ABSTRACT

LI, WEI. Bayesian Inference about Some Geometric Aspects of Nonparametric Functions. (Under the direction of Subhashis Ghosal).

Nonparametric statistics have long been focusing on estimation and inference of some unknown smooth function $f$ itself. However, some geometric features of the function are also of great interest in many applications, such as the level curve $\{x : f(x) = c\}$ or the level set $\{x : f(x) \geq c\}$ and the filaments of functions. A filament, also called “ridge”, consists of local maximizers of $f$ when moving in a certain direction. Intuitively speaking, the objects of interest in these problems are sets of points, collectively forming some lower dimensional features of the function.

The $L^\infty$-convergence rates of some smooth multivariate function $f$ are useful for understanding the statistical properties of the estimation of level sets $\{x : f(x) = c\}$. In the Bayesian framework, we derive new $L^\infty$-posterior contraction rates for multivariate nonparametric function estimation in several different settings. For the multivariate Gaussian white noise model, the $L^\infty$-contraction rates for the functions using trigonometric series and wavelet series priors are separately derived. Since the collection of level sets for the square root of a nonnegative function coincides with that of the original function, the results based on the square roots of the functions can lead to useful results for level sets. For binary regression, Poisson regression and density estimation, the $L^\infty$-contraction rates for the square roots of the mean functions and the square root of the density function respectively, along with their derivatives are obtained. For these three examples, we use random B-spline series prior and the properties of B-splines functions to obtain the $L^\infty$-rates from the $L^2$-rates through a relaxation argument. All these $L^\infty$-rates yield the contraction rates for level sets in the Hausdorff distance and the Lebesgue measure of the symmetric difference. In addition, we study frequentist coverage properties of suitable credible regions for the level sets in both Gaussian white noise model using trigonometric basis prior and multivariate Gaussian nonparametric regression using B-splines basis prior. A simulation study shows that the credible regions proposed have sufficient frequentist coverage.

On filament estimation, there have been some recent theoretical studies in the context of nonparametric kernel density estimation. Our study of filaments contribute to the current literature in two ways. First, we provide a Bayesian approach to filament estimation in the multivariate nonparametric normal regression context and study posterior contraction rates using a finite random series of B-splines basis. Compared with the kernel-estimation method, this has theoretical advantage as the bias can be better controlled when the function is smoother, which allows obtaining better rates. Assuming that $f : \mathbb{R}^2 \mapsto \mathbb{R}$ belongs to an isotropic Hölder class...
of order $\alpha \geq 4$, with the optimal choice of smoothing parameters, the posterior contraction rates for the filament points on some appropriately defined integral curves and for the Hausdorff distance of the filament are both $(n/\log n)^{(2-\alpha)/(2(1+\alpha))}$. Second, we provide a method to construct a credible set with sufficient frequentist coverage for the filaments. We study the performance of our proposed method by simulations and application to an earthquake dataset.
Bayesian Inference about Some Geometric Aspects of Nonparametric Functions

by
Wei Li

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina
2018

APPROVED BY:

_________________________  _________________________
Anastasios Tsiatis           Rui Song

_________________________  _________________________
Brian Reich                 Subhashis Ghosal
Chair of Advisory Committee
The author was born in Guangzhou, China. Upon completion of his Bachelor’s degree from Capital University of Economics and Business in July 2007, he studied Economics at George Mason University, and later at Indiana University Bloomington. He completed a Master’s degree in Mathematics and achieved his Ph.D. candidacy in Economics at Indiana University Bloomington before he started pursuing his Ph.D. in Statistics at North Carolina State University in 2015.
ACKNOWLEDGEMENTS

First and foremost, I would like to offer my deepest gratitude to my advisor, Professor Subhashis Ghosal, for introducing me to the wonderful world of Bayesian nonparametrics and for his invaluable advice and guidance that made this work possible. It is a great privilege and fortune for me to share not only of his exceptional scientific knowledge, but also of his extraordinary professional qualities.

I would also like to thank the rest of my dissertation committee members Professor Anastasios Tsiatis, Professor Rui Song and Professor Brian Reich for their valuable suggestions and encouragement during my research. I would also like to thank Professor Meng Li who is an Assistant Professor at Rice University for sharing his insight regarding a problem I study in Chapter 4. I am grateful to Professors Dennis Boos, Marie Davidian, Len Stefanski and Anastasios Tsiatis for the excellent courses they taught.

This work would not be possible without broad inputs from Professors Anastasios Tsiatis, Len Stefanski, Rui Song, Sujit Ghosh, Howard Bondell and Yichao Wu at the early stage of my Ph.D. study. I have been benefited greatly from their tremendous intellectual insights which broadens my knowledge of statistics. I thank them for introducing me to their research specialties and their time and patience in problem discussions with me. I am also indebted to other faculty and staff members in the Department of Statistics at North Carolina State University for their enthusiastic teachings and assistance.

All my work was also enriched by the Raleigh Chinese Christian Church. During the past three years, I have been extremely fortunate to be blessed with the love and care from a group of wonderful fellows in the Agape Bible Study Group and the student fellowship of the Raleigh Chinese Christian Church. I would also like to offer my deep appreciation for them.

Last but not least, I would like to thank my parents, mother-in-law, my daughter and especially my wife, for their unending support, love and prayers.
# TABLE OF CONTENTS

List of Figures ................................................................. vi

Chapter 1 Introduction ......................................................... 1
  1.1 Bayesian nonparametrics ................................................. 1
  1.2 Posterior contraction ................................................... 3
  1.3 Credible regions and coverage ......................................... 6
  1.4 Research questions and contributions ................................. 8
  1.5 Notations and preliminaries ............................................. 9
    1.5.1 Notations .......................................................... 9
    1.5.2 Trigonometric basis .............................................. 10
    1.5.3 B-Splines Basis .................................................. 11
    1.5.4 Wavelet basis ................................................... 13
  1.6 Roadmap ................................................................. 17
    1.6.1 Chapter 2 ........................................................ 17
    1.6.2 Chapter 3 ........................................................ 17
    1.6.3 Chapter 4 ........................................................ 17
    1.6.4 Appendices ....................................................... 17

Chapter 2 Posterior Contraction and Credible Regions for Level Sets .... 18
  2.1 Introduction ............................................................ 18
  2.2 Preliminaries ........................................................... 19
  2.3 Applications in various inference problems ............................ 21
    2.3.1 Gaussian white noise model .................................... 22
    2.3.2 Gaussian regression .............................................. 24
    2.3.3 Binary regression ............................................... 26
    2.3.4 Poisson regression .............................................. 27
    2.3.5 Density estimation .............................................. 28
    2.3.6 Credible sets ................................................... 29
  2.4 Simulation ............................................................. 31
    2.4.1 Poisson regression .............................................. 32
    2.4.2 Nonparametric regression ..................................... 34
  2.5 Technical Proofs ...................................................... 37
    2.5.1 Proofs of the results for the Gaussian white noise model .... 37
    2.5.2 Proofs of the results for binary regression, Poisson regression and density estimation .................................................. 43
    2.5.3 Proofs of the results for credible sets ........................ 51

Chapter 3 Posterior Contraction and Credible Regions for Filaments of Regression Functions ........................... 55
  3.1 Introduction ............................................................ 55
  3.2 Preliminaries ........................................................... 57
  3.3 Model, prior and posterior ............................................. 58
3.4 Assumptions ................................................................. 59
3.5 Posterior contraction and credible sets for filaments ................. 61
3.6 Simulation ................................................................. 65
3.7 Application ................................................................. 68
3.8 Technical Proofs .......................................................... 70
  3.8.1 Some lemmas ......................................................... 70
  3.8.2 Proofs of the main results ......................................... 73

Chapter 4 Direct Bayesian Learning of Level Curves ......................... 89
  4.1 Introduction ............................................................. 89
  4.2 Model, prior and posterior ............................................ 89
  4.3 Discussion ............................................................... 92
  4.4 Simulation .............................................................. 93
  4.5 Future research ......................................................... 94

References ................................................................. 100

Appendices ................................................................. 108
  Appendix A Some auxiliary lemmas ...................................... 109
  Appendix B Posterior contraction rates using finite random series priors 113
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>B-spline functions. Left plot: $q = 2, N = 3$. Right plot: $q = 4, N = 3$.</td>
</tr>
<tr>
<td>1.2</td>
<td>Examples of some Daubechies scaling functions and wavelet functions.</td>
</tr>
<tr>
<td>2.1</td>
<td>The contour of mean function $f$ and true level curve (orange curve).</td>
</tr>
<tr>
<td>2.2</td>
<td>Poisson regression $\sqrt{f} = b^T \theta$ (effects of the smoothing parameter $J_1$ and $J_2$). The orange circle is the truth, the blue curve is estimated level curve induced by posterior mean. Top left: $J_1 = J_2 = 6$; top right: $J_1 = 8, J_2 = 6$; bottom left: $J_1 = J_2 = 10$; bottom right: $J_1 = J_2 = 12$.</td>
</tr>
<tr>
<td>2.3</td>
<td>Poisson regression $f = b^T \theta$ (effects of the smoothing parameter $J_1$ and $J_2$). The orange circle is the truth, the blue curve is the estimated level curve induced by posterior mean. Top left: $J_1 = J_2 = 6$; top right: $J_1 = 7, J_2 = 8$; bottom left: $J_1 = J_2 = 10$; bottom right: $J_1 = J_2 = 12$.</td>
</tr>
<tr>
<td>2.4</td>
<td>The function $f$ and its level curves.</td>
</tr>
<tr>
<td>2.5</td>
<td>Level curves of a nonparametric regression function (effects of the smoothing parameter $J_1$ and $J_2$). The orange circle is the truth, the blue curve is the estimated level curve induced by the posterior mean. Top left: $J_1 = J_2 = 7$; top right: $J_1 = J_2 = 9$; bottom left: $J_1 = J_2 = 11$; bottom right: $J_1 = J_2 = 15$.</td>
</tr>
<tr>
<td>2.6</td>
<td>Level curves of a nonparametric regression function (uncertainty Quantification). Left: $J_1 = J_2 = 7$. Right: $J_1 = J_2 = 9$.</td>
</tr>
<tr>
<td>3.1</td>
<td>An example of filament curve.</td>
</tr>
<tr>
<td>3.2</td>
<td>The function $f$ and its filament.</td>
</tr>
<tr>
<td>3.3</td>
<td>Filaments of a nonparametric regression function (effects of the smoothing parameter $J_1$ and $J_2$). The orange circle is the truth, the blue curve is the estimated filament induced by the posterior mean. Top left: $J_1 = J_2 = 7$; top right: $J_1 = J_2 = 8$; bottom left: $J_1 = J_2 = 9$; bottom right: $J_1 = J_2 = 15$.</td>
</tr>
<tr>
<td>3.4</td>
<td>Filaments of a nonparametric regression function (uncertainty Quantification). Left: $J_1 = J_2 = 7$. Right: $J_1 = J_2 = 9$.</td>
</tr>
<tr>
<td>3.5</td>
<td>Left: Earthquake data points (in red). Right: 100 filaments in the posterior constructed with high frequentist coverage. The thick blue curves are the filaments induced by the posterior mean.</td>
</tr>
<tr>
<td>3.6</td>
<td>The magnitude surface and filaments induced by the posterior mean.</td>
</tr>
<tr>
<td>4.1</td>
<td>The left figure gives the true level curve and the red and blue dots indicate the directions of the elements of $y$. The right figure gives the initial curve and the red and blue dots indicate the initial directions of the elements of $\mu$.</td>
</tr>
<tr>
<td>4.2</td>
<td>Direct Bayes estimate of level curves ($\epsilon_0 = 0.5$). Going from the left to the right and from the top to the bottom, these figures show the status in the 100-th, 200-th, 300-th, 400-th, 1000-th iteration and the posterior mean with one fifth burn-in.</td>
</tr>
</tbody>
</table>
Figure 4.3  Direct Bayes estimate of level curves ($\epsilon_0 = 1$). Going from the left to the right and from the top to the bottom, these figures show the status in the 100-th, 200-th, 300-th, 400-th, 1000-th iteration and the posterior mean with one fifth burn-in. ................................................. 98

Figure 4.4  Direct Bayes estimate of level curves ($\epsilon_0 = 5$). Going from the left to the right and from the top to the bottom, these figures show the status in the 100-th, 200-th, 300-th, 400-th, 1000-th iteration and the posterior mean with one fifth burn-in. ................................................. 99
Chapter 1

Introduction

1.1 Bayesian nonparametrics

Nonparametric and semiparametric models are powerful tools in modern statistics. These models assume that some or all parameters involved are of infinite dimensions. In contrast, parametric models require parameters belong to certain subset of Euclidean space (hence a finite dimensional space). It is well known that parametric models are prone to misspecification in many applications. Nonparametric and semiparametric models offer protection against misspecification of all or some components of the models thus allowing for much more flexibility. However, this flexibility may lead to lower than root-n convergence rates for fully nonparametric components in nonparametric or semiparametric problems. The study of convergence rates for various nonparametric models in statistics is a very important topic. In the Bayesian paradigm, we study the posterior contraction rates for certain geometric aspects of some smooth functions and also study the uncertainty quantification for these objects. We shall first present a literature review and discuss some of the related concepts and issues in Chapter 1.

Bayesian nonparametrics have been attracting wide interests over the past forty years since the publication of Ferguson (1973). This landmark paper proposed the Dirichlet process, a prior on spaces of probability measures. After that, many priors have been proposed and studied, including many kinds of variants of Dirichlet process, Gaussian process, Pólya tree process, independent increment process and random series prior. Like the convergence of estimators in the frequentist sense, posterior contraction is an important concept for Bayesian approach. It is essentially about the concentration of the posterior mass around the truth as sample size increases (see Section 1.2 for more discussion). The general theory of posterior contraction rates in Bayesian nonparametric models was developed in Ghosal, Ghosh and Van Der Vaart (2000) and Ghosal and Van Der Vaart (2007), respectively for i.i.d (independent and identically distributed) observations and general observations. For detailed account of the history and

The theory of assessing the accuracy of an estimate by confidence statement is of fundamental importance in statistics. The Bayesian approach has a unique advantage in that the whole inference can be carried out using the posterior distribution alone. One can construct a credible set which is a region that the parameter will fall onto with certain posterior probability. A perhaps more interesting question one may ask is whether credible sets can have the right coverage in the frequentist sense. There have been many studies on this important subject recently. The literature in this direction includes the following papers. Szabó, van der Vaart and van Zanten (2015) studied adaptive credible regions in Gaussian white noise model. Yoo and Ghosal (2016) addressed similar issues in the multivariate nonparametric regression setting but known smoothness condition. Knapik, van der Vaart and van Zanten (2011) studied the frequentist coverage of credible sets in nonparametric inverse problems. Belitser (2017) studied credible sets in mildly ill-posed inverse signal-in-white-noise model. Belitser and Nurushev (2015) studied uncertainty quantification for the unknown, possibly sparse, signal in general signal with noise models. van der Pas, Szabó and van der Vaart (2017) studied credible sets using the horseshoe prior in the sparse multivariate normal means model in an adaptive setting. Ray (2017) studied Bernstein–von Mises theorems for adaptive nonparametric Bayesian procedures in the Gaussian white noise model. Yoo and Ghosal (2017) studied Bayesian mode and maximum estimation and provided credible sets with good coverage. Belitser and Ghosal (2018) studied uncertainty quantification for high dimensional linear regression models and their results are also extended to high dimensional additive nonparametric regression models. Using credible regions with sufficiently frequentist coverage, one can obtain confidence regions for the truth in the frequentist sense relatively easily from the posterior distribution. This is especially appealing when the object to be studied is complicated. For instance, the level set or the filament of functions, or other geometric aspects of functions. See Section 1.3 for more discussion on credible sets.

Even though there are many different priors one can use for various problems, we shall mainly study the approach using finite random series prior. In general, given a smooth function
\[ f : \mathbb{R}^d \rightarrow \mathbb{R} \text{ in some regular space (for instance, } L_2), \ f \text{ can be expanded as} \]

\[ f = \sum_{j=1}^{\infty} f_j \psi_j, \]

for some suitable basis functions \( \psi = \{ \psi_1, \psi_2, \ldots \} \), where \( f_j \) may be considered as some generalized Fourier coefficient. This expression has infinitely many terms, while in the model one can express \( f \) as \( f = \sum_{j=1}^{J} \theta_j \psi_{J,j} \), \( \psi_j = \{ \psi_{J,1}, \psi_{J,2}, \ldots, \psi_{J,J} \} \) for some \( J \) going to infinity as sample size increases. For a multivariate problem \( (d > 1) \), the basis \( \psi_j \) is often constructed as tensor-product of bases along each dimension. It should be understood for the multivariate case, that the total number of terms \( J \) in the expansion should then be viewed as the product of the number of terms in each dimension. With this series representation, the estimation problem can be viewed as estimating \( \theta = (\theta_1, \ldots, \theta_J) \), or equivalently, estimating \( f \) from the space \( \mathcal{F}(J) = \{ \sum_{j=1}^{J} \alpha_j f_j : \alpha_j \in \mathbb{R} \} \). This truncated version of estimator is useful when it has good approximation property. That is, for some sufficiently large \( J \),

\[ \inf_{\mathcal{F}(J)} \| f - f^\infty \|_\infty \lesssim J^{-\alpha/d}. \]  

(1.1)

The coefficient \( \alpha \) is normally some smoothness coefficient for the function \( f \). Many choices of bases are available that have this approximation property. We shall discuss with the trigonometric basis, B-spline basis and wavelet basis. For more treatment on approximation of functions, see DeVore and Lorentz (1993); Lorentz (2005); Powell (1981).

When a density function \( p \) is of interest, it can be modeled as

\[ p := \frac{\Psi(f)}{\int \Psi(f) d\mu}, \]

for some positive link function \( \Psi \) and \( f = \sum_{j=1}^{J} \theta_j \psi_j \). A common choice for \( \Psi \) is the exponential function. Modeling \( p = \sum_{j=1}^{J} \theta_j \psi_j \) is also possible, provided that \( \int \sum_{j=1}^{J} \theta_j \psi_j = 1 \) for any \( J \) and \( \psi_j \geq 0 \).

### 1.2 Posterior contraction

In this section, we shall briefly discuss the theory of posterior contraction. Suppose one has observations \( X = (X_i)_{i=1}^{n} \) whose joint distribution is given by \( P_f^{(n)} \) with the corresponding density function \( p_f^{(n)} \). When the observations are independent (not necessarily identically distributed), we write \( P_f^{(n)} = \bigotimes_i^n P_{f_i} \) and \( p_f^{(n)} = \prod_i^n p_{f_i} \), where \( f_i \) denotes \( f(X_i) \). Here \( f \) belongs to some parameter space \( \mathcal{F} \) which is endowed with some semimetric \( d_n \). One can put some prior \( \Pi_n \) on
Let \( \Pi_n(\cdot|X) \) denote the corresponding posterior defined by

\[
\Pi(B|X) = \frac{\int_B p_f^{(n)}(X)d\Pi(f)}{\int p_f^{(n)}(X)d\Pi(f)}.
\]  

(1.2)

The posterior distribution \( \Pi_n(\cdot|X) \) is said to be consistent at the truth \( f^* \in \mathcal{F} \) if for any \( \epsilon > 0 \),

\[
\Pi_n(\theta : d_n(f, f^*) > \epsilon|X) \to 0,
\]

(1.3)
in \( P_f^{(n)} \)-probability as \( n \to \infty \).

We say that a sequence \( \epsilon_n \) is a posterior contraction rate at the parameter \( f^* \) with respect to \( d_n \) if for any \( K_n \to \infty \),

\[
\Pi_n(\theta : d_n(f, f^*) > K_n\epsilon_n|X) \to 0,
\]

(1.4)
in \( P_{f^*}^{(n)} \)-probability as \( n \to \infty \). The smallest \( \epsilon_n \) that satisfies this definition is of particular interest, as it gives us the fastest possible rate. In many classical examples, one would hope that this rate is equal to or close to the “optimal” frequentist rate.

Posterior contraction essentially guarantees that the information in the data can override the prior opinions as the sample size increases. From a subjective Bayesian’s point of view, consistency can also be equivalently interpreted as the merging of predictive distributions of future observations—another desired property. Posterior contraction can also provide point estimators (for many cases, posterior mean) that converge at the same rate.

As mentioned previously, the general theory of posterior contraction rates in Bayesian nonparametric models was developed in Ghosal et al. (2000) and Ghosal and Van Der Vaart (2007), respectively for i.i.d (independent and identically distributed) observations and general observations. In this general theory, the existence of tests of exponentially decaying errors plays a fundamental role. On the other hand, Shen and Wasserman (2001) also studied posterior contraction rates for i.i.d. observations but using different and stronger conditions. These works focused on consistency and contraction rates in the Hellinger metric.

Let the Hellinger distance between \( p \) and \( q \) be defined as \( (\int (\sqrt{p} - \sqrt{q})^2d\mu)^{1/2} \). The Kullback-Leibler divergence (KL divergence) is defined by \( K(p, q) = \int p \log(p/q)d\mu \). The master theorem in general can give rates for the Hellinger distance on the densities induced by \( f \). We may interpret \( d_n(f_1, f_2) \) as \( d_H(p_{f_1}, p_{f_2}) \) when the data are i.i.d, or as \( d_{H,n} := \sqrt{\frac{\sum_{i=1}^{n} d_H^2(p_{f_{1,i}}, p_{f_{2,i}})}{n}} \), the root-average squared Hellinger distance when the data are independently (but not identically) distributed.

The book Ghosal and van der Vaart (2017) has in-depth discussion on general theory. We shall present some of these results.

