \mathbf{x}), \lambda)$. Lemma 3.5.7 and its proof can be restated with $\rho_1(\lambda)$ and $\rho_2(\lambda)$ as follows: For sufficiently large n , there exist constants d_1 and d_2 which depend on λ , such that

$$\begin{aligned} \Pr \left\{ \sup_{x \in [0,1]} |\eta(x)| > M_n \middle| \lambda \right\} &\leq \exp(-d_1 n) \\ \Pr \left\{ \sup_{x \in [0,1]} |\eta'(x)| > M_n \middle| \lambda \right\} &\leq \exp(-d_2 n), \end{aligned}$$

where $d_1 = \frac{1}{4R_0(0;\lambda)}$ and $d_2 = -\frac{1}{4R_0''(0;\lambda)}$.

Furthermore, since we assumed $\rho_1(\lambda)$ and $\rho_2(\lambda)$ are continuous on the compact set A_{Λ} , it is obvious that they are uniformly bounded. In other words, there exist universal constants Ξ_1, Ξ_2, Ξ_3 and Ξ_4 such that

$$0 < \Xi_1 \leq \sup_{\lambda \in A_{\Lambda}} |\rho_1(\lambda)| \leq \Xi_2 \quad \text{and} \quad 0 < \Xi_3 \leq \sup_{\lambda \in A_{\Lambda}} |\rho_2(\lambda)| \leq \Xi_4$$

Consequently, the uniformly asymptotic bound for I_{3n} is achieved as in (4.6)

$$\begin{aligned} \sup_{\lambda \in A_{\Lambda}} \Pr \left\{ \sup_{x \in [0,1]} |\eta(x)| > M_n \right\} &\leq \exp \left(-\frac{n}{4\Xi_2} \right) \\ \sup_{\lambda \in A_{\Lambda}} \Pr \left\{ \sup_{x \in [0,1]} |\eta'(x)| > M_n \right\} &\leq \exp \left(-\frac{n}{4\Xi_4} \right). \end{aligned} \quad (4.6)$$

Hence, it is proven that

$$\sup_{\lambda} \Pi \{ U_n^C | (\mathbf{Y}, \mathbf{x}), \lambda \} \rightarrow 0 \text{ a.s. } [P_0].$$

Finally, let us consider the marginal posterior of (4.3). The marginal posterior is expressed as the integration of the conditional posterior with respect to the posterior of λ given (\mathbf{Y}, \mathbf{x}) as in (4.7)

$$\Pi \{ U_n^C | (\mathbf{Y}, \mathbf{x}) \} = \int_{A_\Lambda} \Pi \{ U_n^C | (\mathbf{Y}, \mathbf{x}), \lambda \} d\Pi(\Lambda | (\mathbf{Y}, \mathbf{x})). \quad (4.7)$$

Since the supremum of the conditional probability in (4.4) converges to zero almost surely $[P_0]$, the marginal posterior converges to zero almost surely $[P_0]$ regardless of the asymptotic distribution of $\Pi(\lambda|D)$ when the sample size increases as follows.

$$\begin{aligned} \Pi \{ U_n^C | Y_1, \dots, Y_n, x_1, \dots, x_n \} &= \int_{A_\Lambda} \Pi \{ U_n^C | (\mathbf{Y}, \mathbf{x}), \lambda \} d\Pi(\Lambda | (\mathbf{Y}, \mathbf{x})) \\ &\leq \int_{A_\Lambda} \sup_{\lambda \in A_\Lambda} \Pi \{ U_n^C | (\mathbf{Y}, \mathbf{x}), \lambda \} d\Pi(\Lambda | (\mathbf{Y}, \mathbf{x})) \\ &\leq \sup_{\lambda \in A_\Lambda} \Pi \{ U_n^C | (\mathbf{Y}, \mathbf{x}), \lambda \} \int_{\lambda \in A_\Lambda} d\Pi(\Lambda | (\mathbf{Y}, \mathbf{x})) \\ &\rightarrow 0, \text{ a.s. } [P_0]. \end{aligned} \quad (4.8)$$

Note that (4.8) holds since $\int_{\lambda \in A_\Lambda} d\Pi(\Lambda | (\mathbf{Y}, \mathbf{x})) = 1$. ■

In general, whenever an additional hyperparameter is introduced to be estimated in the covariance function, we can verify posterior consistency by assuming a compact support of the hyperparameter and the proper continuity condition with respect to the hyperparameter.

4.2 Multi-dimensional Covariates

Up to this point, we assumed that the covariate is one dimensional, but much concern also lies in multi-dimensional covariates. In practice, the GP regression approach has been used in spatial modeling (Paciorek (2003) or Paciorek and Schervish (2004)), where the

covariate is a two dimensional vector. In particular, the empirical evidence of the Bayesian GP spatial models has shown that they outperform other competing approaches. In terms of the posterior consistency, Ghosal and Roy (2005) considered the case of multidimensional covariates for nonparametric binary regression under a certain topology of multidimensional probability functions. In this section, we also extend our results of posterior consistency to the case of multidimensional covariates, under suitable conditions regarding the regression function and design points.

In general, the main difficulty in dealing with multidimensional or high-dimensional regression function lies in the so called phenomenon, the “curse of dimensionality”. Generally speaking, this means that the problem gets harder very quickly as the dimension of the observations increase (Wasserman (2005)). This problem also affects the posterior consistency and makes the results less promising. As the sample size increases, the posterior is also expected to be consistent in high dimensions. However, in order to achieve the posterior consistency in the same fashion as the one-dimensional regression function, the sample size should be much larger than that in one dimension, which is not practical. As a result, stronger assumptions on design points or regression functions are required, or different topologies in the parameter space can be considered. For example, Ghosal and Roy (2005) treated L^1 consistency of a probability function in one dimension with an assumption about the fixed design points, while in higher dimensions, they considered the consistency with respect to the empirical measure of the design points.

Our approach is to use stronger assumptions than those for the case of a one dimensional regression function, stated in Chapter 3. We describe the d -dimensional regression model and our stronger assumptions, here. Consider the same model as in (3.1) with d -dimensional covariates, i.e.

$$Y_i = \eta(X_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where $X_i = (X_{i,1}, \dots, X_{i,d}) \in [0, 1]^d$ is a d -dimensional covariate, $d \geq 2$ and $\eta(\cdot)$ is the

unknown d -dimensional regression function, $\eta(x) : \mathbb{R}^d \rightarrow \mathbb{R}$.

In order to extend previous consistency results to d -dimensional regression function, we must modify the basic assumptions, (Assumption **NRD**, **RD**, **P** and **B**) and adjust them to d -dimensional posterior consistency results. For this purpose, we restate the basic assumptions, tailored to d -dimensional covariates¹.

Assumption \mathbf{RD}_d . The random design points (covariates) $\{X_n\}_{n=1}^\infty$ are independent and identically distributed with probability distribution Q on $[0, 1]^d$.

Assumption \mathbf{NRD}_d . For each hypercube H in $[0, 1]^d$, let $\lambda(H)$ be its Lebesgue measure. Suppose that there exists a constant K_d , $0 < K_d \leq 1$ such that whenever $\lambda(H) \geq \frac{1}{K_d n}$, H contains at least one design point.

Note that the assumption **RD** _{d} is a direct extension to a d -dimensional covariate and the assumption **NRD** _{d} is easily satisfied by the equally spaced design. That is, suppose that the points, $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,d})$, $i = 1, \dots, n$ are spaced so that that $|x_{i,k} - x_{j,k}| = \frac{1}{n^{1/d}}$, $i, j = 1, \dots, n$, $k = 1, \dots, d$ and $0 = x_{i,0} < x_{i,1} \leq \dots \leq x_{i,n} < x_{i,n+1} = 1$. Then each hypercube H with Lebesgue measure at least $1/n$ contains at least one design point.

Assumption \mathbf{P}_d .

P2 _{d} Let $\tilde{\beta} = (\beta_1, \dots, \beta_d)$. The covariance function, $R(\mathbf{x}, \mathbf{x}'; \tilde{\beta})$ is a product of d isotropic and integrable covariance functions, one for each dimension.

$$R(\mathbf{x}, \mathbf{x}'; \tilde{\beta}) = R^{(1)}(x_1, x'_1; \beta_1) R^{(2)}(x_2, x'_2; \beta_2) \dots R^{(d)}(x_d, x'_d; \beta_d),$$

where each $R^{(i)}(x_i, x'_i; \beta_i) = R_{0,i}(\beta_i |x_i - x'_i|)$, where $R_{0,i}$ is a positive multiple of density.

P3 _{d} The mean function $\mu(x)$ of the Gaussian process $\eta(x)$ is continuously differentiable and $R_0(x)$ has continuous partial derivatives up to order $2d + 2$.

¹For notational convenience, we use the subscript d , to indicate the assumptions for d -dimension and only restate the assumptions that need to be modified.

P5_d β_j has a prior distribution, κ_j , with support \mathbb{R}^+ , and there exists $0 < \delta < 1/2$ and $b_1, b_2 > 0$ such that

$$\kappa \left\{ \beta_j > n^\delta \right\} = \Pr \left\{ \beta_j > n^\delta \right\} < b_1 \exp(-b_2 n), \quad \forall n \geq 1, j = 1, \dots, d.$$

Note that the product of isotropic covariance functions is also a valid isotropic covariance function as defined in P2 (See Bochner's Theorem in Breiman (1968, p. 174) or Schabenberger and Gotway (2005, pp. 43–44)).

Assumption B_d1. Let Π_1^* and ν be a Gaussian process and a prior on σ satisfying Assumption P_d. Let V be a constant. $\Omega = \left\{ \eta : \left\| \frac{\partial}{\partial x_j} \eta(x_1, \dots, x_d) \right\|_\infty < V, j = 1, \dots, d \right\}$ with $V > \left\| \frac{\partial}{\partial x_j} \eta_0(x_1, \dots, x_d) \right\|_\infty, j = 1, \dots, d$. Assume that $\Pi_1(\cdot) = \Pi_1^*(\cdot \cap \Omega) / \Pi_1^*(\Omega)$ with $\Pi_1^*(\Omega) > 0$.

Assumption B_d2. Let Π_1^{**} and ν be a Gaussian process and a prior on σ satisfying Assumption P_d. Let V_0 and V be a constant.

$$\Omega = \left\{ \eta : \|\eta\|_\infty < V_0, \left\| \frac{\partial}{\partial x_j} \eta(x_1, \dots, x_d) \right\|_\infty < V, j = 1, \dots, d \right\}$$

with $V_0 > \|\eta_0\|_\infty$ and $V > \left\| \frac{\partial}{\partial x_j} \eta_0(x_1, \dots, x_d) \right\|_\infty, j = 1, \dots, d$. Assume that $\Pi_1(\cdot) = \Pi_1^{**}(\cdot \cap \Omega) / \Pi_1^{**}(\Omega)$.

Note that Assumption B_d1 is weaker than Assumption B_d2. Assumption B_d1 will be used for L^1, d_Q and Hellinger consistency under Assumption NRD_d while Assumption B_d2 will be used for L^1 consistency under Assumption RD_d.

As described earlier, when the dimension of a covariate grows, the problem of estimating the regression function becomes more complicated and so do our results. The most challenging part is to construct appropriate tests, in particular satisfying (3.21) for the d -dimensional case, namely.

$$\sum_{j=1}^n |\eta_{1j} - \eta_{0j}| > rn,$$

The following two lemmas show that under Assumption B_d1 and NRD_d , there exist enough number of design points to acquire (3.21) in d dimensions.