**Theorem 1.2.1** Suppose there exists a sequence of spaces \( \mathcal{F}_n \subset \mathcal{F} \) and constant \( C > 0 \) and a
test \( \phi_n \) such that
\[
P_{f^*}^{(n)} \phi_n \to 0, \quad \sup_{f \in \mathcal{F}_n : d_n(f, f^*) > \epsilon_n} P_f^{(n)} (1 - \phi_n) \leq \exp(-Cn\epsilon_n^2), \tag{1.5}
\]
and that for some constant \( L > 0 \) and some sufficiently large constant \( b > 0 \),
\[
\Pi_n(f \in \mathcal{F}_n : K(p_f^{(n)}, p_{f^*}^{(n)}) \leq n\epsilon_n^2) \geq \exp(-Ln\epsilon_n^2),
\]
\[
\Pi_n(\mathcal{F} \setminus \mathcal{F}_n) \leq \exp(-bne_n^2),
\]
then \( \Pi(d_n(f, f^*) > K_n \epsilon_n|Y) \to 0 \), for every \( K_n \to \infty \).

Condition (1.5) requires the existence of a test whose type I error probability tends to zero and type II error probability is exponentially decaying on the space \( \mathcal{F}_n \). We shall give a sufficient condition for (1.5). To this end, we need the concept of covering number. Given a metric or semimetric space \((T, d)\), for any \( \epsilon > 0 \), its covering number \( N(T, d, \epsilon) \) is defined as the smallest number of closed \( d \)-balls of radius \( \epsilon \) needed to cover \( T \). A sufficient condition for (1.5) is now given in the following:

Suppose for two arbitrary semimetric \( d_n, e_n \) on \( \mathcal{F} \), there exists some constants \( C_1 > 0, C_2 > 0, \xi > 0 \) so that for every \( \epsilon > 0 \) and every \( f_1 \in \mathcal{F}_n \) with \( d_n(f_1, f^*) > \epsilon \), one can find a test \( \phi_n \) with the exponential error probabilities
\[
P_{f^*}^{(n)} \phi_n \leq \exp(-C_1n\epsilon^2), \quad \sup_{f \in \mathcal{F}_n : e_n(f, f_1) < \xi \epsilon} P_f^{(n)} (1 - \phi_n) \leq \exp(-C_2n\epsilon^2), \tag{1.6}
\]
and that
\[
\log N(\epsilon_n, \mathcal{F}_n, e_n) \lesssim n\epsilon_n^2. \tag{1.7}
\]

The condition expressed in (1.6) is the basic testing condition. By the Le Cam–Birgé testing theory, it holds for the Hellinger distance (when the observations are i.i.d) and the root-average squared Hellinger distance (when the observations are independently distributed). The entropy condition (1.7) requires that one can use a collection of ball-like sets with respect to \( e_n \) to cover \( \mathcal{F}_n \) and this collection is not too large.

Though above result allows fairly general metric \( d_n \), it is not automatically extended to the \( L^p \)-norms \( 1 \leq p \leq \infty \), as the Condition (1.6) does not readily hold for these metrics. The contractions rates in the \( L^p \)-norms \( 1 \leq p \leq \infty \) were studied in Giné and Nickl (2011). The minimax contraction rates were obtained for density estimates for \( 1 \leq p \leq 2 \) and for the Gaussian white noise models using conjugate priors for all \( 1 \leq p \leq \infty \). More recently, Castillo (2014) proposed a method to study the sup-norm contraction rates and this method
leads to the minimax contraction rates in the density estimation with log-density prior and the Gaussian white noise models with nonconjugate priors. Yoo and Ghosal (2016) established the optimal $L^2$- and $L^\infty$- rates for the regression function and its mixed partial derivative but with conjugacy. Shen and Ghosal (2017) studied both $L^2$- and $L^\infty$- rates for density and its derivatives. They showed that the minimax rate for the $L^2$- distance is obtained.

Theorem 1.2.1 is a very general one as it does not require any specific type of prior on $f$. For finite random series priors, more explicit results can be presented. We summarize some of those relevant results in Appendix B and shall use them intensively in the proofs of some main results in Chapter 2.

1.3 Credible regions and coverage

Constructing confidence bands in nonparametric problems is a very important topic in itself. Some recent papers include Claeskens and Van Keilegom (2003), Giné and Nickl (2010) and Chernozhukov, Chetverikov and Kato (2013, 2014a). As discussed earlier, an interesting and important question from Bayesian perspective is to ask whether the credible region can have good frequentist coverage. We start with a definition. Consider the parameter $\theta$ in a finite dimensional model ($\Theta \subset \mathbb{R}^p$). Formally, for some positive value $\gamma \in (0, 1)$, a subset $C(X)$ of $\Theta$ is called a $1-\gamma$ credible region if

$$\Pi_n(\theta : \theta \in C(X)|X) = 1 - \gamma.$$ 

Let $X_i \overset{\text{i.i.d.}}{\sim} P_{\theta_0}$. According to the celebrated Bernstein-von-Mises Theorem, under some regularity conditions,

$$\sqrt{n}(\theta - \tilde{\theta})|X \overset{d}{\sim} N(0, I(\theta_0)^{-1}),$$

where $\tilde{\theta}$ is the posterior mean and $I(\theta_0)$ is the Fisher information. Since $\tilde{\theta}$ is asymptotically equivalent to a frequentist efficient estimator (say, the MLE), this guarantees that the credible region $C(X) = \{\theta : \|\theta - \tilde{\theta}\| \leq R_\gamma\}$, where $R_\gamma$ is the $1-\gamma$ quantile of the posterior distribution of $\theta - \tilde{\theta}$, has the right exact asymptotic frequentist coverage. That is,

$$P_{\theta_0}^{(n)}(\|\theta_0 - \tilde{\theta}\| \leq R_\gamma) = 1 - \gamma.$$

Unfortunately, in the infinite dimensional model (when $\Theta = \mathcal{F}$ endowed with distance function $d_n$ on functions), this result may no longer hold (Cox; 1993; Freedman; 1999). However, a weak functional of Bernstein-von-Mises theorem was put forward in Castillo and Nickl (2014); Giné and Nickl (2010) and the credible region constructed in some weaker norm with exact frequentist coverage can be obtained.
A different approach is that, if the requirement of “exact” coverage is relaxed, it is possible to find an “inflated” credible set $C(X) = \left\{ \theta : d_n(\theta, \tilde{\theta}) \leq \rho_n R_\gamma \right\}$, for some $\rho_n \to \infty$ (or possibly a large fixed constant) such that,

$$\liminf_n \inf_{\theta_0 \in \Theta} P_{\theta_0}^{(n)} \left( d_n(\theta_0, \tilde{\theta}) \leq \rho_n R_\gamma \right) \to 1,$$

as $n \to \infty$,

and the posterior quantile $R_\gamma$ has the same order as the minimax rate of estimating the parameter $\theta_0$—a rate usually can be achieved by the posterior mean $\tilde{\theta}$. A recent paper by Szabó et al. (2015) studied adaptive nonparametric regions in the above sense with the full parameter space $\Theta$ replaced by some dense subset in the Gaussian white noise model. Yoo and Ghosal (2016) addressed similar issues in the multivariate nonparametric regression setting with known smoothness condition. Some other papers are mentioned in the Section 1.1. A credible set with sufficient but not exact coverage can also be obtained through undersmoothing without using any inflating factor, but this leads to suboptimal estimation.

To briefly compare the frequentist and Bayesian approaches, we focus on the uniform uncertainty quantification for the function itself. Let $f^*$ denote the true function. Suppose $R_\gamma$ is the $1 - \gamma$ quantile of $\|\hat{f} - f^*\|_\infty$, then

$$P(\|\hat{f} - f^*\|_\infty \leq R_\gamma) = 1 - \gamma.$$

Therefore, a valid $1 - \gamma$ uniform confidence region for $f^*$ is $\{g : \|g - \hat{f}\|_\infty \leq R_\gamma\}$. Note that the uniformity is in the sense that $R_\gamma$ does not depend on any given point $x$ as in contrast to the radius in pointwise confidence interval. The difficulty then is the estimation of the quantiles $R_\gamma$. For some cases, when asymptotic distribution of the $\|\hat{f} - f^*\|_\infty$ can be derived, the quantile can be obtained accordingly. As an alternative, to estimate the quantile $R_\gamma$, bootstrap can be used and some theoretical foundations are established in Chernozhukov et al. (2013, 2014a) and Chernozhukov, Chetverikov and Kato (2014b). However, it is worth to point out that in the bootstrap approach, it is the distribution $\|\hat{f} - E\hat{f}\|_\infty$ that is approximated by the bootstrap distribution. Strictly speaking, it is the quantile of $\|\hat{f} - E\hat{f}\|_\infty$ that can be obtained and confidence region obtained would be for $E\hat{f}$ instead of $f^*$.

We turn to Bayesian approach. Let $\tilde{f}$ be the posterior mean of $f$. One can construct the credible region as $\{g : \|g - \tilde{f}\|_\infty \leq \rho_n R_\gamma\}$ for some inflation constant $\rho_n$, where here $R_\gamma$ is the $1 - \gamma$ quantile of posterior distribution of $\|f - \tilde{f}\|_\infty$. It can be shown that in some cases this credible set can have sufficient frequentist coverage (Yoo and Ghosal; 2016). Additional property may be obtained also. For instance, the radius of the region can have the same rate as the optimal rate of estimating $f^*$ in the $L^\infty$-norm. To ensure an inflation factor of the correct order may be used to achieve coverage, the lower bound of the quantiles would need to be
estimated; see for instance the proof of the results in Section 2.3.6. The benefit of Bayesian approach is that the posterior samples are readily available thus finding the right quantile is easier than the bootstrap procedure. Furthermore, the inferential target is $f^*$, rather than the expected value of the estimator.

1.4 Research questions and contributions

The strategy for uncertainty quantification using confidence region or credible region for the function itself is more or less well understood. In the literature there are some other problems where estimation and inference is performed for a set, or more broadly speaking for some geometric object (Molchanov; 2006). In this dissertation, we consider two features, level sets and filaments. A level set of a multivariate function $f$ is defined as the set $\mathcal{L} := \{x : f(x) = c\}$ or $\mathcal{L} := \{x : f(x) \geq c\}$. The filament, to avoid technicalities here, can be viewed as a collection of local maximizers of $f$ that behaves like a lower dimensional manifold. The study of estimation of level sets dates back to 1990’s while that of filaments is more recent. Both of these problems fall into a broad category of data analytic methods known as topological data analysis, which is used to find structure in data (Wasserman; 2016). Some other examples arise in econometrics literature where partially identified parameters are the objects of interest (Chernozhukov, Hong and Tamer; 2007; Chernozhukov, Kocatulum and Menzel; 2015). Most papers study estimation of levels and filaments in the frequentists setting while there are few recent papers on inference. To our best knowledge, there is only one published Bayesian paper on level set estimation (Gayraud and Rousseau; 2005) and no study on filament using the Bayesian approach.

The $L^\infty$-rates of the estimates of functions $f$ are essential for understanding the statistical properties of the procedures that estimate level sets and filaments. The idea in the last section on credible region of the function may be used to make inference about level sets and filaments. Posterior contraction of the functions in sup-distance can induce credible regions on some features of the functions. This becomes apparent when the Hausdorff distance of level sets or the filaments can be bounded by the sup-distances on the functions or the derivatives of the functions.

In this dissertation, we address the following issues. First, we derive new $L^\infty$-posterior contraction rates for the multivariate functions in several different nonparametric settings. For the multivariate Gaussian white noise model, the $L^\infty$-contraction rates for the functions using trigonometric series and wavelet series priors are separately derived. Since the collection of level sets for the square root of a nonnegative function coincides with that of the original function, the results based on the square roots of the functions can lead to useful results for level sets. For binary, and Poisson regression and density estimation, the $L^\infty$-contraction rates for the square roots of the mean functions and the square root of the density function respectively, along with
their derivatives are obtained. For these three examples, we use random B-spline series prior and the properties of B-splines functions to obtain the $L^\infty$-rates from the $L^2$-rates through a relaxation argument. These rates may not be optimal but they are new in the literature and should be useful as some intermediate results. Second, as by-products, all these $L^\infty$-rates yield the contraction rates for level sets in the Hausdorff distance and the Lebesgue measure of the symmetric difference. Third, we study frequentist coverage properties of suitable credible regions for level sets in both Gaussian white noise model using trigonometric basis prior and multivariate nonparametric normal regression using B-splines basis prior. A simulation study shows that the proposed credible regions have sufficient frequentist coverage.

Our study of filaments contribute to the current literature in two ways. First, we provide a Bayesian approach to filament estimation in the multivariate nonparametric normal regression context and study the posterior contraction rates using a finite random series of B-splines basis. Compared with the kernel-estimation method, this has theoretical advantage as the bias can be better controlled when the function is smoother, which allows obtaining better rates. Assuming that $f : \mathbb{R}^2 \mapsto \mathbb{R}$ belongs to an isotropic Hölder class of order $\alpha \geq 4$, with the optimal choice of smoothing parameters, the posterior contraction rates for the filament points on some appropriately defined integral curves and for the Hausdorff distance of the filament are both $(n / \log n)^{(2 - \alpha) / (2(1 + \alpha))}$. Second, we provide a way to construct a credible set with sufficient frequentist coverage for the filaments. We study the performance of our proposed method by simulations and application to an earthquake dataset.

1.5 Notations and preliminaries

In this section, we shall describe notations and briefly introduce the preliminaries for three common function bases.

1.5.1 Notations

Let $\mathbb{N} = \{1, 2, \ldots\}$, $\mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$. Given two real sequences $a_n$ and $b_n$, $a_n = O(b_n)$ or $a_n \lesssim b_n$ means that $a_n/b_n$ is bounded, while $a_n = o(b_n)$ or $a_n \ll b_n$ means that $a_n/b_n \to 0$. Also $a_n \asymp b_n$ means that both $a_n = O(b_n)$ and $b_n = O(a_n)$. For random element $Z_n$, $Z_n = O_p(a_n)$ means that $P(\|Z_n\| \leq Ca_n) \to 1$ for some constant $C > 0$.

For a vector $x \in \mathbb{R}^d$, we define $\|x\|_p = (\sum_{k=1}^d |x_k|^p)^{1/p}$ for $0 \leq p \leq \infty$, $\|x\|_\infty = \max_{1 \leq k \leq d} |x_k|$ and write $\|x\|$ for $\|x\|_2$. For an $m \times m$ matrix $A$, let $\|A\|_{(r,s)} = \sup\{\|Ax\|_s : \|x\|_r \leq 1\}$. In particular, $\|A\|_{(2,2)} = (\lambda_{\max}(A^T A))^{1/2}$, where $\lambda_{\max}$ denotes the largest eigenvalue; $\|A\|_{(\infty, \infty)} = \max_{1 \leq i \leq m} \sum_{j=1}^m |a_{ij}|$. Let $\|A\|_F = \sqrt{\text{tr}(A^T A)}$ stands for the Frobenius norm of matrix $A$. We also denote the identity matrix by $I$ whose dimension is given in the context.
For two probability densities \( f \) and \( g \), the \( L^p \)-distance \((1 \leq p < \infty)\) is given by \((\int |f - g|^p d\mu)^{1/p}\) assuming the \(\sigma\)-finite measure is given by \(\mu\). The Hellinger distance between \( f \) and \( g \) is defined as \((\int (\sqrt{f} - \sqrt{g})^2 d\mu)^{1/2}\). The Kullback-Leibler divergence (KL divergence) is given by \(K(f, g) = \int f \log(f/g) d\mu\).

Consider \( f : U \mapsto \mathbb{R} \) defined on some bounded set \( U \subset \mathbb{R}^d \). For \(1 \leq p < \infty\), let \(\|f\|_p \) be the \(L^p\)-norm \((\int |f|^p d\nu)^{1/p}\) with respect to some measure \(\nu\). For some probability measure \(\mathcal{G}\), let \(\|f\|_{p,\mathcal{G}} = (\int |f|^p d\mathcal{G})^{1/p}\). The supremum norm is denoted by \(\|f\|_\infty = \sup_{x \in U} |f(x)|\). For \( g : U \mapsto \mathbb{R}^1 \) on some bounded set \( U \subset \mathbb{R}^d \), let \(\nabla g\) be the gradient of \( g \), which is a \(d \times 1\) vector. For a \(d\)-dimensional multindex \( r = (r_1, \ldots, r_d) \in \mathbb{N}_0^d \), let \(D^r\) be the partial derivative operator \(\partial^{|r|}/\partial x_1^{r_1} \cdots \partial x_d^{r_d}\) where \(|r| = \sum_{k=1}^d r_k\).

For some integer \(q \geq 1\), let \(C^q([0,1]^d)\) denote the space of \(q\)-times continuously differentiable functions on \([0,1]^d\). The Hölder Space \(H^\alpha([0,1]^d)\) of order \(\alpha > 0\) consists of functions \( f : [0,1]^d \mapsto \mathbb{R} \) such that \(\|f\|_{\alpha,\infty} < \infty\), where \(\| \cdot \|_{\alpha,\infty}\) is the Hölder norm

\[
\|f\|_{\alpha,\infty} = \max_{r:|r| \leq [\alpha]} \sup_x |D^r f(x)| + \max_{r:|r| = [\alpha]} \sup_{x,y:x \neq y} \frac{|D^r f(x) - D^r f(y)|}{\|x - y\|^{\alpha - |r|}},
\]

where sup is taken over the support of \(f\) and \([\alpha]\) is the largest integer strictly smaller than \(\alpha\).

We shall also need the Hausdorff distance for our study of level sets and filaments. Given two sets \(A\) and \(B\) in Euclidean Space, let \(d(A|B) := \sup_{x \in A} \inf_{y \in B} \|x - y\|\). If \(A = \{x\}\), write \(d(x|B) = d(\{x\}|B) := \inf_{y \in B} \|x - y\|\). The Hausdorff distance between \(A\) and \(B\) is defined as

\[
\text{Haus}(A, B) = \max\{d(A|B), d(B|A)\}. \tag{1.8}
\]

### 1.5.2 Trigonometric basis

Consider the trigonometric basis on \(\mathbb{R}\): \(\phi_1 = 1\), \(\phi_{2j} = \sqrt{2} \cos(2\pi jx)\), \(\phi_{2j+1} = \sqrt{2} \sin(2\pi jx)\), \(j = 1, 2, \ldots\). Functions of period 1 on \(\mathbb{R}\) or the periodic extension of function \(f\) defined on \([0,1]\) can be expressed as

\[
f = \alpha_0 + \sum_{j=1}^\infty [\alpha_j \cos(2\pi jx) + \beta_j \sin(2\pi jx)].
\]

Consider a function \(f : [0,1]^d \mapsto \mathbb{R}\). Let \(j = (j_1, \ldots, j_d) \in \{0,1,2,\ldots\}^d\) and \(\mathcal{K}(j) := \{k \in \{0,1\}^d : k_i = 0 \text{ when } j_i = 0, i = 1, \ldots, d\}\). We denote the tensor product trigonometric basis functions by \(\{\phi_{j,k} : (j,k) \in \{0,1,2,\ldots\}^d \times \mathcal{K}(j)\}\), where

\[
\phi_{j,k}(x) = \prod_{i=1}^d \sqrt{2}[(1-k_i) \cos(2\pi j_i x_i) + k_i \sin(2\pi j_i x_i)].
\]
The function $f$ can be written as

$$f(x) = \sum_{j_1} \cdots \sum_{j_d} \sum_{k \in K(j)} \phi_{jk} \theta_{jk}.$$

For brevity, we may simplify the indices of summation and just write $f = \sum_j \sum_k \phi_{jk} \theta_{jk}$.

Consider the space of the form

$$\mathcal{F}_n = \mathcal{F}_{n,t} := \left\{ \sum_{j_1=0}^{J_n} \cdots \sum_{j_d=0}^{J_n} \sum_{k \in K(j)} \alpha_{jk} \phi_{jk} : \alpha_{jk} \in \mathbb{R} \right\}.$$

If $f \in H^\alpha([0,1]^d)$, for $J_n$ sufficiently large, one has (Schultz; 1969):

$$\inf_{f^\infty \in \mathcal{F}_{n,t}} \|f - f^\infty\|_\infty \lesssim J_n^{-\alpha}.$$

Notice that the order is $(\prod_{l=1}^d J_n)^{-\alpha/d}$ as described in the expression (1.1).

### 1.5.3 B-Splines Basis

Let a sequence of knots $\{t_i : 0 = t_0 < \cdots < t_{N+1} = 1\}$ be a partition of the interval $[0,1]$ into $N + 1$ subintervals: $I_i = [t_i, t_{i+1})$ for $i = 0, \ldots, N$ and $I_N = [t_{N-1}, t_N]$. The space of (polynomials) splines of order $q$ defined on $[0,1]$ is a linear space such that

(i) each element $f$ is a polynomial of order at most $q$ on each interval $I_i$;

(ii) each element $f$ belongs to $C^{q-2}[0,1]$.

Intuitively speaking, a spline function is a piecewise polynomial which has global smoothness over the whole interval. B-splines are one particular basis for the space of splines. Let $\delta_i = t_{i+1} - t_i$ for $i = 0, \ldots, N$ and $\Delta = \max_{0 \leq i \leq N} \delta_i$. We assume a quasi-uniform knot sequence, i.e., $\Delta / \min_{0 \leq i \leq N} \delta_i \leq C$ for some constant $C > 0$. Note that the equidistant knot sequence, i.e., $\delta_i$ all being all equal, is a special case. To construct this basis, extended knots are required and we assume they are all positioned at the end points of the interval. We denote the B-spline functions by $\{B_{1,q}, \ldots, B_{N+q,q}\}$. These functions can be defined using divided difference or by the following convenient recursive formula:

$$B_{i,q}(x) = \frac{x - t_{i-q}}{t_{i-1} - t_{i-q}} B_{i-1,q-1}(x) + \frac{t_i - x}{t_i - t_{i+1-q}} B_{i,q-1}(x), \quad i = 1, \ldots, q + N.$$

When $q = 1$, define $B_{i,q}(x) = \mathbb{1}_{[t_{i-1}, t_i)}(x)$. 