Lemma 4.2.1. *Assume Assumptions NRD_d and B_d1 . Let λ be the Lebesgue measure in \mathbb{R}^d . Let K_d and V be the constants mentioned in Assumption NRD_d and B_d1 . Let A_V be the set of all continuous functions γ such that $\forall \mathbf{x}_1, \mathbf{x}_2 \in [0, 1]^d$, $|\gamma(\mathbf{x}_1) - \gamma(\mathbf{x}_2)| \leq V \|\mathbf{x}_1 - \mathbf{x}_2\|$. For each such function γ and $\epsilon > 0$, define $B_{\epsilon, \gamma} = \{x : |\gamma(x)| > \epsilon\}$. Then for each $\epsilon > 0$ there exists an integer N such that, for all $n \geq N$ and all $\gamma \in A_V$,*

$$\sum_{i=1}^n |\gamma(x_i)| \geq n \lambda(B_{\epsilon, \gamma}) \frac{\epsilon}{2}. \quad (4.9)$$

Proof. Let $\epsilon > 0$ and let N be large enough so that $\frac{\sqrt{d}V}{n^{1/d}} < \epsilon/2$ for all $n \geq N$. Let $\gamma \in A_V$. Since $B_{\epsilon, \gamma}$ is an open set in \mathbb{R}^d , it can be approximated by a finite collection of disjoint and equally-sized hypercubes, $\{H_{\epsilon, \gamma}^k\}_{k=1}^{N_{\epsilon, \gamma}}$. Specifically,

$$\lambda(H_{\epsilon, \gamma}^k) = \frac{1}{K_d n}, \quad H_{\epsilon, \gamma}^k \cap H_{\epsilon, \gamma}^j = \emptyset \quad (j \neq k), \quad H_{\epsilon, \gamma}^k \cap B_{\epsilon, \gamma} \neq \emptyset \quad \text{and} \quad B_{\epsilon, \gamma} \subset \bigcup_{k=1}^{N_{\epsilon, \gamma}} H_{\epsilon, \gamma}^k$$

By the assumption NRD_d , each hypercube of $\{H_{\epsilon, \gamma}^k\}_{k=1}^{N_{\epsilon, \gamma}}$ contains at least one design point. Let x_k^* be a design point in $H_{\epsilon, \gamma}^k$. If x_k^* is also in $B_{\epsilon, \gamma}$, then it is obvious that $|\gamma(x_k^*)| > \epsilon/2$. If $x_k^* \in H_{\epsilon, \gamma}^k \cap B_{\epsilon, \gamma}^c$, then it is still true that $|\gamma(x_k^*)| > \epsilon/2$ as follows: Let x^* be a point in $H_{\epsilon, \gamma}^k \cap B_{\epsilon, \gamma}$. Since x_k^* is in $H_{\epsilon, \gamma}^k$, the distance between x_k^* and x^* is less than $\frac{\sqrt{d}}{(K_d n)^{1/d}}$, i.e.

$$\|x_k^* - x^*\|_d \leq \frac{\sqrt{d}}{(K_d n)^{1/d}},$$

where $\|\cdot\|_d$ is the Euclidean distance.

In addition, by the assumption B_d1 , it is clear that

$$\gamma(x_k^*) \geq \gamma(x^*) - \frac{\sqrt{d}V}{(K_d n)^{1/d}} > \epsilon - \frac{\sqrt{d}V}{(K_d n)^{1/d}} > \frac{\epsilon}{2}.$$

Thus, there are $N_{\epsilon, \gamma}$ design points, $\{x_k^*\}_{k=1}^{N_{\epsilon, \gamma}}$ such that $|\gamma(x_k^*)| > \epsilon/2$. Therefore, $\sum_{k=1}^{N_{\epsilon, \gamma}} |\gamma(x_k^*)| > N_{\epsilon, \gamma} \frac{\epsilon}{2}$ because $\lambda \left(\bigcup_{k=1}^{N_{\epsilon, \gamma}} H_{\epsilon, \gamma}^k \right) = \frac{1}{K_d n} N_{\epsilon, \gamma} \geq \lambda(B_{\epsilon, \gamma})$. Hence,

$$\sum_{i=1}^n |\gamma(x_i)| \geq \{n\lambda(B_{\epsilon, \gamma})\} \frac{\epsilon}{2}, \quad \forall d \geq 2.$$

■

Lemma 4.2.2. *Assume Assumption NRD_d and B_d1 . Let A_V be the set of all continuous functions η such that $\|\eta\|_\infty < M_n$ and $\|\frac{\partial}{\partial x_j} \eta(x_1, \dots, x_d)\|_\infty < V$, $j = 1, \dots, d$. Then for each $\epsilon > 0$ there exist an integer N and $r > 0$ such that, for all $n \geq N$ and all $\eta \in A_V$ such that $\|\eta - \eta_0\|_1 > \epsilon$, $\sum_{i=1}^n |\eta(x_i) - \eta_0(x_i)| \geq rn$.*

Proof. Let V_0 be an upper bound of η_0 and let V be an upper bound on all partial derivatives of η_0 . Let $0 < \delta < \epsilon$. Let $r = (\epsilon - \delta)$ and $D_i = \{x : (i-1)\delta < |\eta(x) - \eta_0(x)| < i\delta\}$.

Let λ be Lebesgue measure. Then, $\|\eta - \eta_0\|_1$ can be bounded as follows.

$$\sum_i i\delta\lambda(D_i) \geq \|\eta - \eta_0\|_1 > \epsilon. \quad (4.10)$$

Let $\zeta(x) = |\eta(x) - \eta_0(x)|$ and $\zeta_m(x) = \min\{m\delta, \zeta(x)\}$, for $m = 0, \dots, n$. Note that $\zeta_{\lceil M_n + V_0/\delta \rceil}(x)$ is the same as $\zeta(x)$.

For $m = 1, \dots, n$, define $B_m \equiv \{x : \zeta_m(x) > (2m-1)\delta/2\}$. Then, for all $x \in B_m$, $\zeta_m(x) - \zeta_{m-1}(x) > \delta/2$. Thus, Lemma 4.2.1 (with $\gamma = \zeta_m - \zeta_{m-1}$ and $\epsilon = \delta/2$) implies

$$\sum_{i=1}^n (\zeta_m(x_i) - \zeta_{m-1}(x_i)) \geq \{n\lambda(B_m)\} \frac{\delta}{4}.$$

Now, write

$$\begin{aligned}
\sum_{i=1}^n |\eta(x_i) - \eta_0(x_i)| &= \sum_{i=1}^n \zeta_{\lceil (M_n + V_0)/\delta \rceil}(x_i) \\
&= \sum_{i=1}^n \sum_{m=1}^{\lceil (M_n + V_0)/\delta \rceil} \{\zeta_m(x_i) - \zeta_{m-1}(x_i)\} \\
&= \sum_{m=1}^{\lceil (M_n + V_0)/\delta \rceil} \sum_{i=1}^n \{\zeta_m(x_i) - \zeta_{m-1}(x_i)\}, \\
&\geq \sum_{m=1}^{\lceil (M_n + V_0)/\delta \rceil} [\lambda(B_m)n] \frac{\delta}{4}.
\end{aligned} \tag{4.11}$$

Also, for $m = 1, \dots, n$,

$$B_m \supset \bigcup_{i=m+1}^{\lceil (M_n + V_0)/\delta \rceil} D_i.$$

It follows that

$$\begin{aligned}
\sum_{m=1}^{\lceil (M_n + V_0)/\delta \rceil} \delta \lambda(B_m) &\geq \sum_{i=2}^{\lceil (M_n + V_0)/\delta \rceil} (i-1) \delta \lambda(D_i) \\
&\geq \left\{ \sum_{i=2}^{\lceil (M_n + V_0)/\delta \rceil} i \delta \lambda(D_i) \right\} - \delta \\
&\geq \epsilon - \delta,
\end{aligned}$$

where the last inequality follows from (4.10). Combining this with (4.11) gives

$$\sum_{i=1}^n |\eta(x_i) - \eta_0(x_i)| \geq n(\epsilon - \delta),$$

for all $n \geq N$. ■

Furthermore, under the assumption P_d , the Gaussian process and its partial derivative processes have continuous sample paths with probability one, which is stated in Lemma 4.2.3.

Lemma 4.2.3. *Given the smoothing parameters, $\tilde{\beta} = (\beta_1, \dots, \beta_d)$, let $\eta(\cdot)$ be a mean zero Gaussian process on $[0, 1]^d$ with a covariance kernel $R(\cdot, \cdot; \tilde{\beta})$ which satisfies Assumption P_d .*

Then $\eta(\cdot)$ has sample paths with continuous partial derivatives and the first order partial derivative processes $\frac{\partial}{\partial x_i}\eta(x_1, \dots, x_d)$, $i = 1, \dots, d$ are also Gaussian processes.

Proof. It is obvious from Theorem 5 in Ghosal and Roy (2005). ■

Now, we state the main theorem that the posterior probability regarding the unknown d -dimensional regression function is also consistent with respect to the L^1 metric for the regression function.

Theorem 4.2.1. *Suppose that the values of the covariate in $[0, 1]^d$ arise according to a design satisfying Assumption NRD_d . Assume that the prior satisfies Assumption B_d1 . Let P_0 denote the joint conditional distribution of $\{Y_n\}_{n=1}^\infty$ assuming that η_0 is the true response function and σ_0^2 is the true noise variance. Assume that the function η_0 has continuous first order partial derivatives. Then for every $\epsilon > 0$,*

$$\Pi \{ W_\epsilon^C \mid Y_1, \dots, Y_n, X_1, \dots, X_n \} \rightarrow 0 \quad a.s.[P_0],$$

Proof. Since the proof involves verifying two conditions (A1) and (A2) of Theorem 3.3.1 as in the one dimensional case, we follow the same steps as in Section 3.5.

(1) Prior positivity condition

Lemma 3.5.1 holds, regardless of the dimensionality of the covariate and Lemma 3.5.3 also holds from the assumption $P3_d$ and Result 3 in Tokdar and Ghosh (2004), which will be stated in Appendices.

(2) Construction a Sieve

Since we deal with set of regression functions, we need to consider a sieve to construct a test. Assumption B_d1 considers the set of all continuously differentiable regression functions with uniformly bounded partial derivatives. Thus, the sieve under consideration can be constructed with growing upper bounds for regression functions, as

follows. Let $M_n = O(n^{\alpha_1})$ for some $1/2 < \alpha_1 < 1$ and define $\Theta_n = \Theta_{2n} \times \mathbb{R}^+$, where

$$\Theta_{2n} = \{\eta(\cdot) : \|\eta\|_\infty < M_n\}. \quad (4.12)$$

From Lemma 3.5.4, we have that there exists K'' such that the ϵ -covering number $\log N(\epsilon, \Theta_{2n}, \|\cdot\|_\infty) \leq (K''M_n)/\epsilon$. In addition, since Θ_{2n} is only based on $\eta(x)$, the probability bound of Θ_{2n}^C does not involve β . Therefore, from Lemmas 3.5.7 and 3.5.8, it follows that Θ_{2n}^C has exponentially small probability.