11
By construction, we immediately have $B_i,q(x) > 0$ on $(t_{i−q}, t_i)$ and $\sum_{i=1}^{N+q} B_i,q = 1$. These functions constitute a linear space of dimension $J = q + N$. In particular, for any given $x \in [t_i, t_{i+1})$, only at most $q$ adjacent B-spline functions are nonzero: $(B_{i+1,q}, \ldots, B_{i+q,q})$. Figure 1.1 gives some examples of B-spline functions.

Consider a linear combination of the B-spline functions given by $f(x) = \sum_{j=1}^{J} \theta_j B_j,q(x)$. For some integer $q > s \geq 1$, the derivative of $D^s f(x)$ is given by $D^s f(x) = \sum_{j=1}^{J-s} \theta_j (s-1)^j B_{j,s}^{-q}(x)$, where $\theta_j = \frac{(q-s)\theta_{j+1}^{(s-1)} - \theta_{j+1}^{(s-1)}}{t_j - t_{j-q+s}}$.

Let $b_{J,q}$ be the column vector of $\{B_{i,q}: 1 \leq i \leq J\}$ and $\theta = (\theta_1, \ldots, \theta_J)^T$. It can be shown that $D^s f(x) = b_{J-s,q-s}^T W_s \theta$ for some $(J - s) \times s$ matrix $W_s$ and each nonzero entry of $W_s$ is of order $J^s$. Here $b_{J-s,q-s}$ consists of $(q - s)$-order B-spline functions defined on the same interval using the same sequence of interior knots.

Figure 1.1: B-spline functions. Left plot: $q = 2, N = 3$. Right plot: $q = 4, N = 3$.

To model a multivariate function $f \in \mathcal{H}^\alpha([0,1]^d)$, we can make use of the tensor product of B-splines basis. Define $b_{J_1, \ldots, J_d, q_1, \ldots, q_d}(x) = (B_{j_1, \ldots, j_d}(x) = \prod_{k=1}^{d} B_{j_k,q_k}(x_k) : 1 \leq j_k \leq J_k)^T$ to be a column vector of tensor product of B-splines functions, with possibly different orders $q_k$ and knot sequences in different directions, i.e., $0 = t_{k,0} < t_{k,1} < \cdots < t_{k,N_k} < t_{k,N_k+1} = 1$ for $k = 1, \ldots, d$; here $N_k$ denotes the number of interior points and $J_k = q_k + N_k$ denotes the number of basis functions on the $j$th coordinate. The elements of this vector are assumed to be in the dictionary order according to their indices. Whenever $q_1, \ldots, q_d$ are considered fixed, we shall suppress the subscripts $q_1, \ldots, q_d$ in our notations of B-spline functions. For each
\( k = 1, \ldots, d \), define \( \delta_{k, l} = t_{k, l+1} - t_{k, l} \) for \( l = 0, \ldots, N_k \). Let \( \Delta_k = \max_{0 \leq l \leq N_k} \delta_{k, l} \). We assume that \( \Delta_k / \min_{0 \leq l \leq N_k} \delta_{k, l} \leq C \) for some \( C > 0 \). This assumption is clearly satisfied for the uniform partition. Therefore we write \( f(x) = b^T_{1, \ldots, d}(x) \theta \), or simply \( f(x) = b^T(x) \theta \). As in the univariate case, we have \( D^r f(x) = b_{J - r, q - r}(x)^T W_r \theta \) for some \( \prod_{k=1}^d (J_k - r_k) \times \prod_{k=1}^d J_k \) matrix \( W_r \).

Several properties of B-splines are important for our purpose and they are listed as follows:

- \( 0 \leq B_{j_1, \ldots, j_d} \leq 1 \) for all \( 1 \leq j_k \leq J_k, k = 1, \ldots, d \) and \( x \in [0, 1]^d \);
- \( \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} B_{j_1, \ldots, j_d}(x) = 1 \) for all \( x \in [0, 1]^d \);
- \( B_{j_1, \ldots, j_d} \) is supported on \( \prod_{k=1}^d [t_{k, j_k - q_k}, t_{k, t_k}] \).

For the moment, we fix \( q_k > \alpha \) and let \( N_k \) (hence \( J_k \)) increase as sample size increases, for \( k = 1, \ldots, d \). Yoo and Ghosal (2016) showed that each nonzero entry of \( W_r \) is uniformly \( O(\prod_{k=1}^d J_k^{r_k}) \). Therefore, \( \|W_r\|_{\infty, \infty} = O(\prod_{k=1}^d J_k^{r_k}) \) and \( \|W_r^T W_r\|_{(2,2)} = O(\prod_{k=1}^d J_k^{r_k}) \). We also have the following approximation property. Consider the space of the form

\[
F_n = F_{n, b} := \left\{ \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \alpha_{j_1, \ldots, j_d} B_{j_1, q_1}(x_1) \cdots B_{j_d, q_d}(x_d) : \alpha_{j_1, \ldots, j_d} \in \mathbb{R} \right\}.
\]

If \( f \in H^\alpha([0,1]^d) \) and \( q_k > \alpha \) for \( k = 1, \ldots, d \), for \( J_n \) sufficiently large, one has

\[
\inf_{f^\infty \in F_{n, b}} \|f - f^\infty\|_{\infty} \lesssim J_n^{-\alpha}.
\]

For more treatment of B-splines functions, readers may refer to De Boor (1978) and Schumaker (2007).

### 1.5.4 Wavelet basis

There is a rich literature on both theory and applications of wavelets. Our discussion here is mainly based on Härdle, Kerkyacharian, Picard and Tsybakov (2012) and Giné and Nickl (2015).

**Definition 1.5.1** A system of real-valued functions \( \{\phi_k : k \in \mathbb{Z}\} \) of \( L^2(\mathbb{R}) \) is an orthonormal system if \( \langle \phi_k, \phi_{k'} \rangle := \int \phi_k(x)\phi_{k'}(x)dx = 1 \) if \( k = k' \), and \( \langle \phi_k, \phi_{k'} \rangle = 0 \) otherwise. It is an orthonormal basis for a subspace \( V \) of \( L^2(\mathbb{R}) \) if for any function \( f \in V \), \( f(x) = \sum k c_k \phi_k(x) \) in \( L^2 \), with \( \sum |c_k|^2 < \infty \).

Let \( \phi \) be some function from \( L_2(\mathbb{R}) \) such that:

1. \( \{\phi_{0k} : k \in \mathbb{Z}\} = \{\phi(\cdot - k) : k \in \mathbb{Z}\} \) forms an orthonormal basis for a subspace \( V_0 \) of \( L_2(\mathbb{R}) \);
2. the collection of linear spaces $V_l := \{ h(x) = f(2^l x) : f \in V_0 \}, l \in \mathbb{Z}$ satisfies that $V_l \subset V_{l+1}, l \in \mathbb{Z}$ and $\cup_{l \geq 0} V_l$ is dense in $L^2(\mathbb{R})$.

The sequence of spaces $\{V_l : l \in \mathbb{Z}\}$ is called a multiresolution analysis (MRA) of $L^2(\mathbb{R})$. The collection of dilations and translations $\{\phi_{lk} := 2^{l/2}\phi(2^l \cdot -k) : k \in \mathbb{Z}\}$ forms an orthonormal basis for $V_l$ for each $l \in \mathbb{N}$. Such a function $\phi$ is called the father wavelet or scaling function that generates this MRA.

If furthermore, for a given MRA, some function $\psi$ can be found such that the collection $\{\psi_{lk} := 2^{l/2}\psi(2^l \cdot -k) : k \in \mathbb{Z}\}$ is an orthonormal basis for the orthogonal complement $W_l := V_{l+1} \ominus V_l$ in Hilbert space for each $l \in \mathbb{N}$. Then one can express $V_l = V_0 \oplus (\oplus_{l=1}^{l-1} W_l)$ and $L^2(\mathbb{R}) = V_0 \oplus (\oplus_{l=1}^{\infty} W_l)$. Here the expression using $\oplus$ means an orthogonal decomposition in Hilbert space. That is, the collection of the functions

$$\phi_k = \phi(\cdot -k), \quad \psi_{lk} = 2^{l/2}\psi(2^l \cdot -k), \quad l \in \{0\} \cup \mathbb{N}, k \in \mathbb{Z},$$

forms an orthonormal basis of $L^2(\mathbb{R})$. Therefore, for any $f \in L^2(\mathbb{R})$, one can write

$$f = \sum_{k \in \mathbb{Z}} \langle \phi_{0k}, f \rangle \phi_{0k} + \sum_{l=0}^\infty \sum_{k \in \mathbb{Z}} \langle \psi_{lk}, f \rangle \psi_{lk} \text{ in } L^2.$$

This expansion is called the wavelet expansion. The function $\psi$ is called the mother wavelet or simply the wavelet. The index $l$ refers to the resolution level.

The simplest orthogonal wavelet basis is Haar wavelet basis which has scaling function $\phi(x) = 1_{[0,1]}(x)$ and wavelet function $\psi(x) = 1_{[0,1/2]}(x) - 1_{[1/2,1]}(x)$. In fact, Haar wavelet is a special case of a more general family of wavelet bases called Daubechies wavelets. The most important features of this family are that the (orthogonal) wavelets have compact support in time domain and can be arbitrarily smooth. To be more precise, for any $N \in \mathbb{N}$, Daubechies wavelet basis of order $N$ can be constructed so that its scaling function and wavelet functions are at least $\lfloor \lambda(N-1) \rfloor$ continuously differentiable for some $\lambda \geq 0.18$ and wavelet functions have $N$ vanishing moments. Figure 1.2 gives some examples of Daubechies scaling functions and wavelet functions.

Another interesting variant is the boundary-corrected wavelet basis of Cohen, Daubechies and Vial (1993), called the CDV-wavelet basis. The coefficients of wavelet expansions in this basis can be naturally used to describe the smoothness of the function with compact support in Besov spaces. The basis is constructed from ordinary Daubechies basis starting from some
sufficiently large resolution level say $J_0$. The expansion can be written as

$$f = \sum_{k=0}^{2^{J_0}-1} \langle \phi_{J_0k}, f \rangle \phi_{J_0k} + \sum_{l=J_0}^{\infty} \sum_{k=0}^{2^l-1} \langle \psi_{lk}, f \rangle \psi_{lk}. $$

Regular wavelet bases have rich theoretical properties. It suffices to point out that the above wavelet expansion does not only converge in $L^2$ but also in $L^\infty$, provided that $f$ is uniformly continuous. Wavelet functions have localization property: $\| \sum_k c_k |\psi_{lk}| \|_\infty \lesssim \max |c_k|^{2l/2}$ for a sequence of real number $\{c_k : k \in \mathbb{Z}\}$ such that $\max_k |c_k| < \infty$. When $f$ is $\alpha$-times continuously differentiable, using sufficiently regular wavelet basis, $\sup_{k \in \mathbb{Z}} |\langle f, \psi_{lk} \rangle| \lesssim 2^{-l(\alpha+1/2)}$ for all $l \geq 0$.

Figure 1.2: Examples of some Daubechies scaling functions and wavelet functions.
Suppose that \( \phi, \psi \) are the scaling function and wavelet of a Daubechies wavelet basis of \( L^2(\mathbb{R}) \). That is, the collection of the functions
\[
\phi_k = \phi(\cdot - k), \quad \psi_{lk} = 2^{l/2} \psi(2^l \cdot - k), \quad l \in \{0\} \cup \mathbb{N}, \quad k \in \mathbb{Z},
\]
forms an orthonormal basis of \( L^2(\mathbb{R}) \). We assume that \( \phi, \psi \) are both \( q \)-times (\( q > \alpha \)) continuously differentiable. A wavelet basis on \([0, 1]^d\) (called CDV-wavelet basis) can be constructed from \( \phi, \psi \) starting from some sufficiently large fixed resolution level \( J_0 \). We shall use the same notations for this basis so that \( \{\phi_k, \psi_{lk}\} : 0 \leq k \leq 2^{J_0} - 1, 0 \leq k' \leq 2^l - 1, l \in \mathbb{N}, l \geq J_0 \} \) form an orthonormal basis for \( L^2([0, 1]) \). Let \( K(l) = \{0, \ldots, 2^l - 1\}^d \) and \( \mathcal{I} \) is the set of sequence \( i = (i_1, \ldots, i_d) \) of zeros and ones excluding \( i = (0, \ldots, 0) \). Standard theory suggests a wavelet series for \( f \in L^2([0, 1]^d) \) can be written as
\[
f = \sum_{k \in K(J_0)} \langle f, \Phi_k \rangle \Phi_k + \sum_{l=J_0}^{\infty} \sum_{i \in \mathcal{I}, k \in K(l)} \langle f, \Psi^i_{l,k} \rangle \Psi^i_{l,k},
\]
where
\[
\Phi_k(x) := \Phi_{J_0,k}(x) = \phi_{J_0,k_1}(x_1) \cdots \phi_{J_0,k_d}(x_d),
\]
\[
\Psi^i_{l,k}(x) := \psi^i_{l,k_1}(x_1) \cdots \psi^i_{l,k_d}(x_d);
\]
here \( \psi^0_{l,k}(\cdot) := \phi_{l,k}(\cdot) = 2^{l/2} \phi(2^l \cdot - k) \) and \( \psi^1_{l,k}(\cdot) := \psi_{l,k}(\cdot) \). In the proof, we will mainly use the following properties of this wavelet basis:
\begin{itemize}
  \item \( \| \sum_{k \in K(J_0)} |\Phi_k|_\infty \| \lesssim 2^{J_0 \alpha/2}, \| \sum_{k \in K(l)} |\Psi^i_{l,k}|_\infty \| \lesssim 2^{ld/2} \),
  \item for \( f \in H^\alpha([0, 1]^d) \) with \( \alpha < q \), \( \sup_{i \in \mathcal{I}, k \in K(l)} |\langle f, \Psi^i_{l,k} \rangle| \lesssim 2^{-l(\alpha + d/2)} \) for all \( l \geq 0 \).
\end{itemize}
Consider the space spanned above tensor wavelet basis
\[
\mathcal{F}_n = \mathcal{F}_{n,w} := \left\{ \sum_{k \in K(J_0)} \alpha_{J_0,k} \Phi_k + \sum_{l=J_0}^{J_n} \sum_{i \in \mathcal{I}, k \in K(l)} \beta_{l,k} \Psi^i_{l,k} : \alpha_{J_0,k}, \beta_{l,k} \in \mathbb{R} \right\}.
\]
If \( f \in H^\alpha([0, 1]^d) \) and wavelet basis is \( q \)-regular (\( q > \alpha \)), for \( J_n \) sufficiently large, one has
\[
\inf_{f^\infty \in \mathcal{F}_{n,w}} \| f - f^\infty \|_\infty \lesssim 2^{-\alpha J_n}.
\]
Compared with the approximation errors using the trigonometric basis and the B-splines basis, the factor \( 2^{J_n} \) may be viewed as the effective number of terms for wavelet basis.
1.6 Roadmap

1.6.1 Chapter 2

In this chapter, we study the $L^\infty$- posterior contraction rates for the nonparametric multivariate functions and/or their derivatives in binary regression, Poisson regression, density estimation and Gaussian white noise models. Contraction rates of level sets in the Hausdorff distance and the Lebesgue measure of the symmetric difference are obtained accordingly. Similar results for level sets in the Gaussian nonparametric regression setting are also presented. In addition, we study credible regions for the level sets in both Gaussian white noise models using trigonometric basis prior and Gaussian nonparametric regression using B-splines basis prior.

1.6.2 Chapter 3

We study the posterior contraction rates for the filament points on some appropriately defined integral curves and for the Hausdorff distance of the filament. We also provide credible sets with sufficient frequentist coverage for the filaments. Finally, we assess the performance of the proposed method in simulations and application to an earthquake dataset.

1.6.3 Chapter 4

In this chapter, we provide a direct Bayesian method to estimate the level curve in a nonparametric regression setting. The level curve is assumed to be a smooth simple closed curve. The squared exponential periodic (SEP) Gaussian process prior is used for the curve. Some simulation results show the method is promising and further study will be conducted in the future.

1.6.4 Appendices

Appendix A provides some elementary lemmas. Appendix B provides some master theorems for posterior contraction rates using finite random series priors.
Chapter 2

Posterior Contraction and Credible Regions for Level Sets

2.1 Introduction

For some constant $c$, the $c$-level set for a smooth function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is defined as the set $\{x \in \mathbb{R}^d : f(x) \geq c\}$, or $\{x \in \mathbb{R}^d : f(x) = c\}$. Level sets estimation have wide applications, as finding observation region that is related to certain level can have important scientific implication in many fields of studies. In addition, it has close relationship with the problems like clustering (Cuevas, Febrero and Fraiman; 2000; Rinaldo and Wasserman; 2010), support estimation of the density function (Biau, Cadre and Pelletier; 2008; Cuevas and Fraiman; 1997) and binary classification (Mammen and Tsybakov; 1999).

The most common approach to the estimation of level sets is the so-called “plug-in” approach, i.e., to estimate the set with $f$ replaced by some nonparametric estimate $\hat{f}$. Under some suitable metrics, convergence rates for density level set estimates were studied in Tsybakov (1997), and later in Baillo, Cuesta-Albertos and Cuevas (2001); Cadre (2006) and Rigollet and Vert (2009) using plug-in methods. Asymptotic normality of some general measure of the symmetric difference between the level set and a plug-in estimator was studied in Mason and Polonik (2009). Cavalier (1997) studied nonparametric regression level sets and Cuevas, González-Manteiga and Rodríguez-Casal (2006) studied level sets of a general smooth function. Besides the plug-in estimation, a direct approach called “excess mass approach” was proposed and studied in Polonik (1995) and Polonik and Wang (2005) for density and regression function respectively. More recent papers studied statistical inference for level sets which include Chen, Genovese and Wasserman (2017); Jankowski and Stanberry (2012) and Mammen and Polonik (2013). The constructions of confidence regions proposed in the first two papers both rely on the estimates of the sup-norm distance of $f$ and $\hat{f}$, while the third paper made use of the es-
timate of variation in the Hausdorff metric directly. Inferential problems of similar types have also been studied in econometrics literature where the objects of interests are identified as sets (Chernozhukov et al.; 2007, 2015).

All above literature is studied in the frequentist framework and the inferences about level sets mainly rely on bootstrap procedure. To the best of our knowledge, there is only one published paper on level set estimation using the Bayesian approach (Gayraud and Rousseau; 2005). The paper studied mainly general contraction rates for the level set \( \{ f \geq c \} \) in density estimation with respect to the Lebesgue measure of symmetric difference. In contrast, we study contraction rates for the level set \( \{ f = c \} \) (or \( \{ \sqrt{f} = c \} \) when \( f \geq 0 \) and \( c > 0 \)) in the same spirit of plug-in approach.

This chapter is organized as follows. Some preliminary materials are given in Section 2.2. In Section 2.3 we obtain the \( L^\infty \)-rates for some nonparametric functions in various settings and obtain the corresponding contraction rates for level sets. To be specific, Section 2.3.1 studies the \( L^\infty \)-contraction rates for the signal function in the Gaussian white noise model for two different random series priors—the trigonometric basis prior and wavelets basis prior. The corresponding rates for level sets in the Hausdorff distance and the Lebesgue measure of symmetric difference are obtained accordingly. For the remaining settings, we shall use tensor product B-splines priors. Section 2.3.2 addresses the same issues in the classical nonparametric regression case. In Section 2.3.3, 2.3.4 and 2.3.5, we focus on the square roots of the mean functions for binary regression, Poisson regression and the square root of density function, since \( \{ x : f(x) = c \} = \{ x : \sqrt{f(x)} = \sqrt{c} \} \) when \( f \geq 0 \) and \( c > 0 \). We derive the \( L^\infty \)-contraction rates and obtain the corresponding rates for level sets. In Section 2.3.6, credible regions for level sets with sufficient frequentist coverage are studied for Gaussian white noise model using trigonometric series priors and nonparametric regression using B-splines series priors. In Section 2.4, we conduct two sets of simulations. The first set of simulation is to compare in Poisson model the proposed strategy of modeling the square root of the mean function with the usual strategy of direct modeling of the mean function. The second set of simulation is to assess the performance of the proposed credible region of level curves in Gaussian nonparametric regression setting. The proof of main results are all given in the Section 2.5.

2.2 Preliminaries

Let \( f : \mathbb{R}^d \to \mathbb{R} \) and consider some value \( c \) that belongs to the range of \( f \). Then the level set at \( c \) is given by

\[
\mathcal{L} := \mathcal{L}(f) = \{ x : f(x) = c \}.
\]

We shall assume the following.
(A1) For some small $\epsilon > 0$ and $\delta_1 > 0$, any $\tilde{c} \in [c - \epsilon, c + \epsilon]$, for all $x$ such that $|f(x) - \tilde{c}| \leq \delta_1$, we have $d(x\{f = \tilde{c}\}) \leq A|f(x) - \tilde{c}|^{\nu_1}$ for some $\nu_1 > 0$.

Assumption (A1) preclude functions arbitrarily flat around the level $c$. In particular, the condition implies that $\mathcal{L}$ does not include any stationary point of the function $f$ or any flat part of $f$ at the level $c$. Hence points of local extrema are excluded. To see this, consider $t_n = c + \epsilon_n$, for some small $\epsilon_n > 0$ such that $\epsilon_n < \delta_1$. Clearly, for $x \in \mathcal{L}$, $|f(x) - t_n| = \epsilon_n < \delta_1$, thus $d(x\{f = t_n\}) \leq A|f(x) - t_n|^{\nu_1}$, thus $\inf_{x_n \in \{f = t_n\}} \|x - x_n\| \leq A\epsilon_n^{\nu_1}$. Therefore, there exists some $x_n$ such that $f(x_n) = t_n$ and $\|x - x_n\| \leq A\epsilon_n^{\nu_1}$. That is, there exists some sequence $x_n \rightarrow x$ and $f(x_n) > c$. Similarly, it is straightforward to see that there also exists some sequence $x_n \rightarrow x$ and $f(x_n) < c$.