(3) Existence of tests

The same test function in Section 3.5.3 is used in the d -dimensional case if we have

$$\sum_{j=1}^n |\eta_{1j} - \eta_{0j}| > rn.$$

This was guaranteed by Lemma 4.2.2. In addition, the existence of continuous sample paths of the Gaussian process and its derivative are also ensured by the assumption P_d and Theorem 5 in Ghosal and Roy (2005).

■

Theorem 4.2.2. *Suppose that the values of the covariate in $[0, 1]^d$ arise according to a design satisfying Assumption RD_d . Assume that the prior satisfies Assumption B_d2 . Let P_0 denote the joint conditional distribution of $\{Y_n\}_{n=1}^\infty$ assuming that η_0 is the true response function and σ_0^2 is the true noise variance. Assume that the function η_0 has continuous first order partial derivatives. Then for every $\epsilon > 0$,*

$$\Pi \{W_\epsilon^C | Y_1, \dots, Y_n, X_1, \dots, X_n\} \rightarrow 0 \quad \text{a.s.}[P_0],$$

Proof. As a corollary of Theorem 4.2.1,

$$\Pi \{U_\epsilon^C | Y_1, \dots, Y_n, X_1, \dots, X_n\} \rightarrow 0 \quad \text{a.s.}[P_0].$$

In addition, by Assumption B_d2, we have the regression functions are uniformly bounded. Therefore, it is obvious from the fact that in-probability convergence is equivalent to L^1 convergence in this case. ■

Chapter 5

Conclusions and Future work

5.1 Summary and Contributions

The main contribution of this thesis is the theoretical justification of GP regression in terms of posterior consistency. We established posterior consistency of GP regression under various model structures, and our major contribution is the extension of the results of Ghosal and Roy (2005). Ghosal and Roy (2005) established in-probability consistency in the nonparametric binary regression problem using Gaussian process priors. We generalized their results to almost sure consistency of nonparametric regression problems with continuous data.

We provided an extension of the posterior consistency theorem of Schwartz (1965) for independent but non-identically distributed observations, and we applied our theorem to the Gaussian process regression problem by stipulating conditions on the Gaussian process prior that allow us to verify the conditions of the consistency theorem, namely the prior positivity of neighborhoods of the true parameter and the existence of tests with suitable properties. The tests that we constructed for this purpose were designed independently of the specific prior distribution that we used, and these tests can be used for posterior

consistency of other nonparametric regression models with additive noise as long as the prior distribution and the regression function under consideration meet our conditions.

We proved almost sure consistency of posterior probabilities of three different joint neighborhoods of the true regression function and a suitable neighborhood of an unknown noise variance in additive nonparametric regression problems using Gaussian process priors. As a corollary, posterior consistency of the marginal distribution of the regression function is also obtained. We have also verified that the conditions for consistency hold for several classes of priors that are already used in practice.

We found that the case of random covariates is more challenging than nonrandom covariates. This is due to the fact that it is difficult to insure that the covariates will spread themselves uniformly enough to obtain consistency at all smooth true regression functions. The problem arises when we consider regression functions with an arbitrarily large upper bound. In this case, we can find functions η that are far from the true regression function η_0 in L^1 distance but differ from η_0 very little over almost all of the covariate space. A random sample of covariates will not have much chance of containing sufficiently many points x such that $|\eta(x) - \eta_0(x)|$ is large. Such covariates are needed in order to define uniformly consistent tests. The metric d_Q declares such η functions to be close to η_0 while the L^1 metric might declare them to be far apart. Also, when the noise distribution is normal, functions that are close to η_0 in d_Q distance produce similar joint distributions for the covariate and the response in terms of Hellinger distance. These distinctions disappear when the space of possible regression functions is known to be uniformly bounded a priori.

In addition, we also studied how much posterior consistency can be extended into other regression model structures and provided some additional results under appropriate conditions. We included cases when there are additional hyperparameters of a covariance function to be estimated, when the covariate is multidimensional.

5.2 Future Work

We discuss several open issues that are worth further consideration but have not been studied in this dissertation.

First, we have said nothing about rates of convergence. When the posterior consistency is achieved, the next inherent question is to investigate the convergence rate of the posterior distribution. Ghosal and van der Vaart (2004) present general results on convergence rates for non i.i.d observations which include nonparametric regression cases. They mention the general results for nonparametric regression with the assumption that the regression function is uniformly bounded. We would like to extend their results to our model structures. Similar conditions to those used to prove posterior consistency are required to study the rate of convergence. The test functions that we constructed could also be used to analyze the rate that the posterior probability concentrates around the true parameter. The challenge is to find a rate that a prior probability shrinks as we consider a decreasing sequence of Kullback-Leibler neighborhoods. To be specific, this problem involves finding lower tail probabilities for Gaussian processes, known as a small ball problem or a small deviation problem. Historically, this problem has been studied mainly in the setting of Brownian motion or its variants. Most recently, Li and Shao (2004) derived lower tail probabilities for general Gaussian processes and stated suitable conditions on the covariance functions. However, their conditions for covariance functions are not compatible with ours, and we could not pursue their directions. Therefore, the important question to be answered is obtaining the exponentially small lower bound for small uniform balls of Gaussian processes.

Secondly, Assumption NRD is perhaps a bit strong. Ghosal and Roy (2005), in a problem with uniformly bounded regression functions, use a condition on the design points that is weaker than Assumption NRD. We have not tried to find the weakest condition that guarantees almost sure consistency.