**Lemma 2.2.1** A sufficient condition for (A1) with $\nu_1 = 1$ is that there exists some $\epsilon_1 > 0, c_0 > 0$ such that $\inf_{x \in \{f - c\leq \epsilon_1\}} \|\nabla f(x)\| > c_0$.

The readers now may want to first recall the definition of the Hausdorff distance in (1.8). The following two lemmas are slight generalization of Theorem 1 and Theorem 2 of Cuevas et al. (2006).

**Lemma 2.2.2** Suppose that the continuous functions $f$ satisfies (A1), and there is a sequence of continuous function $f_n$ such that $\|f - f_n\|_{\infty} \rightarrow 0$. Then for $n$ sufficiently large,

$$\text{Haus}(\{f = c\}, \{f_n = c\}) \leq C\|f - f_n\|_{\infty}^{\nu_1}.$$  

for some constant $C > 0$ which can be taken as $6A$.

To consider the Lebesgue measure of the symmetric difference as metric, we assume the following.

(A1') Let $\lambda$ be the Lebesgue measure and $\lambda\{c - \epsilon < f < c + \epsilon\} \leq A_2\epsilon^{\nu_2}$, for some $\nu_2 > 0$.

**Lemma 2.2.3** Suppose that the continuous functions $f$ satisfies (A1'), and there is a sequence of continuous function $f_n$ such that $\|f - f_n\|_p \rightarrow 0$ for $1 \leq p < \infty$. Then for some constant $C > 0$,

$$\lambda(\{f \geq c\} \triangle \{f_n \geq c\}) \leq C\|f - f_n\|_p^{\nu_2p/(\nu_2+p)}.$$  

In particular, if $\|f - f_n\|_{\infty} \rightarrow 0$, then for some constant $C > 0$,

$$\lambda(\{f \geq c\} \triangle \{f_n \geq c\}) \leq C\|f - f_n\|_{\infty}^{\nu_2}.$$  

In view of Lemma 2.2.1 and to simplify the exposition, we shall use $\nu_1 = 1$ and $\nu_2 = 1$ in the following sections.
2.3 Applications in various inference problems

We shall use superscript * to indicate the true values. Let \( f^* \in \mathcal{H}^\alpha([0,1]^d) \) for some real \( \alpha > 0 \). We model \( f \) as \( f = \sum_{j=1}^J \theta_j \psi_{j,j} \) for some basis functions \( \psi_{j,j} = \{\psi_{j,1}, \ldots, \psi_{j,J}\} \) and impose some prior on the coefficients \( \theta \). For Gaussian white noise models, we consider two different priors—random trigonometric basis prior and random wavelet basis prior.

For the Gaussian nonparametric regression, binary regression, Poisson regression and density estimation, we shall use the tensor product B-splines functions. Some basics about B-splines functions have been already discussed in Section 1.5. For Gaussian white noise models, we consider two different priors—random trigonometric basis prior and random wavelet basis prior.

For regression applications such as the Gaussian nonparametric regression, binary regression and Poisson regression but allow only the fixed design in Poisson regression. For the Gaussian nonparametric regression, binary regression, Poisson regression and density estimation settings, we would also make use of some general theory of posterior contraction rates given in Appendix B in order to obtain some preliminary rates. Some brief discussion of the general theory of posterior contraction has also been given in Section 1.2. We let \( d \) to be the \( L^\infty \)-metric, \( S(M) \subset [-M, M] \) for some large positive \( M \). We list the conditions here.
(P1) \( \Pi(\|\theta_j - \theta_0\| \leq \epsilon) \geq \exp(-c_1J \log(\sqrt{J}/\epsilon)) \) for every \( \theta_0 \in \mathbb{R}^J \) with \( \|\theta_0\|_\infty \leq H \) for some positive constants \( c_1 \) and \( H \), and every sufficiently small \( \epsilon > 0 \).

(P2) There exists some \( M_0 > 0 \) such that \( \Pi(\theta_j \notin S(M)^J) \leq J \exp(-c_2M^\tau) \) for some positive constant \( c_2 \) and all \( M \geq M_0 > 0 \) for some large constant \( \tau > 0 \).

(C1) For every \( \theta_1, \theta_2 \in \mathbb{R}^J \), \( d(\theta_1^T \psi, \theta_2^T \psi) \leq \rho(J)\|\theta_1 - \theta_2\| \) for some positive increasing function \( \rho \) which is some multiple of polynomial in \( J \).

Lastly, to present the contraction rates we shall use \( \mathbb{D}_n \) to denote all observations.

### 2.3.1 Gaussian white noise model

Consider the multivariate white noise model which is defined through the stochastic differential equation \( dY(t) = f(t)dt + n^{-1/2}dW(t) \) for \( t \in [0,1]^d \) and \( W \) is independent standard Brownian motions \( (W_1(t_1), \ldots, W_d(t_d)) \) (see Section 1.2.2 of Giné and Nickl (2015)). Let \( \{e_i : i = 1, 2, \ldots, J\} \) be an orthonormal basis of \( L^2([0,1]^d) \). Then the equivalent sequence space model is \( y_i = \theta_i + n^{-1/2}e_i \), \( y_i := \int e_i(t)Y(t) dt, \theta_i := \langle e_i, f \rangle = \int e_i(t)f(t)dt \) and \( e_i \overset{i.i.d.}{\sim} N(0,1), i = 1, 2, \ldots \). We shall consider two separate random series priors in this sections. One is using tensor product wavelet basis and the other one is using tensor product trigonometric basis.

**Trigonometric basis**

First recall some facts about trigonometric basis from Section 1.5.2. Let \( j = (j_1, \ldots, j_d) \in \{0, 1, 2, \ldots\}^d \) and \( \mathcal{K}(j) := \{k \in \{0, 1\}^d : k_i = 0 \text{ when } j_i = 0, i = 1, \ldots, d\} \). We denote the tensor product trigonometric basis functions by \( \{\phi_{jk} : (j, k) \in \{0, 1, 2, \ldots\}^d \times \mathcal{K}(j)\} \), where

\[
\phi_{jk}(x) = \prod_{i=1}^d \sqrt{2}[(1 - k_i) \cos(2\pi j_i x_i) + k_i \sin(2\pi j_i x_i)].
\]

Assume that the function \( f^* : [0, 1]^d \mapsto \mathbb{R} \) can be written as

\[
f^*(x) = \sum_{j_1} \cdots \sum_{j_d} \sum_{k \in \mathcal{K}(j)} \phi_{jk} \theta_{0,jk},
\]

for some collection of coefficients \( \{\theta_{0,jk} : (j, k) \in \{0, 1, 2, \ldots\}^d \times \mathcal{K}(j)\} \). The equivalent sequence model is \( y_{jk} = \theta_{jk} + n^{-1/2}e_{jk} \). If the prior is that \( \theta_{jk} \overset{\text{ind}}{\sim} N(0, \mu_j) \), the posterior distribution of \( \theta_{jk} \) is Gaussian with mean \( \hat{\theta}_{jk} = n\mu_j y_{jk}/(n\mu_j + 1) \) and variance \( \mu_j/(n\mu_j + 1) \). The following theorem gives the posterior contraction rate for \( d = 2 \). For \( d > 2 \), we believe the same rate (with modification due to dimensions) may be obtained but the derivation can be cumbersome.
Proposition 2.3.1 Assume that $\sum_{j_1=0}^{\infty} \sum_{j_2=0}^{\infty} \sum_{k \in K(j_1,j_2)} \max_{i_1+i_2=\alpha,i_1,i_2 \geq 0} |\theta_{0,jk}| \leq L$ for some positive constant $L$ and $\alpha > 2$. Suppose prior is imposed so that $\theta_{jk} \overset{\text{ind}}{\sim} N(0,\mu_{j_1},\mu_{j_2})$, where $\mu_{j_i} = j_i^{-(1+\alpha)}$ if $j_i \neq 0$ but $\mu_{j_i} = 1$ if $j_i = 0$, for all $j_i \in \{0,1,\ldots\}$, $i = 1,2$. We have for arbitrary $K_n \to \infty$,

$$
\Pi\left(\|f - f^*\|_\infty > K_n n^{-\alpha/(2\alpha+2)} \log n \|D_n\| \right) \overset{P_0}{\to} 0.
$$

Theorem 2.3.2 Under the assumptions of Proposition 2.3.1, for every $K_n \to \infty$,

(i) if $f^*$ satisfies Assumption (A1), then

$$
\Pi\left(\text{Haus}\{f = c\}, \{f^* = c\}\right) > K_n n^{-\alpha/(2\alpha+2)} \log n \|D_n\| \overset{P_0}{\to} 0.
$$

(ii) if $f^*$ satisfies Assumption (A1'), then

$$
\Pi\left(\lambda\{f \geq c\} \triangle \{f^* \geq c\}\right) > K_n n^{-\alpha/(2\alpha+2)} \log n \|D_n\| \overset{P_0}{\to} 0.
$$

Wavelet basis

Recall some basics about wavelet basis from Section 1.5.4. Let $\{\phi_k, \psi_{lk} : 0 \leq k \leq 2^{J_0} - 1, 0 \leq k' \leq 2^l - 1, l \in \mathbb{N}, l \geq J_0\}$ denotes the CDV-wavelet basis for $L^2([0,1])$. Let $K(l) = \{0,\ldots,2^l-1\}^d$ and $I$ be the set of sequence $i = (i_1,\ldots,i_d)$ of zeros and ones excluding $i = \{0,\ldots,0\}$. A wavelet series for $f \in L^2([0,1]^d)$ can be written as

$$
f = \sum_{k \in K(J_0)} \langle f, \Phi_k \rangle \Phi_k + \sum_{l=J_0}^{\infty} \sum_{i \in I, k \in K(l)} \langle f, \psi_{l,k}^i \rangle \psi_{l,k}^i,
$$

where $\Phi_k(x) := \phi_{J_0 k}(x_1) \cdots \phi_{J_0 k}(x_d)$ and $\psi_{l,k}^i(x) := \psi_{l,k}^{i_1}(x_1) \cdots \psi_{l,k}^{i_d}(x_d); \text{ here } \psi_{l,k}^{i_1}(\cdot) := \phi_{l,k}^{i_1}(\cdot) = 2^{l/2} \phi(2^l \cdot - k)$ and $\psi_{l,k}^{i_d}(\cdot) := \psi_{l,k}^{i_d}(\cdot)$. In the proof, we shall mainly use the following properties of this wavelet basis:

1. $\|\sum_{k \in K(J_0)} |\Phi_k||_\infty \lesssim 2^{J_0d/2}$, $\|\sum_{k \in K(l)} |\psi_{l,k}||_\infty \lesssim 2^{ld/2}$,

2. for any $f \in \mathcal{H}^\alpha([0,1]^d)$ with $\alpha < q$, we have $\sup_{i \in I, k \in K(l)} |\langle f, \psi_{l,k}^i \rangle| \lesssim 2^{-l(\alpha+d/2)}$ for all $l \geq 0$.

Now with above choice of the basis, the equivalent sequence space model is given by

$$
y_k = \theta_k + \frac{1}{\sqrt{n}} \varepsilon_k, \quad k \in \{0,\ldots,2^{J_0} - 1\}^d,
$$

$$
y_{l,i,k} = \theta_{l,i,k} + \frac{1}{\sqrt{n}} \varepsilon_{l,i,k}, \quad k \in \{0,\ldots,2^l - 1\}^d, l \geq J_0, i \in I,
$$
where the parameters are given by \((\theta_k, \theta_{l,i,k}) := (\langle f, \Phi_k \rangle, \langle f, \Psi_{l,i,k}^j \rangle)\) and \(\varepsilon_k, \varepsilon_{l,i,k}\) all i.i.d. \(N(0,1)\).

We shall put prior \(\theta_k \overset{i.i.d.}{\sim} N(0,1), \theta_{l,i,k} \overset{ind}{\sim} N(0,\mu_l)\) for some constant \(\mu_l\) to be defined in the theorem below. The posterior distribution of \(f\) given \(D_n\) then is given by the law of

\[
\sum_{k \in K(J_0)} \left( \frac{n}{n+1} y_k + \left( \frac{1}{n+1} \right)^{1/2} \tilde{\varepsilon}_k \right) \Phi_k + \sum_{l=J_0}^{\infty} \sum_{i \in I, k \in K(l)} \left( \frac{n \mu_l}{n \mu_l + 1} y_{l,i,k} + \left( \frac{\mu_l}{n \mu_l + 1} \right)^{1/2} \tilde{\varepsilon}_{l,i,k} \right) \Psi_{l,i,k}^j,
\]

with the variables \(\tilde{\varepsilon}_k, \tilde{\varepsilon}_{l,i,k}\) all i.i.d \(N(0,1)\), independent of all \(\varepsilon_k, \varepsilon_{l,i,k}\). The following proposition is our multivariate version of Theorem 1 of Giné and Nickl (2011).

**Proposition 2.3.3** Using \(q\)-times continuously differentiable \((q > \alpha)\) CDV-wavelet tensor basis with above prior, \(\mu_l = l^{-1/2} - l^{(2\alpha + d)}\) we have for arbitrary \(K_n \to \infty,\)

\[
\Pi(\|f - f^*\|_{\infty} > K_n \left( \frac{\log n}{n} \right)^{\frac{\alpha}{d+2\alpha}} | D_n) \overset{P_0}{\to} 0.
\]

**Theorem 2.3.4** Under the assumptions of Proposition 2.3.3, for every \(K_n \to \infty,\)

(i) if \(f^*\) satisfies Assumption (A1), then

\[
\Pi\left(\text{Haus}([f = c], [f^* = c]) > K_n \left( \frac{\log n}{n} \right)^{\frac{\alpha}{d+2\alpha}} | D_n) \overset{P_0}{\to} 0.
\]

(ii) if \(f^*\) satisfies Assumption (A1'), then

\[
\Pi\left(\lambda([f \geq c] \triangle [f^* \geq c]) K_n \left( \frac{\log n}{n} \right)^{\frac{\alpha}{d+2\alpha}} | D_n) \overset{P_0}{\to} 0.
\]

### 2.3.2 Gaussian regression

Consider the nonparametric regression model

\[Y_i = f(X_i) + \varepsilon_i,\]

where \(\varepsilon_i \overset{i.i.d.}{\sim} N(0,\sigma^2)\) for \(i = 1,\ldots,n\). Both \(f\) and \(\sigma\) are unknown parameters. Let \(Y = (Y_1,\ldots,Y_n)^T, X = (X_1^T,\ldots,X_n^T)^T, F = (f(X_1),\ldots,f(X_n))^T\) and \(\varepsilon = (\varepsilon_1,\ldots,\varepsilon_n)^T\), then we can write \(Y = F + \varepsilon\). We model \(f = b^T \theta\), thus model becomes

\[Y|(X,\theta,\sigma^2) \sim N(B\theta,\sigma^2 I_n).\]

We assume that \(Y_i = f^*(X_i) + \varepsilon_i\), where \(\varepsilon_i\) are i.i.d. sub-Gaussian with mean 0 and variance \(\sigma_0^2\) for \(i = 1,\ldots,n\). In this section, the covariates \(X_1,\ldots,X_n\) can be either fixed or random. If
they are random, we assume that \((Y_i, X_i)_{i=1}^n\) are i.i.d. and independent of errors \(\varepsilon_1, \ldots, \varepsilon_n\). We assign \(\theta|\sigma^2 \sim N(\theta_0, \sigma^2 \Lambda_0)\), assuming that for some constants \(0 < c_1 \leq c_2 < \infty\) we have that
\[ c_1 I \leq \Lambda_0 \leq c_2 I, \]
where \(I\) is the \(\prod J_k \times \prod J_k\) identity matrix. It follows that
\[ \theta|Y, \sigma^2 \sim N((\Lambda_0^{-1} + B^T B)^{-1}(B^T Y + \Lambda_0^{-1} \theta_0), \sigma^2 (B^T B + \Lambda_0^{-1})^{-1}). \]

The posterior distribution for \(f(x)\) and its partial derivatives are then obtained accordingly. Define the vector
\[
\begin{align*}
b_J^{(r)}(x) &= \left( \frac{\partial^r}{\partial x_j^{q_1}} B_{j_1, q_1}(x_1) \cdots \frac{\partial^r}{\partial x_d^{q_d}} B_{j_d, q_d}(x_d) : 1 \leq j_k \leq J_k, k = 1, \ldots, d \right)^T.
\end{align*}
\]
Therefore,
\[
\Pi(D^r f|\mathbb{D}_n, \sigma^2) \sim GP(A_r Y + C_r \theta_0, \sigma^2 \Sigma_r),
\]
where \(GP\) denotes a Gaussian process and
\[
\begin{align*}
A_r(x) &= b_J^{(r)}(x)^T (B^T B + \Lambda_0^{-1})^{-1} B^T, \\
C_r(x) &= b_J^{(r)}(x)^T (B^T B + \Lambda_0^{-1})^{-1} \Lambda_0^{-1}, \\
\Sigma_r(x, y) &= b_J^{(r)}(x)^T (B^T B + \Lambda_0^{-1})^{-1} b_J^{(r)}(y).
\end{align*}
\]
Note that under \(P_0\), \(A_r(x)\varepsilon/\sigma_0\) is a mean zero process with sub-Gaussian tail.

Finally, regarding \(\sigma^2\), we can either put a conjugate inverse-gamma prior such that \(\sigma^2 \sim IG(a/2, b/2)\) with shape parameter \(a/2 > 2\) and rate parameter \(b/2 > 0\) or plug-in an estimate for \(\sigma^2\). Since the choice will not affect the theory too much, for ease of exposition, we shall use the second approach—also called the empirical Bayes approach. The empirical Bayes has the following posterior distribution
\[
\Pi(D^r f|\mathbb{D}_n, \sigma^2) |_{\sigma^2 = \hat{\sigma}^2} \sim GP(A_r Y + C_r \theta_0, \hat{\sigma}^2 \Sigma_r),
\]
where
\[
\hat{\sigma}^2_n = n^{-1} (Y - B \theta_0)^T (B \Lambda_0 B^T + I_n)^{-1} (Y - B \theta_0).
\]
The following result is from Yoo and Ghosal (2016). Let \(\tilde{f} := AY + C \theta_0\) be the posterior mean of \(f\).
Proposition 2.3.5 For $J_k \asymp (n/\log n)^{1/(2\alpha + d)}$, $k = 1, \ldots, d$, we have for any $K_n \to \infty$,

\[
\Pi\left(\|D^r f - D^r f^*\|_\infty > K_n \left(\frac{n}{\log n}\right)^{\frac{|r| - \alpha}{d + 2\alpha}}\|D_n\|_n\right) \to 0,
\]

\[
\Pi\left(\|D^r f - D^r \tilde{f}\|_\infty > K_n \left(\frac{n}{\log n}\right)^{\frac{|r| - \alpha}{d + 2\alpha}}\|D_n\|_n\right) \to 0,
\]

and

\[
\|D^r \tilde{f} - D^r f^*\|_\infty = O_P\left(\left(\frac{n}{\log n}\right)^{\frac{|r| - \alpha}{d + 2\alpha}}\right).
\]

Theorem 2.3.6 Under the assumptions of Proposition 2.3.5, for every $K_n \to \infty$,

(i) if $D^r f^*$ satisfies Assumption (A1),

\[
\Pi\left(\text{Haus}(\{D^r f = c\}, \{D^r f^* = c\}) > K_n \left(\frac{n}{\log n}\right)^{\frac{|r| - \alpha}{d + 2\alpha}}\|D_n\|_n\right) \to 0,
\]

(ii) if $D^r f^*$ satisfies Assumption (A1'),

\[
\Pi\left(\lambda(\{D^r f \geq c\} \triangle \{D^r f^* \geq c\}) > K_n \left(\frac{n}{\log n}\right)^{\frac{|r| - \alpha}{d + 2\alpha}}\|D_n\|_n\right) \to 0.
\]

2.3.3 Binary regression

Consider binary regression with the random design, with observations $(Y_i, X_i)_{i=1}^n$ where $Y_i | X_i \sim \text{Bin}(1, f(X_i))$. We denote the conditional density by $p_f(y|x) = f(x)^y(1 - f(x))^{1-y}$. The results in this section hold for both fixed design and random design cases.

Proposition 2.3.7 Suppose $\sqrt{f^*} \in \mathcal{H}^\alpha([0,1]^d)$ for $\alpha > d/2$. Let $\sqrt{f} = \theta^T b$ and $\theta$ satisfy Condition (P1) with each element restricted to $(0,1)$. Then with $J_k \asymp (n/\log n)^{1/(2\alpha + d)}$ for $k = 1, \ldots, d$, we have for every $K_n \to \infty$,

\[
\Pi\left(\|\sqrt{f} - \sqrt{f^*}\|_\infty > K_n \left(\frac{n}{\log n}\right)^{\frac{d - 2\alpha}{2(\alpha + 2\alpha)}}\|D_n\|_n\right) \to 0.
\]

Also, for $r$ such that $2|r| < 2\alpha - d$, we have

\[
\Pi\left(\|D^r \sqrt{f} - D^r \sqrt{f^*}\|_\infty > K_n \left(\frac{n}{\log n}\right)^{\frac{d + 2|r| - 2\alpha}{2(\alpha + 2\alpha)}}\|D_n\|_n\right) \to 0.
\]

Note the the Condition (P2) trivially holds for $S(M) = (0,1)$ and $M_0 = 1$ and thus is not stated in the above proposition.
Theorem 2.3.8 Under the assumptions of Proposition 2.3.7, for \( r \) such that \( 2|r| < 2\alpha - d \), we have for every \( K_n \to \infty \),

(i) if \( D^r \sqrt{f^*} \) satisfies Assumption (A1), then

\[
\Pi \left( \text{Haus}(\{D^r \sqrt{f} = c\}, \{D^r \sqrt{f^*} = c\}) > K_n \left( \frac{n}{\log n} \right)^{\frac{d+2|r| - 2\alpha}{2(d+2\alpha)}} | \mathbb{D}_n \right) \xrightarrow{P_0} 0;
\]

(ii) if \( D^r \sqrt{f^*} \) satisfies Assumption (A1'), then

\[
\Pi \left( \lambda(\{D^r \sqrt{f} \geq c\} \searrow \{D^r \sqrt{f^*} \geq c\}) > K_n \left( \frac{n}{\log n} \right)^{\frac{d+2|r| - 2\alpha}{2(d+2\alpha)}} | \mathbb{D}_n \right) \xrightarrow{P_0} 0.
\]

For the fixed design case, \( Y_i \overset{\text{ind}}{\sim} \text{Bin}(1, f^*(X_i)) \), the same contraction rates results as described above can be obtained.

2.3.4 Poisson regression

For the Poisson regression model, we consider the fixed design case where \( Y_i \overset{\text{ind}}{\sim} \text{Poi}(f(X_i)) \). We denote the conditional density by \( p_f(y|x) = \exp \left( -f(x) \right) \frac{f(x)^y}{y!} \).

Proposition 2.3.9 Suppose \( \sqrt{f^*} \) is a positive function that belongs to \( H^\alpha([0, 1]^d) \) for \( \alpha > \frac{d}{2} \). Let \( \sqrt{f} = \theta^T b \) with each element of \( \theta \) restricted to \( (0, \infty) \). Then in the fixed design, under Conditions (P1) and (P2) with \( J_k \propto (n/\log n)^{1/(2\alpha + d)} \) for \( k = 1, \ldots, d \), we have for every \( K_n \to \infty \),

\[
\Pi \left( \|\sqrt{f} - \sqrt{f^*}\|_\infty > K_n \left( \frac{n}{\log n} \right)^{\frac{d-2\alpha}{2(d+2\alpha)}} | \mathbb{D}_n \right) \xrightarrow{P_0} 0.
\]

Also, for \( r \) such that \( 2|r| < 2\alpha - d \), we have

\[
\Pi \left( \|D^r \sqrt{f} - D^r \sqrt{f^*}\|_\infty > K_n \left( \frac{n}{\log n} \right)^{\frac{d+2|r| - 2\alpha}{2(d+2\alpha)}} | \mathbb{D}_n \right) \xrightarrow{P_0} 0.
\]

Note a sufficient condition for the Condition (P2) with \( S(M) = [1/M, M] \) is that for each element \( \theta_j \) of \( \theta \), \( \Pi(\theta_j > M) \leq \exp(-a_1 M^r) \) for \( M > 0 \) large and \( \Pi(\theta_j < M) \leq \exp(-a_2 M^r) \) for \( M > 0 \) small. One possible family of the prior of this type would be the generalized inverse-Gaussian distribution whose density is proportional to \( x^{a_3} e^{-(a_1 x + a_2/x)} \) for some \( a_1, a_2 > 0 \) and \( a_3 \in \mathbb{R} \).

Theorem 2.3.10 Under the assumptions of Proposition 2.3.9, for \( r \) such that \( 2|r| < 2\alpha - d \), we have for every \( K_n \to \infty \),
(i) if $D'\sqrt{f}$ satisfies Assumption (A1), then
\[
\Pi\left(\text{Haus}\{D'\sqrt{f} = c\}, \{D'\sqrt{f} = c\}\right) > K_n\left(\frac{n}{\log n}\right)^{\frac{d+2r(\alpha-2\alpha)}{2(d+2\alpha)}} |D_n| \xrightarrow{P} 0;
\]
(ii) if $D'\sqrt{f^\ast}$ satisfies Assumption (A1’), then
\[
\Pi\left(\lambda\{D'\sqrt{f} \geq c\} \triangle \{D'\sqrt{f^\ast} \geq c\}\right) > K_n\left(\frac{n}{\log n}\right)^{\frac{d+2r(\alpha-2\alpha)}{2(d+2\alpha)}} |D_n| \xrightarrow{P} 0.
\]

### 2.3.5 Density estimation

We have $X_i \overset{i.i.d.}{\sim} P_0$, whose density function is denoted by $p_0$. For density function $p$ supported on $[0,1]^d$, it may be given a prior of the form $p(x) = \Psi(f)/\int \Psi(f)du$ for a dominating measure $\mu$ and some link function $\Psi: \mathbb{R} \to (0,\infty)$, where $f$ may be written as some form of basis expansion. Here we take a similar approach. We use normalized B-spline functions defined as
\[
\tilde{B}_{j_1,q_1,\ldots,j_d,q_d}(x) := \left(\frac{B_{j_1,q_1}(x_{j_1})}{\int_{[0,1]} B_{j_1,q_1}^2(u_{j_1})du_{j_1}}\right)^{1/2} = \frac{B_{j_1,q_1}(x_{j_1})}{C_{j_1,q_1}},
\]
\[
\hat{b}_{j_1,\ldots,j_d,q_1,\ldots,q_d}(x) := \left(\tilde{B}_{j_1,\ldots,j_d}(x) = \prod_{k=1}^d \tilde{B}_{j_k,q_k}(x_{j_k}) : 1 \leq j_k \leq J_k\right)^T.
\]

The elements in the vector $\hat{b}$ are sorted in the lexicographical order according to their indices. Since $\theta^T(\int \hat{b}\hat{b}^T)\theta = \int (\hat{b}^T\theta)^2 = 0$ if and only if $\theta = 0$, the matrix $\int \hat{b}\hat{b}^T$ is positive definite. We can write
\[
\int \hat{b}\hat{b}^T = \Gamma^{-1/2}\Xi\Gamma^{-1/2},
\]
where $\Gamma$ and $\Xi$ both are $\prod_{k=1}^d J_k$ by $\prod_{k=1}^d J_k$ matrix. Given two $d$-multi-index $j$ and $j'$, the $(j,j')$-th element of $\Gamma$ is given by $\Gamma_{j,j'} = C_{j_1,q_1}^2 \times \cdots \times C_{j_d,q_d}^2 = \prod_{k=1}^d \int B_{j_k,q_k}^2(u_{j_k})du_{j_k}$; and the $(j,j')$-th element of $\Xi$ is given by $\Xi_{j,j'} = \prod_{k=1}^d \int B_{j_k,q_k}(u_{j_k})B_{j_k,q_k}'(u_{j_k})du_{j_k}$. Clearly, $\Gamma$ is a diagonal matrix. Because the local support property of the B-splines, the matrix $\Xi$ is $q$-banded (Yoo and Ghosal; 2016) which means that $\Xi_{j,j'} = 0$ whenever $|j_k - j'_k| \geq q_k$ for some $k = 1,\ldots,d$.

We shall write $\sqrt{p} := \sqrt{f} = \hat{b}^T\theta$. To ensure that $\int p = 1$, $\theta$ should satisfy that $\theta^T\left(\int \hat{b}\hat{b}^T\right)\theta = 1$, specifically, $\theta^T\Gamma^{-1/2}\Xi\Gamma^{-1/2}\theta = 1$.

**Proposition 2.3.11** Suppose $\sqrt{f} \in H^\alpha([0,1]^d)$ for $\alpha > d/2$. Let $\sqrt{f} = \hat{b}^T\theta$ for $\theta$ satisfying $\theta^T\Gamma^{-1/2}\Xi\Gamma^{-1/2}\theta = 1$. Under the Conditions (P1) and (P2) with $J_k \asymp (n/\log n)^{1/(2\alpha+d)}$ for
\[ k = 1, \ldots, d, \text{ we have for every } K_n \to \infty \]
\[
\Pi\left( \| \sqrt{f} - \sqrt{f^*} \|_{\infty} > K_n \left( \frac{n}{\log n} \right)^{\frac{d+2\alpha}{2(d+2\alpha)}} \| D_n \| \right) \xrightarrow{P_0} 0.
\]

Also, for \( r \) such that \( 2|r| < 2\alpha - d \), we have
\[
\Pi\left( \| D^r \sqrt{f} - D^r \sqrt{f^*} \|_{\infty} > K_n \left( \frac{n}{\log n} \right)^{\frac{d+2|r|-2\alpha}{2(d+2\alpha)}} \| D_n \| \right) \xrightarrow{P_0} 0.
\]

Note that for Condition (P2), it suffices to consider \( S(M) = (0, \prod_{k=1}^{d} J_{1/2}) \). This is explained in the proof of Proposition 2.3.11.

**Theorem 2.3.12** Under the assumptions of Proposition 2.3.11, for \( r \) such that \( 2|r| < 2\alpha - d \), we have for every \( K_n \to \infty \),

(i) if \( D^r \sqrt{f^*} \) satisfies Assumption (A1), then
\[
\Pi\left( \text{Haus}\{ D^r \sqrt{f} = c \}, \{ D^r \sqrt{f^*} = c \} \right) > K_n \left( \frac{n}{\log n} \right)^{\frac{d+2|r|-2\alpha}{2(d+2\alpha)}} \| D_n \| \xrightarrow{P_0} 0;
\]

(ii) if \( D^r \sqrt{f^*} \) satisfies Assumption (A1'), then
\[
\Pi\left( \lambda\{ D^r \sqrt{f} \geq c \} \triangle \{ D^r \sqrt{f^*} \geq c \} \right) > K_n \left( \frac{n}{\log n} \right)^{\frac{d+2|r|-2\alpha}{2(d+2\alpha)}} \| D_n \| \xrightarrow{P_0} 0.
\]

### 2.3.6 Credible sets

In this section, we provide credible sets with sufficient coverage for the Gaussian white noise model using trigonometric series prior and the Gaussian nonparametric regression using B-splines series priors.

We start with the first model. Let \( \hat{C} \) be the induced level curve of \( \hat{f} \). For some \( 0 < \gamma < 1/2 \), let \( R_{n,\gamma} \) denotes the \( 1-\gamma \) quantile of the posterior distribution of \( \| f - \hat{f} \|_{\infty} \). Let \( C_{f,\gamma} := \{ f : \| f - \hat{f} \|_{\infty} \leq \rho_n R_{n,\gamma} \} \), where \( \rho_n = \rho \sqrt{\log n} \) for some large constant \( \rho > 0 \). The following theorem provides two valid credible sets with sufficiently high frequentist coverage in the Gaussian white noise model using trigonometric series prior. A similar method is used in Yoo and Ghosal (2017) where they provided credible sets for the mode and the maximum of a regression function.

**Theorem 2.3.13** Under the assumptions of Proposition 2.3.1 and using the same prior, if \( f^* \)
satisfies (A1), then for the following two sets,

\[ C_\mathcal{L} = \{ \mathcal{L}(f) : f \in C_{f,\gamma} \}, \]
\[ \bar{C}_\mathcal{L} = \{ \mathcal{L} : \text{Haus}(\mathcal{L}, \mathcal{\tilde{L}}) \leq C \rho n R_n,\gamma \}, \]

the credibility of \( C_\mathcal{L} \) for \( \mathcal{L} \) and its coverage probability for \( \mathcal{L}(f^*) \) both tend to 1 and \( C_\mathcal{L} \subset \bar{C}_\mathcal{L} \) with high posterior probability with \( P_0 \)-probability tending to 1.

**Remark 2.3.1** If \( f^* \) satisfies (A1'), in view of Lemma 2.2.3, we can set above set \( \bar{C}_\mathcal{L} \) to be \( \{ \mathcal{L}(f) : \lambda(\{ f \geq c \} \Delta \{ \tilde{f} \geq c \}) \leq C \rho n R_n,\gamma \} \).

**Remark 2.3.2** An alternative use of Bayesian posterior to quantify uncertainty is to observe that the true level set (a curve) will be contained in \( \{ x : |\tilde{f}(x) - c| < \rho n R_n,\gamma \} \) with probability tending to 1. But it does not say anything about how close the truth will be to the level curve of \( \tilde{f} \) in Hausdorff metric. Consider the following set,

\[ C'_\mathcal{L} = \{ x : |\tilde{f}(x) - c| < \rho n R_n,\gamma \} \]

Examine the credibility:

\[
\Pi(\mathcal{L} \in C'_\mathcal{L}|D_n) = \Pi(f : \{ x : f(x) = c \} \subset \{ x : |\tilde{f}(x) - c| < \rho n R_n,\gamma \}|D_n)
= \Pi(f : |\tilde{f}(x) - f(x)| < \rho n R_n,\gamma, \forall x \in \{ x : f(x) = c \}|D_n)
\geq \Pi(f : ||\tilde{f} - f||_\infty < \rho n R_n,\gamma|D_n)
\]

which tends to one by the proof of Theorem 2.3.13. Similarly, \( P_0(\mathcal{L}^* \in C'_\mathcal{L}) \geq P_0(||\tilde{f} - f^*||_\infty < \rho n R_n,\gamma) \) which also tends to 1. The advantage of this set \( C'_\mathcal{L} \) is that the posterior mean and \( R_{n,\gamma} \) can be obtained from the posterior. If one computes the set \( \{ x : |\tilde{f}(x) - c| < \rho n R_n,\gamma \} \) numerically, it will serve as a confidence set. Here instead of relying on bootstrap to get quantiles, the cut-off \( R_{n,\gamma} \) is obtained from the posterior.

The following theorem provides two valid credible sets with sufficiently high frequentist coverage in the nonparametric regression settings.

**Theorem 2.3.14** Assume that \( f^* \) satisfies (A1), using \( J_k \asymp (n/\log n)^{1/2(\alpha+1)}, k = 1, \ldots, d \), and some \( \rho > 0 \) sufficiently large, for the following two sets,

\[ C_\mathcal{L} = \{ \mathcal{L}(f) : f \in C_{f,\gamma} \}, \]
\[ \bar{C}_\mathcal{L} = \{ \mathcal{L} : \text{Haus}(\mathcal{L}, \mathcal{\tilde{L}}) \leq C \rho R_n,\gamma \}, \]

30
the credibility of \( C_L \) for \( L \) and its coverage probability for \( L(f^*) \) both tend to 1 and \( C_L \subset \tilde{C}_L \) with high posterior probability with \( P_0 \)-probability tending to 1.

**Remark 2.3.3** Both Remark 2.3.1 and Remark 2.3.2 for the Gaussian white noise model also hold for nonparametric regression. Furthermore, the same method can be used to construct credible region for \( L(D^r f) \) by replacing \( f \) with \( D^r f \) in the Theorem 2.3.14. The proof is essentially the same by noting the rates for the derivatives as given in the Proposition 2.3.5.

### 2.4 Simulation

We give a description of the two generic algorithms we use for finding level sets.

**Algorithm 1: Fixed Point Algorithm**

Set \( \epsilon > 0, \tau > 0, \bar{a} > 0 \) and select a collection of starting points \( \{x_1, ..., x_n\} \), compute \( f(x_i) \) and keep only those points \( x_i \) for which \( f(x_i) > \tau \). Now for each \( x_i \), let \( x_i^{(1)} = x_i \); iterate through the following the following steps from \( t = 1 \):

1. evaluate \( \nabla f(x_i^{(t)}) \);
2. compute \( (f(x_i^{(t)}) - c)\nabla f(x_i^{(t)}) \);
3. update \( x_i^{(t+1)} = -\bar{a}(f(x_i^{(t)}) - c)\nabla f(x_i^{(t)}) + x_i^{(t)} \);
4. stop if \( \|x_i^{(t+1)} - x_i^{(t)}\| < \epsilon \) or \( \|dF(x_i^{(t+1)})\| < \epsilon \).

The second algorithm is based on the idea of simulated annealing and posterior sampling from the unnormalized density \( \exp(-(f(x) - c)^2/\bar{a}) \), where \( \bar{a} \) is a small positive tuning parameter.

**Algorithm 2: Simulated-annealing Based Algorithm**

Set \( \tau > 0 \), some small positive \( \bar{a} \) and burn-in time \( B \). Select a collection of starting points \( \{x_1, ..., x_n\} \), compute \( f(x_i) \) and keep only those points for which \( f(x_i) > \tau \).

1. For \( i = 1, ..., n \), draw the posterior samples \( \{x_i^{(1)}, ..., x_i^{(B+1)}\} \) from \( \exp(-(f(x) - c)^2/\bar{a}) \).
2. Collect all points \( \{x_i^{(B+1)}\}_{i=1}^n \).
2.4.1 Poisson regression

In the simulation, we consider the following mean function

\[ f(x_1, x_2) = 4 \exp \left( -5(x_1 + 1/2)^2 - 5(x_2 - 1/3)^2 \right) + 5 \exp \left( -5(x_1 - 1/2)^2 - 5(x_2 + 1/3)^2 \right) \]

We generate i.i.d. data \( X_i \) uniformly on \([0, 1] \times [0, 1] \). Then generate \( Y_i \) independently according to Poisson distribution with mean \( f(X_i) \). The sample size is 2000. Figure 2.1 gives the contour of the mean function and the level curve \( \{ x : f(x) = 3 \} \).

First, we model \( \sqrt{f} = b^T \theta \) and use fourth-order B-splines functions, that is \( q_1 = q_2 = 4 \). We put independent Generalized Inverse Gaussian prior on each element of \( \theta \). Specifically, we let each element distributed as

\[ \theta_j \sim \theta_j^{1/2} \exp \left( - \left( \frac{\theta_j}{2} + \frac{1}{2\theta_j} \right) \right). \]

The smoothing parameter \( (J_1, J_2) \) plays an important role as in any nonparametric estimation problem. They can be chosen by different methods. Here, the smoothing parameter \( (J_1, J_2) \) are chosen using the AIC which suggests \( J_1 = 8, J_2 = 6 \). Alternatively, they can be chosen using posterior mode of \( (J_1, J_2) \) (see for instance the nonparametric regression in the next section). We experimented with different smoothing values. To estimate the level curves, we use simulated-annealing based algorithm by setting \( \tau = \sqrt{2}, \bar{a} = 10^{-5} \) and \( B = 200 \). Figure 2.2 gives some results. In this figure, the contour of the posterior mean of the function \( f = (b^T \theta)^2 \) is also displayed.
Figure 2.2: Poisson regression $\sqrt{f} = b^T\theta$ (effects of the smoothing parameter $J_1$ and $J_2$). The orange circle is the truth, blue curve is estimated level curve induced by posterior mean. Top left: $J_1 = J_2 = 6$; top right: $J_1 = 8, J_2 = 6$; bottom left: $J_1 = J_2 = 10$; bottom right: $J_1 = J_2 = 12$.

To compare, we also put a prior directly on the functions using the B-spline series, that is $f = b^T\theta$. We also use fourth-order B-splines functions. We put independent gamma distribution (shape=1.5, rate=.5) prior on each element of $\theta$. The AIC suggests that the smoothing parameter ($J_1, J_2$) can be chosen to be $J_1 = 7, J_2 = 8$. We experimented with difference smoothing values. We use the same values of $\hat{a}$ and $B$ as before but $\tau = 2$. Some results are given in Figure 2.3.

For both models, the results seem to suggest that using the values for smoothing parameters suggested by the AIC can give fairly good results for the estimated level curves. Choosing smaller values for smoothing parameters (oversmoothing) seems more detrimental than choosing larger values (undersmoothing). From the top left graph in Figure 2.3 ($f = b^T\theta, J_1 = J_2 = 6$), the small level curve is completely missed and the approximation to the big level curve seems off quite a bit. At any rate, modeling in the two different ways does not seem to give much different outcomes and this is consistent with what the theory suggests.
The true level curves consist of two circles, as given in Figure 2.4. In the simulation, we consider the following function

\[ f(X_1, X_2) = 1 + \left( g\left( \sqrt{x_1^2 + x_2^2} \right) \right)^{1 + \cos^2(\tan^{-1}(x_2/x_1))}, \]

where \( g \) is the normal density function with mean 0.5 and standard deviation 0.3. We generate i.i.d. data \( X_i \) uniformly on \([0, 1] \times [0, 1] \) and i.i.d. \( \varepsilon_i \) from normal with mean 0 and standard deviation 0.1 and then set \( Y_i = f(X_{i,1}, X_{i,2}) + \varepsilon_i \). The level curve of interest is \( \{ x : f(x) = 2.1 \} \). The true level curves consist of two circles, as given in Figure 2.4.

The sample size is 2000. We use fifth-order B-splines functions, that is \( q_1 = q_2 = 5 \). One can choose the pair \( (J_1, J_2) \) by their posterior mode (in a logarithmic scale) by maximizing the following

\[ \log \Pi(J_1, J_2 | D_n, \sigma^2) = -2n \log \hat{\sigma} - \log(\det(BA_0B^T) + I_n) + \text{const}. \]
We give the results using a simulated-annealing based algorithm. We choose $\tau = 0.5$ and $\bar{a} = 10^{-5}$. The turning parameters $(J_1, J_2) = (9, 9)$ is chosen throughout in pilot experiments.

We experimented with different choices of $J_1$ and $J_2$. Figure 2.5 shows that the posterior mean under different smoothing levels. When $J_1 = J_2 = 7$, just slightly smaller than the 9, it can be seen that the posterior mean completely fails to approximate the inner circle. The approximation seems quite reasonable with larger $J$.

Figure 2.6 gives uncertainty quantification for two cases $J_1 = J_2 = 7$ and $J_1 = J_2 = 9$. We choose $\gamma = 0.1$ and $\rho = 1.2$. This choice of $\rho$ should give sufficiently large credibility but not too high. This can be done by some pilot simulation using the posterior samples. To evaluate the $1 - \gamma$ quantile $R_{n,\gamma}$, we first draw 200 posterior samples of $\theta$ to compute their posterior mean $\tilde{\theta}$. Then we compute $\sup_x |b_{J_1,J_2}(x)^T(\theta - \tilde{\theta})|$ by searching on a crude grid and pick the maximum point on the grid and then starting from this maximum point apply gradient ascent or descent method to check if nearby points can achieve greater (absolute) value. We keep the largest value as the supremum. The $(1 - \gamma)$-empirical quantile over all these suprema gives our estimate of $R_{n,\gamma}$. Finally we draw 100 posterior level curves and keep those that fall in the set $C_L$.