Third, we would like to extend our results to non-normal data. For example, nonlinear Poisson regression or nonlinear autoregressive regression approaches have been attempted in practice but their theoretical questions have not been answered, in particular about posterior consistency. For posterior consistency, the prior positivity can be handled in the same way as in Gaussian process regression, regardless of the models under consideration whereas the existence of tests must be investigated carefully. By following our approach, we can construct test functions in a nonlinear Poisson regression problem in a similar way because the observations will be assumed to be independent. The difficulty comes from nonlinear autoregressive models where the observations are dependent and the Hoeffding type inequalities do not work in general. Ghosal and van der Vaart (2004) briefly mention the nonlinear autoregressive model (NLAR) for convergence rates but much remains to be done on NLAR. Tang and Ghosal (2003) deal with the general consistency theorem for ergodic Markov processes considering the Dirichlet mixture model for Markov processes. We can also try to apply or modify their results in NLAR under Gaussian process prior, but it may be a challenging approach.

Finally, there are subtle issues that we could not finish in this dissertation. A general approach to Bayesian misspecification in the infinite dimensional model can be found in Kleijn and van der Vaart (2002). The results of Kleijn and van der Vaart (2002) suggest that misspecification of the error distribution does not affect the asymptotic results of Bayesian nonparametric regression under suitable conditions. It would be interesting to investigate the extent to which misspecification of the error distribution matters in Gaussian process regression, following their direction. Another issue with potential for improvement is to establish L^1 consistency results under a weaker condition than the uniform boundedness of the regression function, such as the uniform L^1 or L^2 boundedness, which we have not tried yet.

We plan to address some of these issues in the near future, and are hopeful about where

the search will lead. We invite our colleagues, both those who are experienced as well as newcomers to the field, to add their insights to the research.

Appendices

A : Useful Facts

1. Mill's ratio

If $X \sim N(0, 1)$, then

$$\Pr(|X| \geq c) \leq 2\phi(c)/c \text{ or } 1 - \Phi(x) < \frac{1}{x}\phi(x),$$

where $\phi(x)$ is a standard normal density function and $\Phi(x)$ is a standard normal distribution function.

2. Noncentral chi-square distribution

(a) Ravishanker and Dey (2002, Result 5.3.2)

Let $\mathbf{x} \sim N_k(\mu, I)$ where $\mu' = (\mu_1, \dots, \mu_k) \neq \mathbf{0}$, and let $U = \mathbf{x}'\mathbf{x}$. The pdf of U is

$$f(u) = \sum_{j=0}^{\infty} \frac{\exp(-\lambda)\lambda^j}{j!} \frac{u^{(k+2j-2)/2} \exp\{-u/2\}}{2^{j+\frac{k}{2}} \Gamma(\frac{1}{2}(k+2j))}, \quad u > 0,$$

where $\lambda = \frac{1}{2}\mu'\mu = \frac{1}{2} \sum_{j=1}^k \mu_j^2$. We say $U \sim \chi^2(k, \lambda)$, i.e., U has a noncentral chi-square distribution with k degrees of freedom and noncentrality parameter equal to λ .

(b) **Ravishanker and Dey (2002, p. 166)**

The moment generating function of a random variable U is given as $M_U(t) = E[\exp(Ut)]$.

$$M_U(t) = (1 - 2t)^{-k/2} \exp\{-\lambda[1 - (1 - 2t)^{-1}]\}, \quad t < 1/2.$$

(c) **Ravishanker and Dey (2002, Example 5.3.3)**

Suppose $U \sim \chi^2(k, \lambda)$ and $V \sim \chi_k^2$. Then for some constant c ,

$$P(U > c) \geq P(V > c).$$

B : Useful Theorems1. **van der Vaart and Wellner (1996, Theorem 2.7.1)**

For $\alpha > 0$, define for any vector $k = (k_1, \dots, k_d)$ of d integers the differential operator

$$D^k = \frac{\partial^k}{\partial x_1^{k_1} \dots \partial x_d^{k_d}},$$

where $k_{\cdot} = \sum k_i$. Then for a function $f : \mathcal{X} \rightarrow \mathbb{R}$, let

$$\|f\|_{\alpha} = \max_{k \leq \underline{\alpha}} \sup_x |D^k f(x)| + \max_{k = \underline{\alpha}} \sup_{x, y} \frac{|D^k f(x) - D^k f(y)|}{\|x - y\|^{\alpha - \underline{\alpha}}},$$

where the suprema are taken over all x, y in the interior of \mathcal{X} with $x \neq y$ and $\underline{\alpha}$ is the greatest integer smaller than α . Let $C_M^{\alpha}(\mathcal{X})$ be the set of all continuous functions $f : \mathcal{X} \rightarrow \mathbb{R}$ with $\|f\|_{\alpha} \leq M$.

Let \mathcal{X} be a bounded, convex subset of \mathbb{R}^d with nonempty interior. There exists a constant K depending only on α and d such that

$$\log N(\epsilon, C_1^{\alpha}(\mathcal{X}), \|\cdot\|_{\infty}) \leq K \lambda(\mathcal{X}^1) \left(\frac{1}{\epsilon}\right)^{d/\alpha},$$

for every $\epsilon > 0$, where $\lambda(\mathcal{X}^1)$ is the Lebesgue measure of the set $\{x : \|x - \mathcal{X}\| < 1\}$.

2. van der Vaart and Wellner (1996, Proposition A.2.7)

For a given Gaussian process X indexed by an arbitrary set T , we denote by ρ its “natural semimetric” $\rho(s, t) = \sigma(X_s - X_t)$. Define $\sigma^2(X) = \sup_{t \in T} \text{var} X_t$ and $\sigma(X)$ is the supremum of the standard deviations $\sigma(X_t)$. Let $\bar{\Phi}(z) = \int_z^\infty \phi(x) dx \leq z^{-1} \phi(z)$ be the tail probability of a standard normal variable.