Now we assess the credibility and coverage performance over 100 iterations. Some experiments suggests that $\rho = 1.2$ is a reasonable choice, giving 97.13% credibility averaging over all iterations. To evaluate the coverage performance, we compute the Hausdorff distance between $L^*$ and $\tilde{C}_L$. Simulation shows that $L^*$ belongs to $\tilde{C}_L$ for 88%, 90%, 94%, 96%, 98% time when $C$ takes value 0.21, 0.22, 0.23, 0.24 and 0.25 respectively. Inspecting the proof of Lemma 2.2.1, the constant $A$ may be estimated by the $1/\inf_{x \in \{f = c\}} ||\nabla \tilde{f}(x)||$. When we compute with $C = 6A$ in this way, the frequentist coverage turns out to be 100% as the theory predicts.

Figure 2.4: The function $f$ and its level curves.
Figure 2.5: Level curves of a nonparametric regression function (effects of the smoothing parameter $J_1$ and $J_2$). The orange circle is the truth, the blue curve is the estimated level curve induced by the posterior mean. Top left: $J_1 = J_2 = 7$; top right: $J_1 = J_2 = 9$; bottom left: $J_1 = J_2 = 11$; bottom right: $J_1 = J_2 = 15$.

Figure 2.6: Level curves of a nonparametric regression function (uncertainty Quantification). Left: $J_1 = J_2 = 7$. Right: $J_1 = J_2 = 9$. 

36
2.5 Technical Proofs

This section includes the proofs of main results.

**Proof of Lemma 2.2.1.** Let \( x \) be such that \( |f(x) - \hat{c}| < \delta_1 \), take \( \hat{c} \) between \( c - \epsilon_1 \) and \( c + \epsilon_1 \). Consider \( \tilde{y} \) the closest projection of \( x \) onto \{ \( f = \hat{c} \) \} (just take one if not unique), therefore \( \tilde{y} \in \{ f = \hat{c} \} \) and \( d(x|\{ f = \hat{c} \}) = \|\tilde{y} - x\| \). It must be that \( x = \tilde{y} + \tilde{\ell} \nabla f(\tilde{y}) \) for some \( \tilde{\ell} \) and \( \|\nabla f(\tilde{y})\| > c_0 \). Without loss of generality, take \( \delta_1 \) to be a small value. By Taylor’s expansion,

\[
\left| f(x) - f(\tilde{y}) \right| = \left| (x - \tilde{y})^T \nabla f(\tilde{y}) + o(x - \tilde{y}) \right| = \left| \|x - \tilde{y}\|^2 / \tilde{\ell} + o(\|x - \tilde{y}\|) \right| = \|x - \tilde{y}\| \|\nabla f(\tilde{y})\| / \tilde{\ell} + o(1) \right| = \|x - \tilde{y}\| \|\nabla f(\tilde{y})\| / \tilde{\ell} + o(1).
\]

Therefore, \( \|x - \tilde{y}\| < A|f(x) - f(\tilde{y})| \) for some \( A \), since \( \|\nabla f(\tilde{y})\| > c_0 > 0 \). □

2.5.1 Proofs of the results for the Gaussian white noise model

**Proof of Proposition 2.3.1.** We show the proof in three parts.

1. \( \mathbb{E}\|f - \tilde{f}\|_\infty \lesssim n^{-\alpha/(2\alpha+2)} \log n; \)
2. \( \mathbb{E}\|\tilde{f} - E\tilde{f}\|_\infty \lesssim n^{-\alpha/(2\alpha+2)} \log n; \)
3. \( \|E\tilde{f} - f^\star\|_\infty \lesssim n^{-\alpha/(2\alpha+2)}. \)

This then will give the desired claim using Markov’s inequality. Throughout the proof, given some non-integer real number \( m \), we may slightly abuse notation by using \( \sum_{j=1}^m \) for \( \sum_{j=1}^m \) where \( \lfloor m \rfloor \) is the smallest integer greater than \( m \); and \( \sum_{j=m+1}^\infty \) for \( \sum_{j=\lfloor m \rfloor + 1}^\infty \) where \( \lfloor m \rfloor \) is the smallest integer greater than \( m \). Let \( n_\alpha = n^{1/(\alpha-1)} \). We may also write \( \sum_j \) for \( \sum_{j_1} \sum_{j_2} \). To avoid heavy notation, whenever the range for the indices of the summation is omitted, it should be viewed as taking over all permissible values in the context. For instance, when \( j_1, j_2 \in \mathbb{N}, \sum_{j_1, j_2 < n_\alpha} \) means \( \sum \sum_{j_1, j_2 < \lfloor n_\alpha \rfloor, 1 \leq j_1, j_2} \).

Part (1). Let \( Z = f - \tilde{f} \). Here is the outline of proof. We first prove that \( \|\mathbb{E}(Z^2(\cdot)|\mathcal{D}_n)\|_\infty \lesssim n^{-\alpha/(1+\alpha)} \log n \). We will then apply Lemma A.11 of Yoo and Ghosal (2016) to bound \( \mathbb{E}(\|Z\|_\infty^2|\mathcal{D}_n) \) by \( \log n \times \|\mathbb{E}(Z^2(\cdot)|\mathcal{D}_n)\|_\infty \). Thus \( \mathbb{E}(\|Z\|_\infty) = \mathbb{E}(\|Z\|_\infty|\mathcal{D}_n) \leq \sqrt{\mathbb{E}(\|Z\|_\infty^2|\mathcal{D}_n)} \) gives the claimed rate.

Note that \( Z = \sum_j \sum_k \phi_{jk} (\theta_{jk} - \hat{\theta}_{jk}) \). The posterior of \( Z \) given data is the centered Gaussian process and \( \mathbb{E}(Z^2(x)|\mathcal{D}_n) = \sum_j \sum_k h_j/(n \mu_j + 1) \phi_{jk}^2 \). By the uniform boundedness of the basis
functions, \( E(Z^2(x)|D_n) \) is bounded by

\[
n + \sum_{j_2=1}^{\infty} \sum_{n + j_2^{\alpha+1}} \frac{1}{n + j_2^{\alpha+1}} + \sum_{j_1=1}^{\infty} \frac{1}{n + j_1^{\alpha+1}} + \sum_{j_1=1}^{\infty} \sum_{j_2=1}^{\infty} \frac{1}{n + (j_1j_2)^{\alpha+1}}.
\]

The second term can be bounded by

\[
\sum_{j_1=1}^{n_\alpha} \frac{1}{n + (j_1j_2)^{\alpha+1}} \leq \sum_{j_1=1}^{n_\alpha} j_1^{-1(\alpha)} \sum_{j_2=1}^{\infty} j_2^{-1(\alpha)} \leq n^{-\alpha/(1+\alpha)}.
\]

The third term can be bounded in a similar way. For the fourth term, we split into two cases. One is where \( j_1j_2 \leq n_\alpha \) and the other \( j_1j_2 > n_\alpha \). For the first case,

\[
\sum_{j_1j_2 \leq n_\alpha} \frac{1}{n + (j_1j_2)^{\alpha+1}} \leq \frac{1}{n} \# \{ (j_1, j_2) : j_1, j_2 \leq n_\alpha, j_1, j_2 \in \mathbb{N} \}.
\]

For any fixed \( j_1 \) so that \( 1 \leq j_1 \leq n_\alpha \), the number of \( j_2 \) such that \( j_2 \leq n_\alpha/j_1 \) is bounded by \( n_\alpha/j_1 \). Therefore, \( n^{-1} \# \{ (j_1, j_2) : j_1, j_2 \leq n_\alpha, j_1, j_2 \in \mathbb{N} \} \leq n^{-1} n_\alpha \sum_{j_1=1}^{n_\alpha} j_1^{-1} \leq n^{-\alpha/(1+\alpha)} \log n \).

For the second case, we can have two scenarios: (i) when \( j_1 \geq n_\alpha, j_2 \geq 1 \),

\[
\sum_{j_1j_2 > n_\alpha} \frac{1}{n + (j_1j_2)^{\alpha+1}} \leq \sum_{j_1=1}^{n_\alpha} j_1^{-1(\alpha)} \sum_{j_2=1}^{\infty} j_2^{-1(\alpha)} = n^{-\alpha/(1+\alpha)},
\]

while for (ii) \( j_1 < n_\alpha, j_2 \geq n_\alpha/j_1 \),

\[
\sum_{j_1=1}^{n_\alpha} \sum_{j_2 > n_\alpha/j_1} \frac{1}{(j_1j_2)^{1+\alpha}} = \sum_{j_1=1}^{n_\alpha} j_1^{-1(\alpha)} \sum_{j_2 = n_\alpha/j_1}^{\infty} j_2^{-1(\alpha)} = \sum_{j_1=1}^{n_\alpha} j_1^{-2(\alpha)} \left( \frac{n_\alpha}{j_1} \right)^{-\alpha},
\]

which is equal to \( n^{-\alpha/(1+\alpha)} \sum_{j_1=1}^{n_\alpha} 1/j_1 \) thus bounded by \( n^{-\alpha/(1+\alpha)} \log n \). This then completes the proof of \( \|E(Z^2(x)|D_n)\|_\infty \leq n^{-\alpha/(1+\alpha)} \log n \).

We next will show that \( E(|Z(x) - Z(y)|^2|D_n) \) is bounded by some power of \( n \) multiplied by \( \|x - y\|^2 \). Note that

\[
E(|Z(x) - Z(y)|^2|D_n) = \sum_{j_1=0}^{\infty} \sum_{j_2=0}^{\infty} \sum_{k \in \mathcal{K}(j)} \frac{1}{n + (\mu_{j_1} \mu_{j_2})^{-1}} |\phi_{j_1j_2k}(x) - \phi_{j_1j_2k}(y)|^2.
\]

Clearly, when \( j_1 = j_2 = 0 \), the summand is bounded by \( n^{-1} \). Consider the summation over \( j_1 = 0, j_2 \in \mathbb{N} \). Using the fact that \( |\phi_{0j_2k}(x) - \phi_{0j_2k}(y)| \lesssim |x_2 - y_2| j_2^2 \), the summation is
bounded by
\[
\sum_{j_2=1}^{\infty} \frac{1}{n + j_2^{2+\alpha}} |\phi_{0j_2k}(x) - \phi_{0j_2k}(y)|^2 \lesssim \sum_{j_2=1}^{\infty} \frac{j_2^2}{n + j_2^{2+\alpha}} |x_2 - y_2|^2
\]
\[
\lesssim |x_2 - y_2|^2 \left( \sum_{j_2=1}^{\infty} \frac{j_2^2}{n} + \sum_{j_2>n_\alpha}^{\infty} \frac{j_2^2}{j_2^{2+\alpha}} \right)
\]
\[
\lesssim |x_2 - y_2|^2 \left( n^{(1-\alpha)/(1+\alpha)} + n^{-\alpha} \right)
\]
\[
\lesssim |x_2 - y_2|^2 n^{(1-\alpha)/(1+\alpha)}.
\]

Notice that for the third line above to hold, \( \alpha \) is required to be greater than 2. Bound of the same order can be obtained by the summation over \( j_1 \in \mathbb{N}, j_2 = 0. \) Consider the summation over \( j_1, j_2 \in \mathbb{N}. \) Using \( |\prod_{i=1}^{n} a_i - \prod_{i=1}^{n} b_i| \leq \sum_{i=1}^{n} |a_i - b_i| \) for \( |a_i|, |b_i| \leq 1, \) we have

\[
|\phi_{j_1j_2k}(x) - \phi_{j_1j_2k}(y)| = |\phi_{j_1k_1}(x_1)\phi_{j_2k_2}(x_2) - \phi_{j_1k_1}(y_1)\phi_{j_2k_2}(y_2)|
\]
\[
= 2 \sum_{i=1}^{2} \left| \frac{1}{\sqrt{2}} \phi_{j_1k_1}(x_i) - \frac{1}{\sqrt{2}} \phi_{j_1k_1}(y_i) \right|
\]
\[
\lesssim \sum_{i=1}^{2} j_i |x_i - y_i|.
\]

Therefore,

\[
|\phi_{j_1j_2k}(x) - \phi_{j_1j_2k}(y)|^2 \lesssim \left( \sum_{i=1}^{2} j_i |x_i - y_i| \right)^2 \leq (j_1^2 + j_2^2) \|x - y\|^2
\]

by the Cauchy-Schwartz’s inequality. Hence

\[
\sum_{j_1=1}^{\infty} \sum_{j_2=1}^{\infty} \sum_{k} \frac{1}{n + (\mu_{j_1\mu_{j_2}} - 1)} |\phi_{j_1j_2k}(x) - \phi_{j_1j_2k}(y)|^2 \lesssim \|x - y\|^2 \sum_{j_1=1}^{\infty} \sum_{j_2=1}^{\infty} \frac{j_1^2 + j_2^2}{n + (j_1j_2)^{1+\alpha}}.
\]

As before, we consider two cases for the summation. One is the summation over \( j_1, j_2 \leq n_\alpha \) and the other is over \( j_1, j_2 > n_\alpha. \) For the first case,

\[
\sum_{j_1, j_2 \leq n_\alpha} \frac{j_1^2 + j_2^2}{n + (j_1j_2)^{\alpha+1}} \leq \frac{1}{n} \sum_{j_1, j_2 \leq n_\alpha} (j_1^2 + j_2^2) \leq \frac{1}{n} \sum_{j_1, j_2 \leq n_\alpha} n_\alpha \sum_{j_2 \leq n_\alpha/j_1} (j_1^2 + j_2^2).
\]

By a straightforward calculation, it is bounded by \( n^{(2-\alpha)/(1+\alpha)}. \) For the second case, we again consider two scenarios (i) \( j_1 \geq n_\alpha, j_2 \geq 1 \) and (ii) \( j_2 < n_\alpha, j_2 \geq n_\alpha/j_1. \) Proceeding similarly as
Therefore, we can get the bound $n^{(2-\alpha)/(1+\alpha)}$. In summary,

$$E(\|Z(x) - Z(y)\|^2 | D_n) \lesssim n^{(2-\alpha)/(1+\alpha)} \|x - y\|^2.$$  

Now, by Lemma A.11 of Yoo and Ghosal (2016) with $\delta_n = n^{-p}$ for $p > 0$ sufficiently large, we obtain $E(\|Z\|_\infty^2 | D_n) \lesssim \log n \times \|E(Z^2(\cdot) | D_n)\|_\infty$. By the remark in the beginning of the proof, these implies $E\|Z\|_\infty \lesssim n^{-\alpha/(2\alpha+2)} \log n$.

Part (2). Let $V = \tilde{f} - E\tilde{f}$.

$$V = \sum_{j_1=0}^{\infty} \sum_{j_2=0}^{\infty} \sum_{k \in K(j)} \frac{\sqrt{n\varepsilon_{jk}}}{n + (\mu_{j_1,j_2})^{-1}} \phi_{jk}$$

is also a mean zero Gaussian process. The same argument as for $Z$ is used. We omit the details. But one can establish that $\|E(V^2(\cdot))\|_\infty \lesssim n^{-\alpha/(1+\alpha)} \log n$ and $E(\|V(x) - V(y)\|^2) \lesssim n^{(2-\alpha)/(1+\alpha)} \|x - y\|^2$. Therefore, $E\|V\|_\infty \lesssim n^{-\alpha/(2\alpha+2)} \log n$.

Part (3). Note that

$$E(\tilde{f}) - f^* = \sum_{j_1=0}^{\infty} \sum_{j_2=0}^{\infty} \sum_{k \in K(j)} \left( \frac{n}{n + \mu_{j_1,j_2}} - 1 \right) \phi_{jk} \theta_{0,jk}.$$

The summand of $j_1 = j_2 = 0$ is bounded by $n^{-1}$. Considering the summation over $j_1, j_2 \in \mathbb{N}$,

$$\left\| \sum_{j_2 \geq 1} \sum_{k \in K(j)} \left( \frac{n}{n + \mu_{j_1,j_2}} - 1 \right) \phi_{jk} \theta_{0,jk} \right\|_\infty \leq \left\| \sum_{j_2 \geq 1} \sum_{k \in K(0,j_2)} \left( \frac{j_2^{1+\alpha}}{n + j_2^{1+\alpha}} \right) \phi_{jk} \theta_{0,jk} \right\|_\infty$$

$$+ \left\| \sum_{j_2 > n \alpha} \sum_{k \in K(0,j_2)} \left( \frac{n}{n + j_2^{1+\alpha}} - 1 \right) \phi_{jk} \theta_{0,jk} \right\|_\infty$$

where the first term of the right hand side can be bounded as

$$\left\| \sum_{j_2 \geq 1} \sum_{k \in K(0,j_2)} \frac{j_2^{0}}{n \alpha} \phi_{jk} \theta_{0,jk} \right\|_\infty \leq \frac{n \alpha}{n} \left( \sum_{j_2 \geq 1} \sum_{k \in K(0,j_2)} \frac{j_2^{0}}{n \alpha} \right) \theta_{0,jk} \lesssim n^{-\alpha}/(1+\alpha),$$

by the uniform boundedness of basis functions and the assumption on the smoothness of $f^*$ (take $i_1 = 0, i_2 = \alpha$). Since for $j_2 > n \alpha$, $\sum_{j_2 > n \alpha} \sum_{k \in K(0,j_2)} |\theta_{0,jk}| \lesssim n^{-\alpha}$. The second term of the right hand side of above display can be bounded as $\sum_{j_2 > n \alpha} \sum_{k \in K(0,j_2)} |\theta_{0,jk}| \lesssim n^{-\alpha} = n^{-\alpha}/(\alpha+1)$.

Therefore $\left\| \sum_{j_2 \geq 1} \sum_{k \in K(0,j_2)} \left( \frac{n}{n + j_2^{1+\alpha}} - 1 \right) \phi_{jk} \theta_{0,jk} \right\|_\infty \lesssim n^{-\alpha}/(\alpha+1)$.

Finally, consider the sum over $j_1, j_2 \in \mathbb{N}$. We consider two scenarios (i) $j_1j_2 > n \alpha = n^{1/(\alpha+1)}$ and (ii) $j_1j_2 \leq n \alpha$. For the first scenario, by the uniform boundedness of basis func-
We shall bound all the three terms $E\| E\sum_{j_1j_2>n_n} \sum_{k \in K(j)} \phi_{jk} \theta_{0,jk} \|$ is of the order $n^{-\alpha/2}$. It is easy to see that

$$\left\| \sum_{j_1j_2>n_n} \sum_{k \in K(j)} \left( \frac{n}{n + \mu_j^{-1}} - 1 \right) \phi_{jk} \theta_{0,jk} \right\|_\infty$$

is of the order $n^{-\alpha/2} = n^{-\alpha/(2(1+\alpha))}$. In the second scenario, we bound the term as

$$\left\| \sum_{j_1j_2 \leq n_n} \sum_{k \in K(j)} \frac{(j_1j_2)^{1+\alpha/2}}{n + (j_1j_2)^{1+\alpha/2}} \phi_{jk} \theta_{0,jk} \right\|_\infty \lesssim \frac{n^{1+\alpha/2}}{n} \left\| \sum_{j_1j_2 \leq n_n} \sum_{k \in K(j)} \frac{(j_1j_2)^{\alpha/2}}{n^{1+\alpha/2}} \phi_{jk} \theta_{0,jk} \right\|_\infty,$$  

which is of the order $n^{-\alpha/(2+2\alpha)}$ by the uniform boundedness of basis functions and the assumption on the smoothness of $f^*$.

In summary, $\|E\tilde{f} - f^*\|_\infty \lesssim n^{-\alpha/(2+2\alpha)}$ as claimed. \(\blacksquare\)

**Proof of Proposition 2.3.3.** Let $E_{f^*}$ denotes the expectation with respect to variables $\varepsilon_k$, $\varepsilon_{l,i,k}$, and $E_\Pi$ with respect to variables $\bar{\varepsilon}_k$, $\bar{\varepsilon}_{l,i,k}$.

We shall prove that $E_{f^*} - E_{\Pi}(\|f - f^*\|_\infty \|\mathcal{D}_n\|) \lesssim \varepsilon_n = (\log n/n)^{\alpha/2\alpha}$, then the result of the theorem is obtained by invoking Markov’s inequality. Throughout we let $E$ denote the expectation with respect to all $\varepsilon_k$, $\varepsilon_{l,i,k}$, $\bar{\varepsilon}_k$, $\bar{\varepsilon}_{l,i,k}$. Let $\tilde{f}$ denotes the posterior mean of $f$. Write $f - f^* = (f - \bar{f}) + (\tilde{f} - E\bar{f}) + (E\tilde{f} - f^*)$. Choose some $J_n$ such that $2^{J_n} \propto (n/\log n)^{1/(2\alpha+d)}$.

We shall bound all the three terms $E\|f - \bar{f}\|_\infty$, $E\|\tilde{f} - E\bar{f}\|_\infty$, and $\|E\tilde{f} - f^*\|_\infty$.

Firstly,

$$f - \bar{f} = \sum_{k \in K(J_0)} \left( \frac{1}{n+1} \right)^{1/2} \varepsilon_k \Phi_k + \sum_{\ell = J_0}^\infty \sum_{i \in I, k \in K(\ell)} \left( \frac{\mu_i}{n\mu_i + 1} \right)^{1/2} \varepsilon_{l,i,k} \Psi^i_{l,k},$$

One can show that,

$$E\left\| \sum_{k \in K(J_0)} \frac{1}{(n+1)^{1/2}} \varepsilon_k \Phi_k \right\|_\infty \leq \frac{1}{\sqrt{n}} \left( \left\| \sum_{k \in K(J_0)} \Phi_k \right\|_\infty \right) E \left( \max_{k \in K(J_0)} |\varepsilon_k| \right) \lesssim \frac{1}{\sqrt{n}},$$

and in view of Lemma A.0.1,

$$E\left\| \sum_{\ell = J_0}^\infty \sum_{i \in I, k \in K(\ell)} \left( \frac{\mu_i}{n\mu_i + 1} \right)^{1/2} \varepsilon_{l,i,k} \Psi^i_{l,k} \right\|_\infty.$$
can be bounded as

\[
\sum \sum \left( \left\| \sum_{k \in K(l)} |\Psi_{l,k}^{i}| \right\|_\infty \mathbb{E} \left( \max_{k \in K(l)} |\tilde{\epsilon}_{l,i,k}| \right) \left( \frac{\mu_l}{n\mu_l + 1} \right)^{1/2} \right)
\]

\[
\lesssim \sum_l (2^d - 1)2^{ld/2} \sqrt{\log(2^{ld})} \left( \frac{\mu_l}{n\mu_l + 1} \right)^{1/2}
\]

\[
\lesssim \sum_{l=J_0}^{J_n} \sqrt{2^{ld}d \left( \frac{\mu_l}{n\mu_l + 1} \right)^{1/2}} + \sum_{l > J_n} \sqrt{2^{ld}d} \left( \frac{\mu_l}{n\mu_l + 1} \right)^{1/2}
\]

\[
\lesssim \sum_{l=J_0}^{J_n} \frac{\sqrt{2^{ld}d}}{n} + \sum_{l > J_n} \sqrt{2^{ld}d} \mu_l,
\]

which is bounded by \(\epsilon_n\).

Next,

\[
\tilde{f} - E\tilde{f} = \sum_{k \in K(J_0)} \frac{n}{\sqrt{n(1 + n)}} \varepsilon_k \Phi_k + \sum_{l=J_0}^{\infty} \sum_{i \in I, k \in K(l)} \frac{n\mu_l}{\sqrt{n(n\mu_l + 1)}} \varepsilon_{l,i,k} \Psi_{l,k}^{i}.
\]

Arguing similarly, it is not hard to see that for the first term, \(E \left| \sum_{k \in K(J_0)} \frac{n}{\sqrt{n(1 + n)}} \varepsilon_k \Phi_k \right|_\infty \lesssim n^{-1/2}\). For the second term,

\[
E \left| \sum_{l=J_0}^{\infty} \sum_{i \in I, k \in K(l)} \frac{n\mu_l}{\sqrt{n(n\mu_l + 1)}} \varepsilon_{l,i,k} \Psi_{l,k}^{i} \right|_\infty \lesssim \sum_{l=J_0}^{J_n} \sqrt{2^{ld}d} \frac{n\mu_l}{\sqrt{n(n\mu_l + 1)}} + \sum_{l > J_n} \sqrt{2^{ld}d} \frac{n\mu_l}{\sqrt{n(n\mu_l + 1)}}
\]

\[
\lesssim \sum_{l=J_0}^{J_n} \frac{\sqrt{2^{ld}d}}{n} + \sum_{l > J_n} \sqrt{2^{ld}d} \mu_l.
\]

The first term of the last expression is bounded by \(\epsilon_n\). Since for \(l \geq J_n\), \(\mu_l \lesssim n^{-1}\), the second term is of smaller order than \(\sum_{l=J_n}^{\infty} \sqrt{2^{ld}d} \mu_l\) which is bounded by \(\epsilon_n\).

Turning lastly to

\[
E\tilde{f} - f^* = \sum_{k \in K(J_0)} \frac{-1}{(1 + n)} \langle f^*, \Phi_k \rangle \Phi_k + \sum_{l=J_0}^{\infty} \sum_{i \in I, k \in K(l)} \frac{-1}{\left( n\mu_l + 1 \right)} \langle f^*, \Psi_{l,k}^{i} \rangle \Psi_{l,k}^{i},
\]
it is easy to see that \( \| \sum_{k \in K(J_0)} \frac{1}{(1+n)} \left( f^*, \Phi_k \right) \Phi_k \|_\infty \lesssim n^{-1} \). For the second term,

\[
\left\| \sum_{\ell=J_0}^{\infty} \sum_{i \in I, k \in K(\ell)} \frac{-1}{(n \mu_\ell + 1)} \left( f^*, \Psi_{i,k}^\ell \right) \Psi_{i,k}^\ell \right\|_\infty \leq \sum_{\ell=J_0}^{\infty} \frac{2^{-(\alpha + \frac{d}{2})}}{n \mu_\ell + 1} \sum_{i} \sum_{k} \| \Psi_{i,k}^\ell \|_\infty \\
\lesssim \sum_{\ell=J_0}^{\infty} (2^d - 1)2^{d/2} \frac{2^{-(\alpha + \frac{d}{2})}}{n \mu_\ell + 1} \\
\lesssim \sum_{\ell=J_0}^{J_n} \frac{2^{d-\alpha}}{n \mu_\ell} + \sum_{\ell>J_n}^{\infty} 2^{d-\alpha},
\]

which can be shown to be bounded by \( \epsilon_n \). This then completes the proof.

2.5.2 Proofs of the results for binary regression, Poisson regression and density estimation

In our proofs for binary regression, Poisson regression and density estimation, we shall make use of some master theorems in Appendix B to obtain some intermediate results. Recall the definition of the Hellinger distance between \( p \) and \( q \) as \( \left( \int (\sqrt{p} - \sqrt{q})^2 d\mu \right)^{1/2} \). Let the metric \( d \) to be the \( L^\infty \)-metric and let \( d_n \) denote the metric with respect to which the contraction rates will be available in the master theorem. We may interpret \( d_n(f_1, f_2) \) as \( d_H(p_{f_1}, p_{f_2}) \) when the data are i.i.d (say the random design case), or as \( d_{H,n} := \sqrt{\frac{1}{n} \sum_{i=1}^{n} d_H^2(p_{f_i,1}, p_{f_i,2})} \) the root-average squared Hellinger distance when the data are independently (but not identically) distributed (for the fixed design case). Note that for the fixed design case, the probability distribution will only refer to that of \( Y_i \) as indexed through \( f(X_i) \).

For example, in binary regression, the Hellinger distance \( d_H = d_{H,G} \) on the law of one copy \( (Y_i, X_i) \) is given by

\[
d_H(p_f, p_g) := d_{H,G}(p_f, p_g) = \sqrt{\int \sum_{y=0}^{1} \left( \sqrt{p_f(y|x)} - \sqrt{p_g(y|x)} \right)^2 \d G(x)}.
\]

Such a distance is defined similarly in Poisson regression. The root average square Hellinger distance \( d_{H,n} \) is similarly defined but with \( G \) replaced by \( G_n \), that is, \( d_{H,n}(p_f, p_g) := d_{H,G_n}(p_f, p_g) \). Let \( \mathcal{F}_n := \{ \theta^T \psi, \theta \in \mathbb{R}^{J_n}, \theta \in S(M_n)^{J_n} \} \).

**Proof of Proposition 2.3.7.** Note first that because of the tensor product presentation, \( \prod_{k=1}^{d} J_k \) described here is equivalent to the term \( J \) in Theorem B.0.1. We let \( J_k \asymp J_n^* \) for \( k = 1, ..., d \).
We first consider the case where \( X \) is the fixed design. When \( d_n \) is the root-average squared Hellinger metric \( d_{H,n}(p_f,p_{f^*}) \), the general theory guarantees that Assumption 1.6 holds. We first apply Theorem B.0.1 with \( d \) the supremum distance and the sieve \( \mathcal{F}_n := \{ \theta^T b : \theta \in (0,1)^{k \cdot J_k} \} \). We specify \( \sqrt{\mathcal{T}} = \theta^T b \). In view of Lemma 2.8 in Ghosal and van der Vaart (2017), we have that for any pair of \( f_1^{1/2}, f_2^{1/2} \in \mathcal{F}_n \),

\[
d_{H,n}(p_{f_1}, p_{f_2}) \leq \left[ \sup \left( (\sqrt{f_1} - \sqrt{f_2})^2 + (\sqrt{1 - f_1} - \sqrt{1 - f_2})^2 \right) \right]^{1/2}.
\]

Note also that due to \( \|\sqrt{f} - \sqrt{1}\|_\infty \leq 2 \|\sqrt{f} - \sqrt{1}\|_2 \), we have

\[
\|\sqrt{f_1} - \sqrt{f_2}\|_\infty \leq 2 \|\sqrt{f_1} - \sqrt{f_2}\|_2.
\]

Consider for every \( f \) such that \( \|\sqrt{\mathcal{T}} - \sqrt{f}\|_\infty \) sufficiently small, we also have

\[
K(p_f(n), p_f(n)) \lesssim \sum_{i=1}^n K(p_{f^*_i}, p_{f_i}) \lesssim n\|f - f^*\|_\infty^2 \lesssim n\|\sqrt{\mathcal{T}} - \sqrt{f}\|_\infty^2.
\]

Therefore, (B.5) and (B.6) hold. With \( J_n^* \asymp (n/\log n)^{1/(d+2\alpha)} \) one can obtain the contraction rate \( \eta_n = (n/\log n)^{-\alpha/(d+2\alpha)} \) relative to \( d_{H,n}(p_f,p_{f^*}) \). Note that

\[
\|\sqrt{\mathcal{T}} - \sqrt{f^*}\|_{2,G_n} = \left\| \frac{1}{n} \sum_{i=1}^n \left( \sqrt{f(X_i)} - \sqrt{f^*(X_i)} \right) \right\|_2.
\]

Since \( \|\sqrt{\mathcal{T}} - \sqrt{f^*}\|_{2,G_n} \leq d_{H,n}(p_f,p_{f^*}) \), the same rate \( \eta_n \) applies to \( \|\sqrt{\mathcal{T}} - \sqrt{f^*}\|_{2,G_n} \).

Now since \( \sqrt{\mathcal{T}} \in \mathcal{H}^\alpha([0,1]^d) \), by Lemma A.0.3, there exists some \( \theta^\infty \) such that \( \|((\theta^\infty)^T b - \sqrt{\mathcal{T}})\|_\infty \lesssim (J_n^*)^{-\alpha} \). Let \( \sqrt{\mathcal{T}}^\infty := (\theta^\infty)^T b \). We have

\[
\|\sqrt{\mathcal{T}} - \sqrt{f^*}\|_\infty \lesssim \|\sqrt{\mathcal{T}}^\infty - \sqrt{f^*}\|_\infty + \|\sqrt{\mathcal{T}}^\infty - \sqrt{\mathcal{T}}\|_\infty
\]

\[
\lesssim (J_n^*)^{-\alpha} + \|((\theta^\infty - \theta)^T b\|_\infty
\]

\[
\lesssim (J_n^*)^{-\alpha} + \|\theta^\infty - \theta\|_\infty
\]

\[
\lesssim (J_n^*)^{-\alpha} + \|\theta^\infty - \theta\|,
\]

44
where the third line follows because \( \| b \|_1 = \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} B_{j_1}(x) \cdots B_{j_d}(x) = 1 \). Now by Lemma A.0.4,

\[
\| \sqrt{f^\infty} - \sqrt{f} \|_{2,G_n}^2 = \| (\theta^\infty - \theta)^T b \|_{2,G_n}^2 = (\theta^\infty - \theta)^T B^T B (\theta^\infty - \theta) \asymp \left( \prod_{k=1}^d J_k^{-1} \right) \| \theta^\infty - \theta \|^2,
\]

thus \( \| \theta^\infty - \theta \| \asymp \left( \prod_{k=1}^d J_k^{1/2} \right) \| \sqrt{f} - \sqrt{f^\infty} \|_{2,G_n} \). Therefore

\[
\| \sqrt{f} - \sqrt{f^\infty} \|_\infty \lesssim (J_n^*)^{-\alpha} + (J_n^*)^{d/2}(\| \sqrt{f} - \sqrt{f^\infty} \|_{2,G_n} + \| \sqrt{f^*} - \sqrt{f^\infty} \|_{2,G_n}) \\
\lesssim (J_n^*)^{-\alpha} + (J_n^*)^{d/2}(\eta_n + (J_n^*)^{-\alpha}) \\
\lesssim (J_n^*)^{d/2}(\eta_n + (J_n^*)^{-\alpha}).
\]

With the choice of \( J_n^* \), the contraction rate for \( \| \sqrt{f} - \sqrt{f^\infty} \|_\infty \) is \( (n/\log n)^{(d-2\alpha)/(2(d+2\alpha))} \).

To consider the derivatives, let

\[
b^{(r)} = b_{J_1,\ldots,J_d,q_{d1},\ldots,q_{dJ_d}}^{(r)}(x) = (D^{r_1} B_{j_1,q_{j_1}}(x_1) \cdots D^{r_d} B_{j_d,q_{dJ_d}}(x_d) : 1 \leq j_k \leq J_k, k = 1, \ldots, d)^T.
\]

Using \( \| (\theta^\infty - \theta)^T b^{(r)} \|_\infty \lesssim \| \theta^\infty - \theta \|_\infty \prod_{k=1}^d J_k^{\tau_k} \), which is true as

\[
sup_x \sum_{j_1=1}^{J_1} |D^{r_1} B_{j_1}(x_1)| \cdots \sum_{j_d=1}^{J_d} |D^{r_d} B_{j_d}(x_d)| \lesssim \prod_{k=1}^d J_k^{\tau_k},
\]

we have

\[
\| D^r \sqrt{f} - D^r \sqrt{f^*} \|_\infty \lesssim \| D^r \sqrt{f} - D^r \sqrt{f^\infty} \|_\infty + \| D^r \sqrt{f^*} - D^r \sqrt{f^\infty} \|_\infty \\
\lesssim \| (\theta^\infty - \theta)^T b^{(r)} \|_\infty + \sum_{k} J_k^{\tau_k - \alpha} \\
\lesssim \prod_{k} J_k^{\tau_k} \| \theta^\infty - \theta \| + \sum_{k} J_k^{\tau_k - \alpha}.
\]

Using \( \| \theta^\infty - \theta \| \asymp \prod_{k=1}^d J_k^{1/2} \| \sqrt{f} - \sqrt{f^\infty} \|_{2,G_n} \) as previously,

\[
\| D^r \sqrt{f} - D^r \sqrt{f^*} \|_\infty = (\eta_n + (J_n^*)^{-\alpha})(J_n^*)^{d/2+|r|}.
\]

By the choice of \( J_n^* \), we obtain the rate as claimed.

For the case of random design, \( d_n \) is the Hellinger metric \( d_H(p_f,p_{f^*}) \). Applying Theorem B.0.1 we have same contraction rate for \( d_H(p_f,p_{f^*}) \) and hence for \( \| \sqrt{f} - \sqrt{f^*} \|_{2,G} \). The con-
traction rate for \( \| \sqrt{f} - \sqrt{f^*} \| \infty \) remains the same as the fixed design case by following similar argument as above and noting that
\[
\| \sqrt{f} - \sqrt{f^*} \|_{2, G}^2 = (\theta^\infty - \theta)^T E_G (bb^T) (\theta^\infty - \theta) = (\theta^\infty - \theta)^T \frac{B^T B}{n} (\theta^\infty - \theta) + o_p(\| \theta^\infty - \theta \|^2).
\]

\[\blacksquare\]

**Proof of Proposition 2.3.9.** We shall verify the conditions of Theorem B.0.2 for Poisson model for the fixed design with \( S(M_n) = [M_n^{-1}, M_n] \) in (P2). Because of the tensor product, \( \prod_{k=1}^d J_k \) described here is equivalent to the term \( J \) in Theorem B.0.2. We let \( J_k \sim J^*_n \) for \( k = 1, ..., d \). The space \( F_n = \{ \theta^T b : \theta \in \mathbb{R} \prod_{k=1}^d J_k, \theta \in [M_n^{-1}, M_n] \prod_{k=1}^d J_k \} \). Let
\[
e_n(\theta_1, \theta_2) := e_n(f^{1/2}_{\theta_1}, f^{1/2}_{\theta_2}) = \max_j \left| \frac{\theta_{1,j}}{\theta_{2,j}} - 1 \right| \vee \max_j \left| \frac{\theta_{2,j}}{\theta_{1,j}} - 1 \right|,
\]
where \( f^{1/2}_{\theta} = \theta^T b \in F_n \). It suffices to show the contraction rate is \( \epsilon_n = (n/ \log n)^{-\alpha/(d+2\alpha)} \) with respect to \( d_n(f^{1/2}_0, f^{1/2}_1) := \| \sqrt{f^0} - \sqrt{f^1} \|_{2, G_n} \). The sup-norm contraction rate then follows by the same argument as in the proof of Proposition 2.3.7. The argument to derive the rates for \( D^* \sqrt{f} \) is also similar to that proof.

Throughout the proof, the design points \( (X_i)_{i=1}^n \) are fixed and take \( \bar{\epsilon}_n = C \epsilon_n > 0 \) small for some constant \( C \). We set \( f_0 = f^* \) and let \( f_{0,i} = f_0(X_i) \) and \( f_{1,i} = f_1(X_i) \). Assume that for some fixed \( f_1 := f_{\theta_1} \) such that \( \sqrt{f_{\theta_1}} \in F_n \),
\[
d^2_n(f^{1/2}_0, f^{1/2}_1) := \frac{1}{n} \sum_{i=1}^n (\sqrt{f_{0,i}} - \sqrt{f_{1,i}})^2 > \bar{\epsilon}_n^2.
\]
Consider the Neyman-Pearson test for \( P_{f_0} \) versus \( P_{f_1} \) given by the rejection region
\[
\frac{\prod_{i=1}^n f_{Y_i} e^{-f_{1,i}}}{\prod_{i=1}^n f_{Y_i} e^{-f_{0,i}}} > 1.
\]
Let \( a_i = \log(f_{0,i}/f_{1,i}) \) and \( b = \sum_{i=1}^n (f_{0,i} - f_{1,i}) \). The rejection region can be written as
The type I error probability against $P_{f_0}$, by Markov’s inequality, is bounded as

$$P_{f_0} \left( \sum_{i=1}^{n} Y_i a_i < b \right) = P_{f_0} \left( - \frac{1}{2} \sum_{i=1}^{n} Y_i a_i > -\frac{b}{2} \right)$$

$$\leq e^{b \sqrt{2}} E_{f_1} \left[ e^{\sum_i (Y_i(-a_i)/2)} \right]$$

$$= e^{b / 2} \prod_{i=1}^{n} \exp \left( f_{1,i} (e^{-a_i/2} - 1) \right)$$

$$= \exp \left( \sum_{i=1}^{n} \left( f_{1,i}^{1/2} f_{1,i}^{1/2} - \frac{1}{2} f_{1,i} - \frac{1}{2} f_{0,i} \right) \right).$$

Noting that $\sqrt{ab} - (a + b)/2 = -(\sqrt{a} - \sqrt{b})^2 / 2$, the expression is equal to

$$\exp \left( \sum_{i=1}^{n} - \frac{(\sqrt{f_{0,i}} - \sqrt{f_{1,i}})^2}{2} \right) = \exp(-n d_n (f_0, f_1)^2 / 2) \leq \exp(-n \tilde{\epsilon}_n^2 / 2).$$

The same bound holds for Type II error probability $P_{f_1} (\sum_i Y_i a_i > b)$ by same argument when the roles of $f_0$ and $f_1$ are flipped.