Let X be a separable mean-zero Gaussian process such that for some $K > \sigma(X)$, some $V > 0$ and some $0 < \epsilon_0 \leq \sigma(X)$,

$$N(\epsilon, T, \rho) \leq \left(\frac{K}{\epsilon} \right)^V, \quad 0 < \epsilon < \epsilon_0.$$

Then there exists a universal constant D such that, for all $\lambda \geq \sigma^2(X)(1 + \sqrt{V})/\epsilon_0$,

$$P \left(\sup_{t \in T} X_t \geq \lambda \right) \leq \left(\frac{DK\lambda}{\sqrt{V}\sigma^2(X)} \right)^V \bar{\Phi} \left(\frac{\lambda}{\sigma(X)} \right).$$

3. **Tokdar and Ghosh (2004, Theorem 3.4)** Let $\sigma(t, s)$ be a covariance function on $\mathbb{R} \times \mathbb{R}$. Given a $\beta \in \mathbb{R}^+$, let $W(t)$ be a separable Gaussian process on $[0, 1]$ with $\text{Cov}(W(t), W(s)|\beta) = \sigma_\beta(t, s) = \sigma(\beta t, \beta s)$. Assume

(A1) : $\exists M, m > 0$ such that $m \leq \sigma(t, t) \leq M$, $\forall t \in \mathbb{R}^+$.

(A2) : $\exists C > 0$ such that $[\sigma(t, t) + \sigma(s, s) - 2\sigma(t, s)]^{1/2} \leq C|s - t|$, $\forall s, t \in \mathbb{R}^+$.

(A3) : For any $n \geq 1$ and any $t_1, \dots, t_n \in \mathbb{R}^+$, $\Sigma = ((\sigma(t_i, t_j)))$ is non-singular.

(A4) : The mean function of $W(t)$, $\mu(t)$ is continuous.

Define the set

$$\mathcal{A}_R = \{f(t) = \sum_{i=1}^k a_i R(t, t_i^*; \beta), \text{ for some } \beta \in B, k \in \mathcal{N}, a_i \in \mathbb{R}, t_i^* \in Q \cap [0, 1]\},$$

where Q is the set of rationals in \mathbb{R} .

Let $\bar{\mathcal{A}}_R$ be the sup-norm closure of \mathcal{A} . Then, under the assumptions A1 through A4,

$$w_0 \in \bar{\mathcal{A}} \iff \forall \epsilon > 0, \Pr \left(\sup_{t \in [0,1]} |W(t) - w_0(t)| < \epsilon \right) > 0.$$

4. **Tokdar and Ghosh (2004, Result 2)** Suppose ϕ is a nowhere zero continuous density function on \mathbb{R} and define $R(t, s) = \phi(t - s)$. Then $\bar{\mathcal{A}}_R = C[0, 1]$.
5. **Tokdar and Ghosh (2004, Result 3)** Suppose $d \geq 2$, and

$$R(t, s) = R^{(1)}(t_1, s_1) R^{(2)}(t_2, s_2) \dots R^{(d)}(t_d, s_d),$$

for some functions $R^{(i)}(t, s)$, $1 \leq i \leq d$, on $[0, 1] \times [0, 1]$. If $\bar{\mathcal{A}}_{R^{(i)}} = C[0, 1]$ for each i , then $\bar{\mathcal{A}}_R = C[0, 1]$.

6. **Ghosal and Roy (2004, Theorem 4)**

Assume that $\{W(t), t \in T\}$ is a Gaussian process with continuous sample paths having mean function $\mu(t)$ and continuous covariance kernel $\sigma(s, t)$. Assume that $\mu(t)$ and a function $w(t)$ belongs to the Reproducing Kernel Hilbert Space (RKHS) of the kernel $\sigma(s, t)$, $\bar{\mathcal{A}}_\sigma$. Then

$$\Pr \left(\sup_{t \in T} |W(t) - w(t)| < \epsilon \right) > 0, \text{ for all } \epsilon > 0.$$

7. **Ghosal and Roy (2005, Theorem 4)**

Let $\eta(\cdot)$ be a Gaussian process on \mathcal{X} , a bounded set of \mathbb{R}^d . Assume that the mean function $\mu(\cdot)$ is in $C^\alpha(\mathcal{X})$ and the covariance kernel $\sigma(\cdot, \cdot)$ has $2\alpha + 2$ mixed partial derivatives for some $\alpha \geq 1$. Then $\eta(\cdot)$ has differentiable sample paths with mixed partial derivatives up to order α and the successive derivative processes $D^w \eta(\cdot)$ are

also Gaussian with continuous sample paths. Further, there exists a constant d_w such that

$$P\left(\sup_{x \in \mathcal{X}} |D^w \eta(x)| > M\right) \leq e^{-d_w M^2 / \sigma_w^2(\eta)}$$

for $w = \langle w_1, w_2, \dots, w_d \rangle$, $w_i \in \{0, 1, 2, \dots, \alpha\}$, $|w| \leq \alpha$ and $\sigma_w^2(\eta) = \sup_{x \in \mathcal{X}} \text{var}(D^w \eta(x)) < \infty$ and the covariance function of the derivative processes $D^w \eta(x)$ are functions of the derivatives of the covariance kernel $\sigma(\cdot, \cdot)$.

Bibliography

- Abrahamsen, P. (1997). A review of gaussian random fields and correlation functions. Technical Report 917, Norwegian Computing Center.
- Adler, R. J. (1981). *The Geometry of Random Fields*. Wiley, New York.
- Adler, R. J. (1990). *An introduction to continuity, extrema, and related topics for Gaussian processes*. IMS Lecture notes-monograph series vol. 12, Hayward, CA.
- Amewou-Atisso, M., Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (2003). Posterior consistency for semiparametric regression problems. *Bernoulli*, 9: 291–312.
- Anderson, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proc. Amer. Math. Soc.*, 6:170–176.
- Barron, A., Schervish, M. J., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27:536–561.
- Belitser, E. and Ghosal, S. (2003). Adaptive bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.*, 31: 536–559.
- Billingsley, P. (1995). *Probability and measure, Third ed.* Wiley, New York.
- Breiman, L. (1968). *Probability*. Addison-Wesley, MA.

- Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24:2384–2398.
- Choudhuri, N., Ghosal, S., and Roy, A. (2003). Bayesian methods for function estimation. In Dey, D., editor, *Handbook of Statistics*. Marcel Dekker, New York.
- Choudhuri, N., Ghosal, S., and Roy, A. (2004). Bayesian estimation of the spectral density of time series. *JASA*, 99.
- Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.*, 21: 903–923.
- Cramer, H. and Leadbetter, M. R. (1967). *Stationary and related stochastic processes*. Wiley, New York.
- Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *JRSS B.*, 60:333–350.
- Diggle, P. J., Ribeiro Jr., P. J., and Christensen, O. F. (2003). An introduction to model-based geostatistics. In Møller, J., editor, *Lecture notes in statistics*. 173. Springer-Verlag, New York.
- DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika.*, 88:183–201.
- Doob, J. L. (1949). Application of the theory of martingales. *Coll. Int. du C. N. R. S. Paris.*, pages 23–27.
- Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, 90.
- Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist.*, 27:1119–1141.

- Fuentes, M. and Smith, R. (2001). A new class of nonstationary spatial models. Department of Statistics, North Carolina state university.
- Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.*, 27:143–158.
- Ghosal, S. and Roy, A. (2005). Posterior consistency of Gaussian process prior for nonparametric binary regression. Preprint.
- Ghosal, S. and van der Vaart, A. W. (2004). Convergence rates of posterior distributions for noniid observations. Preprint.
- Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Sprinber-Verlag, New York.
- Gibbs, M. N. (1997). *Bayesian Gaussian processes for regression and classification*. PhD thesis, University of Cambridge.
- Girard, A. (2004). *Approximate methods for propagation of uncertainty with Gaussian processes models*. PhD thesis, University of Glasgow.
- Gu, C. (2002). *Smoothing splines ANOVA models*. Sprinber-Verlag, New York.
- Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Sprinber-Verlag, New York.
- Higdon, D., Swall, J., and Kern, J. (1998). Non-stationary spatial modeling. In Bernardo, J. M., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics 6*. Oxford Univ. Press, Oxford, U.K.
- Huang, T. M. (2004). Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.*, 32: 1556–1593.

- Ibrahim, J. G., Chen, M. H., and Sinha, D. (2001). *Bayesian survival analysis*. Springer-Verlag, New York.
- Kleijn, B. and van der Vaart, A. W. (2002). Misspecification in infinite dimensional Bayesian statistics. Preprint.
- Lenk, P. J. (1988). The logistic normal distribution for bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.*, 83.
- Lenk, P. J. (1991). Towards a practicable bayesian nonparametric density estimator. *Biometrika*, 78.
- Leonard, T. (1978). Density estimation, stochastic processes and prior information. *J. Roy. Statist. Soc. Ser. B.*, 40.
- Li, W. V. and Shao, Q. M. (2004). Lower tail probabilities for gaussian processes. *Ann. Probab.*, 33:216–242.
- Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83.
- Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag, New York.
- Neal, R. M. (1997). Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Department of Statistics and Department of Computer Science, University of Toronto.
- O’Hagan, A. (1978). Curve fitting and optimal design for prediction (with discussion). *JRSS B.*, 40:1–42.
- Paciorek, C. J. (2003). *Nonstationary Gaussian processes for regression and spatial modeling*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

- Paciorek, C. J. and Schervish, M. J. (2004). Nonstationary covariance functions for gaussian process regression. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems 16*. MIT Press, MA.
- Rasmussen, C. (1996). *Evaluation of Gaussian processes and other methods for non-linear regression*. PhD thesis, University of Toronto, Toronto, Canada.
- Ravishanker, N. and Dey, D. K. (2002). *A First Course in Linear Model Theory*. Chapman & Hall/CRC, New York.
- Schabenberger, O. and Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC, New York.
- Schwartz, L. (1965). On Bayes procedures. *Z. Wahr. Verw. Gebiete*, 4: 10–26.
- Seeger, M. (2003). *Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds*. PhD thesis, University of Edinburgh.
- Seeger, M. (2004). Bayesian Gaussian processes for machine learning. *International Journal of Neural Systems*, 14.
- Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.*, 29: 666–686.
- Shi, J., Murray-Smith, R., and Titterton, D. (2005). Hierarchical Gaussian process mixtures for regression. *Statistics and Computing*, 15.
- Stein, M. L. (1999). *Interpolation of Spatial Data : Some Theory for Kriging*. Springer, New York.
- Tang, Y. and Ghosal, S. (2003). Dirichlet mixture of normal models for markov process. Preprint.

- Tokdar, S. and Ghosh, J. K. (2004). Posterior consistency of Gaussian process priors in density estimation. Preprint.
- van de Geer, S. (2000). *Empirical Processes in M-Estimation*. CAMBRIDGE Univ. Press, Cambridge, UK.
- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.
- Wahba, G. (1978). Improper priors, spline smoothing and the problems of guarding against model errors in regression. *JRSS B.*, 40:364–372.
- Walker, S. G. (2003). Bayesian consistency for a class of regression problems. *South African Stat. Journal*, 37: 149–167.
- Walker, S. G. (2004). New approaches to bayesian consistency. *Ann. Statist.*, 32:2028–2043.
- Walker, S. G. and Damien, P. (2000). On Bayesian models and consistency. *Ann. Statist.*, 32:2028–2043.
- Wasserman, L. (2005). *All of Nonparametric statistics*. Springer, New York.
- Zhao, L. H. (2000). Bayesian aspects of some nonparametric problems . *Ann. Statist.*, 28: 532–552.