Now consider the Type II error rate against $P_{f_2}$ for arbitrary $f_2 := f_{\theta_2}$ such that $\sqrt{f_{\theta_2}} \in \mathcal{F}_n$ with $e_n(\theta_1, \theta_2) \leq \delta_n$ which is assumed to be less than $\tilde{\epsilon}_n^2 / 16 M_n^2$. Applying the Cauchy-Schwarz inequality,

$$P_{f_2} \left( \sum_{i=1}^{n} Y_i a_i > b \right) = E_{f_2} \left[ \mathbb{1} \left( \sum_{i=1}^{n} Y_i a_i > b \right) \right]$$

$$= E_{f_1} \left[ \mathbb{1} \left( \sum_{i=1}^{n} Y_i a_i > b \right) \prod_{i=1}^{n} \left( \frac{f_{2,i}}{f_{1,i}} \right)^{Y_i} \exp \left( - \sum_{i=1}^{n} (f_{2,i} - f_{1,i}) \right) \right]^{1/2}$$

$$= \left\{ E_{f_1} \left[ \mathbb{1} \left( \sum_{i=1}^{n} Y_i a_i > b \right) \right] \right\}^{1/2}$$

$$\times \left\{ E_{f_1} \left[ \prod_{i=1}^{n} \left( \frac{f_{2,i}}{f_{1,i}} \right)^{2Y_i} \exp \left( - \sum_{i=1}^{n} (f_{2,i} - f_{1,i}) \right) \right] \right\}^{1/2}$$

Clearly,

$$\left\{ E_{f_1} \left[ \mathbb{1} \left( \sum_{i=1}^{n} Y_i a_i > b \right) \right] \right\}^{1/2} \leq \exp(-n \tilde{\epsilon}_n^2 / 4). \quad (2.12)$$

47
We then have
\[
\left\{ E_{f_1} \left[ \prod_{i=1}^{n} \left( \frac{f_{2,i}}{f_{1,i}} \right)^{2Y_i} \right] \right\}^{1/2} = \left\{ \prod_{i=1}^{n} E_{f_1} \left[ e^{2Y_i \log(f_{2,i}/f_{1,i})} \right] \right\}^{1/2}
\]
\[= \prod_{i=1}^{n} \exp \left[ \frac{1}{2} f_{1,i} \left( e^{2\log(f_{2,i}/f_{1,i})} - 1 \right) \right]
\]
\[= \exp \left( \sum_{i=1}^{n} \frac{1}{2} f_{1,i} \left( \frac{f_{2,i}^2}{f_{1,i}^2} - 1 \right) \right).
\]

Since \( f_1^{1/2} = \theta_1^T b \) and \( f_2^{1/2} = \theta_2^T b \), we get
\[
\left\| f_1^{1/2} \right\|_\infty \sqrt{\left\| f_2^{1/2} \right\|_\infty - 1} = \left\| \sum (\frac{\theta_{1,j}}{\theta_{2,j}} - 1) \theta_{2,j} B_j \right\|_\infty \sqrt{\left\| \frac{\sum \theta_{1,j} B_j}{\sum \theta_{2,j} B_j} \right\|_\infty} \leq e_n(\theta_1, \theta_2).
\]

We then have
\[
\left\| \frac{f_1}{f_2} - 1 \right\|_\infty \leq \left\| \frac{f_1^{1/2}}{f_2^{1/2}} - 1 \right\|_\infty \left\| \frac{f_1^{1/2}}{f_2^{1/2}} - 1 + 2 \right\|_\infty \leq \delta_n(2 + \delta_n) \lesssim \delta_n,
\]
\[
\left\| \frac{f_1}{f_2} - 1 \right\|_\infty \leq \left\| \frac{f_1}{f_2} - 1 \right\|_\infty \left\| \frac{f_1}{f_2} - 1 + 2 \right\|_\infty \leq \delta_n.
\]

Also, \( \sum_{i=1}^{n} f_{1,i}/n = \sum_{i=1}^{n} (\sqrt{f_i})^2 / n = \|\theta_1^T b\|_{2,G_n}^2 \). By Lemma A.0.4, we have
\[
\|\theta_1^T b\|_{2,G_n} = \|\theta_1\| \prod_{k=1}^{d} J_k^{-1/2}.
\]

Since \( \|\theta_1\| \leq \prod_{k=1}^{d} J_k^{1/2} \|\theta_1\|_\infty \), \( \sum_{i=1}^{n} f_{1,i} \leq n M_n^2 \). We then obtain the following
\[
\left\{ E_{f_1} \left[ \prod_{i=1}^{n} \left( \frac{f_{2,i}}{f_{1,i}} \right)^{2Y_i} \right] \right\}^{1/2} \lesssim \exp\{n\delta_n M_n^2\}. \tag{2.13}
\]

Note also that \( \|f_1 - f_2\|_\infty \leq \|f_1/f_2 - 1\|_\infty \|f_2\|_\infty = \|f_1/f_2 - 1\|_\infty \|(\theta_1^T b)^2\|_\infty \leq \delta_n \|\theta_2\|_\infty^2 \lesssim \delta_n M_n^2 \).

Thus
\[
\exp \left( - \sum_{i=1}^{n} (f_{2,i} - f_{1,i}) \right) \leq \exp(n\|f_1 - f_2\|_\infty) \leq \exp(n\delta_n M_n^2). \tag{2.14}
\]

Therefore, combining equations (2.12), (2.13) and (2.14), \( P_{f_2} \left( \sum_{i=1}^{n} Y_i a_i > b \right) \leq \exp(-n\epsilon_n^2/8) \).
We next verify that \((n\varepsilon_n^2/8) - \log N_n \to \infty\). Notice also that for all \(j\), then

\[|\theta_{1,j}/\theta_{2,j} - 1| = |\theta_{1,j} - \theta_{2,j}|/|\theta_{2,j}| < |\theta_{1,j} - \theta_{2,j}| M_n.\]

Likewise, \(|\theta_{2,j}/\theta_{1,j} - 1| < |\theta_{1,j} - \theta_{2,j}| M_n.\) Thus \(e_n(\theta_1, \theta_2) \leq M_n\|\theta_1 - \theta_2\|_\infty.\) Recall that \(\mathcal{F}_n = \{\theta^T b : \theta \in \mathbb{R}^{\prod_{k=1}^d J_k}, M_n^{-1} \leq \|\theta\|_\infty \leq M_n\}.\) Consider a sequence of \(B_{n,k}\) sets of the form \(\{\theta^T b \in \mathcal{F}_n : e_n(\theta^T b, \theta_k^T b) \leq \delta_n\}\) for some sequence of \(\theta_k\) such that \(\theta_k^T b \in \mathcal{F}_n \cap \{\theta^T b : d_n(\sqrt{\|T\}}, \theta^T b) > \tilde{\epsilon}_n\},\) for \(k = 1, 2, \ldots.\) The \(\theta_1\) in the previous discussion may be viewed as one of these \(\theta_k.\) Since

\[\mathcal{N}(\delta_n/M_n, \{\theta \in \mathbb{R}^{\prod_{k=1}^d J_k} : M_n^{-1} \leq \|\theta\|_\infty \leq M_n\}, \|\cdot\|_\infty \leq (3M_n^2/\delta_n)\prod_{k=1}^d J_k,\]

the minimal number of sets \(B_{n,k}\) needed to cover \(\mathcal{F}_n\) is less than \((3M_n^2/\delta_n)\prod_{k=1}^d J_k,\) so is the minimal number \(N_n\) of sets \(B_{n,k}\) needed to cover \(\mathcal{F}_n \cap \{\theta^T b : d_n(\sqrt{\|T\}}, \theta^T b) > \epsilon\}.\) Recall that \(\delta_n = \tilde{\epsilon}_n^2/16M_n^2.\) By choosing \(M_n\) to be some power of \(n\) and given the choices of \(J_n^*\) and \(\epsilon_n,\) with \(\tilde{\epsilon}_n = C\epsilon_n\) for some constant \(C > 0, (n\varepsilon_n^2/8) - \log N_n \to \infty\) can be guarantee to hold when \(C\) is chosen to be sufficiently large. Indeed, suppose \(M_n = n^m\) for some \(m > 0,\) then \(\log N_n = (n/\log n)^d/(2\alpha + d) \log (48M_n^4/C^2\epsilon_n^2),\) which can be shown to be less than \(n^d/(2\alpha + d) \log (n^{2\alpha}/(2\alpha + 2\alpha)/(2\alpha + d))\) when \(C^2 > 48,\) while \(n\varepsilon_n^2/8 = (C^2n^d/(2\alpha + d) \log n^{2\alpha}/(2\alpha + d))/8.\) When \(C^2 > \max(48, 8(8\alpha + 4md + 2\alpha)/(2\alpha + d))\), we have the desired result.

Now it remains to check the upper bound condition for the KL divergence. Recall the notations \(p_{f_i}^* = p_{f_i^*(X_i)}\) and \(p_{f_i} = p_{f_i}(X_i).\) By the assumption, \(f^*\) is a positive continuous function with a compact support. Thus it is bounded away from 0 and from above. Consider \(f\) such that \(\|\sqrt{T} - \sqrt{\|T\|}_\infty\) is sufficiently small. Thus \(f\) is also bounded away from 0 and from above. By Taylor’s expansion, for some \(0 < t < 1,\)

\[K(p_{f_i^*}, p_{f_i}) = f_i - f_i^* + f_i^*(\log(f_i^*) - \log(f_i)) = \frac{f_i^*}{2(tf_i^* + (1-t)f_i)^2}(f_i^* - f_i)^2,\]

which can be bounded as \(\|f^* - f\|_\infty^2 \lesssim \|\sqrt{T} - \sqrt{\|T\|}_\infty^2\). This implies that \(K(P_{f_i^*}(X_i), P_{f_i}(X_i)) = \sum_{i=1}^n K(p_{f_i^*}, p_{f_i}) \lesssim n\|\sqrt{T} - \sqrt{\|T\|}_\infty^2.\]

**Proof of Proposition 2.3.11.** We shall first see Condition (C1) holds. In the following we
use \( j \) as a multi-index. Let \( \tilde{B}_{j,q}(x) = \prod_{k=1}^{d} \tilde{B}_{j_k,q_k}(x_k) \) and \( C_{j,q} = \prod_{k=1}^{d} C_{j_k,q_k} \).

\[
\|\tilde{b}^T \theta_1 - \tilde{b}^T \theta_2\|_\infty = \left\| \sum_{j=1}^{d} J_k (\theta_{1,j} - \theta_{2,j}) \tilde{B}_{j,q} \right\|_\infty = \left\| \sum_{j=1}^{d} \frac{\theta_{1,j} - \theta_{2,j}}{C_{j,q}} B_{j,q} \right\|_\infty \\
\lesssim \left( \prod_{k=1}^{d} J_k^{1/2} \right) \|\theta_1 - \theta_2\|_\infty, \tag{2.15}
\]

where the last expression is due to the fact that \( \sum_j B_{j,q} = 1 \) (the summation is from \( j = 1 \) to \( \prod_{k=1}^{d} J_k \)) and \( C_{j_k,q_k} \lesssim J_k^{1/2} \) in view of Lemma A.0.2. Therefore, when (P1) and (P2) hold as well, one can apply Theorem B.0.1 with \( d_n(\sqrt{T_1}, \sqrt{T_2}) = \|\sqrt{T_1} - \sqrt{T_2}\|_{2,\mu} \), which is equal to the Hellinger metric \( d_H(p_{f_1}, p_{f_2}) \). Notice that \( \|\sqrt{T_1} - \sqrt{T_2}\|_{2,\mu} \leq \|T_1 - T_2\|_\infty \). By the assumption, as a continuous function with a compact support and is strictly positive, \( f^* \) is bounded away from 0 and from above. Consider \( f \) such that \( \|\sqrt{T} - \sqrt{T^*}\|_\infty \) is sufficiently small. Thus \( f \) is also bounded away from 0 and from above. Now by Lemma B.2 of Ghosal and van der Vaart (2017),

\[
K(p_{f^*}, p_f) \leq 2d_H^2(p_{f^*}, p_f) \left\| \frac{p_{f^*}}{p_f} \right\|_\infty \lesssim d_H^2(p_{f^*}, p_f) \leq \|\sqrt{T} - \sqrt{T^*}\|^2_\infty.
\]

Thus \( K(p_{f_1}^{(n)}, p_f^{(n)}) \leq nK(p_{f^*}, p_f) \leq n\|\sqrt{T^*} - \sqrt{T}\|^2_\infty \). Applying Theorem B.0.1, \( d_n(\sqrt{T}, \sqrt{T}) = \|\sqrt{T} - \tilde{b}^T \theta\|_{2,\mu} \) contracts at \( \eta_n = (n/\log n)^{-\alpha/(2\alpha+d)} \).

Now to derive the rates for \( \|\sqrt{T} - \sqrt{T}\|_\infty \). Due to the tensor product representation, \( \prod_{k=1}^{d} J_k \) described here is equivalent to the term \( J \) in Theorem B.0.1. We let \( J_k \propto J_n^* \) for \( k = 1, \ldots, d \). By Lemma A.0.3, there exists some \( \theta^* \) such that \( \|\sqrt{T} - \tilde{b}^T \theta^*\|_\infty \lesssim (J_n^*)^{-\alpha} \) and \( (\theta^*)^{T} \Gamma^{-1/2} \Xi \Gamma^{-1/2} \theta^* = 1 \). Let \( \sqrt{T^\infty} := \tilde{b}^T \theta^* \) and define \( \tilde{\theta}_{j,q} = \theta_{j,q}/C_{j,q} \). Since

\[
\sum_{j=1}^{d} \left( \theta_{j}^\infty - \theta_{j} \right) \frac{B_{j,q}(x)}{C_{j,q}} = \sum_{j=1}^{d} \left( \frac{\theta_{j}^\infty - \theta_{j}}{C_{j,q}} \right) B_{j,q}(x) = \sum_{j=1}^{d} \left( \tilde{\theta}_{j,q} - \tilde{\theta}_{j} \right) B_{j,q}(x),
\]

it follows that

\[
\|\sqrt{T^\infty} - \sqrt{T}\|_\infty = \|\left( \theta^\infty - \theta \right)^T \tilde{b} \|_\infty = \|\left( \tilde{b}^\infty - \tilde{\theta} \right)^T \tilde{b} \|_\infty \lesssim \|	ilde{b}^\infty - \tilde{\theta}\|_\infty \leq \|\tilde{b}^\infty - \tilde{\theta}\|.
\]

By the property that

\[
(\tilde{b}^\infty - \tilde{\theta})^{T} \frac{B^T B}{n} (\tilde{b}^\infty - \tilde{\theta}) \propto \|\tilde{b}^\infty - \tilde{\theta}\|^2_{T,\Xi} = \|\left( \theta^\infty - \theta \right)^T \tilde{b} \|^2_{2,\Gamma},
\]

and the left hand side has the same order as \( \left( \prod_{k=1}^{d} J_k^{-1} \right) \|\tilde{b}^\infty - \tilde{\theta}\|^2 \).
Therefore \( \| \tilde{\theta}^\infty - \hat{\theta} \| \lesssim \left( \prod_{k=1}^d J_k^{1/2} \right) \| \sqrt{f^\infty} - \sqrt{f} \|_{2, G} \). So
\[
\| \sqrt{f^\infty} - \sqrt{f} \|_{\infty} \leq (J_n^*)^{d/2} (\| \sqrt{f^\infty} - \sqrt{f^*} \|_{2, G} + \| \sqrt{f^*} - \sqrt{f} \|_{2, G}) \lesssim (J_n^*)^{d/2} ((J_n^*)^{-\alpha} + \eta_n).
\]
Finally, since \( \| \sqrt{f^\infty} - \sqrt{f} \|_{\infty} \leq \| \sqrt{f^\infty} - \sqrt{f^*} \|_{\infty} + \| \sqrt{f^*} - \sqrt{f} \|_{\infty} \), we have \( \| \sqrt{f^\infty} - \sqrt{f} \|_{\infty} \lesssim (J_n^*)^{-\alpha} + (J_n^*)^{d/2} ((J_n^*)^{-\alpha} + \eta_n) \). With the choice of \( J_n^* \), the contraction rate of \( \| \sqrt{f^\infty} - \sqrt{f} \|_{\infty} \) is \((n/ \log n)^{(d-2\alpha)/(2(d+2\alpha))}\). The rates for derivatives follow similarly as in the proof of Proposition 2.3.7 and the details are omitted.

At last, for Condition (P2) to hold, it suffices to consider \( S(M_n) = (0, \prod_{k=1}^d J_k^{1/2}) \). We shall first show that (i) \( D \) does not depend on \( D \).

Now we prove claim (ii). Let \( 1 \leq l_1, l_2, l'_1, l'_2 \leq n_\alpha \), where \( n_\alpha = n^{1/(1+\alpha)} \) and \((l_1, l_2) \neq (l'_1, l'_2)\).

2.5.3 Proofs of the results for credible sets

**Proof of Theorem 2.3.13.** Let \( Z = f - \tilde{f} \). Note that the posterior distribution of \( Z \) given \( \mathbb{D}_n \) does not depend on \( \mathbb{D}_n \), \( E(\| Z \|_\infty | \mathbb{D}_n) = E(\| Z \|_\infty) \). The proof consists of two parts. In Part (1) we shall first show that (i) \( R_{n, \gamma} \lesssim (\log n)n^{-\alpha/(2(1+\alpha))} \) and (ii) \( E(\| Z \|_\infty | \mathbb{D}_n) \gtrsim (\log n)^{1/2}n^{-\alpha/(2(1+\alpha))} \).

In Part (2) we prove the claims of this theorem.

Part (1). Claim (i) is straightforward by the third assertion of Proposition A.2.1 of Van der Vaart and Wellner (1996), i.e., \( R_{n, \gamma} \) is bounded by \( \sqrt{8E(\| Z \|_\infty^2 | \mathbb{D}_n) \log(2/\gamma)} \), which is of the order \( n^{-\alpha/(2(1+\alpha))} \log n \) in view of the Part (1) in the proof of Proposition 2.3.1.

Now we prove claim (ii). Let \( 1 \leq l_1, l_2, l'_1, l'_2 \leq n_\alpha \), where \( n_\alpha = n^{1/(1+\alpha)} \) and \((l_1, l_2) \neq (l'_1, l'_2)\).
Without loss of generality, we take $l_1 > l_1'$. Consider
\[
E \left( \left| Z\left( \frac{l_1}{n_\alpha}, \frac{l_2}{n_\alpha}\right) - Z\left( \frac{l_1'}{n_\alpha}, \frac{l_2'}{n_\alpha}\right) \right|^2 \right)
\]
\[= \sum_{j_1=0, j_2=0}^{\infty} \sum_{k \in K(j)} \frac{1}{n + (\mu_{j_1}, \mu_{j_2})^T} \left| \phi_{j_1,j_2k}\left( \frac{l_1}{n_\alpha}, \frac{l_2}{n_\alpha}\right) - \phi_{j_1,j_2k}\left( \frac{l_1'}{n_\alpha}, \frac{l_2'}{n_\alpha}\right) \right|^2.
\]
It is clearly greater than the sum over $j_1 \in \mathbb{N}$ and $j_2 = 0$. Therefore it is lower bounded by
\[
\sum_{j_1 \in \mathbb{N}} \frac{1}{n + j_1^{1+\alpha}} \left| 2\sin\left( \frac{2\pi l_1 j_1}{n_\alpha} \right) - 2\sin\left( \frac{2\pi l_1' j_1}{n_\alpha} \right) \right|^2.
\] (2.16)

Pick some $\epsilon > 0$ arbitrarily small. Note that
\[
\left| \sin\left( \frac{2\pi l_1 j_1}{n_\alpha} \right) - 2\sin\left( \frac{2\pi l_1' j_1}{n_\alpha} \right) \right| > \epsilon
\]
whenever the following two inequalities hold:
\[
\frac{2j_1(l_1 - l_1')}{n_\alpha} - m > \delta \text{ for } m \in \mathbb{Z} \text{ even,}
\] (2.17)
\[
\frac{2j_1(l_1 + l_1')}{n_\alpha} - m > \delta \text{ for } m \in \mathbb{Z} \text{ odd,}
\] (2.18)
for some very small positive constant $\delta$. Consider $j_1 \in \{1, \ldots, n_\alpha\}$ and let $j_1/n_\alpha = s$. So $0 < s \leq 1$. Inequality (2.17) holds when $s(l_1 - l_1')$ is not in some $\delta$-neighborhood of an integer up to $l_1 - l_1'$. Thus $s$ has to avoid some $\delta/(l_1 - l_1')$ neighborhood of a finite number of rational numbers. Similarly, Inequality (2.18) holds when $s(l_1 + l_1')$ is not in some $\delta$-neighborhood of a half integer up to $l_1 + l_1'$. Thus $s$ has to avoid some $\delta/(l_1 + l_1')$-neighborhood of a finite number of rational numbers. Altogether, when $s$ is not in a set which is an union of finitely many very small intervals, (2.17) and (2.18) hold. The complement of the set, being also a union of finitely many intervals, has positive Lebesgue measure. Note also that the set is not changing with $n_\alpha$. Therefore, the fraction of $\{j_1 : 1 \leq j_1 \leq n_\alpha\}$ so that $j_1/n_\alpha$ falls in the admissible region is bounded away from zero. Thus (2.16) is lower bounded by the sum over $j_1$ from 1 to $n_\alpha$ which is further lower bounded by upper bound $c n_\alpha/(2n)$ for some small positive constant $c$. This gives the order $cn^{-\alpha/(1+\alpha)}$.

Let $U_{l_1,l_2} = \sqrt{2e^{-1/2}n^{\alpha/(2(1+\alpha))}} Z\left( \frac{l_1}{n_\alpha}, \frac{l_2}{n_\alpha}\right)$. This then gives
\[
E(U_{l_1,l_2} - U_{l_1',l_2}')^2 \geq 2 = E(V_{l_1,l_2} - V_{l_1',l_2}')^2,
\]
52
for some i.i.d standard normal random variable $V_{l_1,l_2}$. By Slepian’s lemma and Lemma A.0.1,

$$E(\max_{l_1,l_2} U_{l_1,l_2}) \gtrsim E(\max_{l_1,l_2} V_{l_1,l_2}) \gtrsim \sqrt{\log(n^2_{\alpha})} \asymp \sqrt{\log n}.$$ 

Since $E(\max_{l_1,l_2} Z(l_1/n_{\alpha}, l_2/n_{\alpha})) \gtrsim n^{-\alpha/(2(1+\alpha))} E(\max_{l_1,l_2} U_{l_1,l_2})$, we obtain

$$E\|Z\|_{\infty} \geq E\left( \max_{l_1,l_2} Z \left( \frac{l_1}{n_{\alpha}}, \frac{l_2}{n_{\alpha}} \right) \right) \gtrsim n^{-\alpha/(2(1+\alpha))} \sqrt{\log n}.$$

Part (2). Note that since for $0 < \gamma < 1/2$, $R_{n,\gamma}$ is greater than the posterior median of $\|Z\|_{\infty}$. By the property of Gaussian process, the posterior mean of $\|Z\|_{\infty}$ and the posterior median of $\|Z\|_{\infty}$ are of the same order. Then (i) and (ii) of Part (1) imply that

$$R_{n,\gamma} \gtrsim E(\|Z\|_{\infty} | D_n) = E(\|Z\|_{\infty}) \asymp n^{-\alpha/(2(1+\alpha))} \sqrt{\log n}.$$ 

We are now ready to prove the claim of the theorem. Let $\epsilon_n = n^{-\alpha/(2(1+\alpha))} \log n$. By Borell’s inequality (see Proposition A.2.1 of Van der Vaart and Wellner (1996)),

$$\Pi(\mathcal{L} \notin \mathcal{C}_L | D_n) \leq \Pi(\mathcal{L} \text{ corresponds to } f \notin \mathcal{C}_{f,\gamma} | D_n)$$

$$= \Pi(\|Z\|_{\infty} > \rho_n R_{n,\gamma} | D_n)$$

$$= \Pi(\|Z\|_{\infty} - E(\|Z\|_{\infty} | D_n) > \rho_n R_{n,\gamma} - E(\|Z\|_{\infty} | D_n) | D_n)$$

$$\leq \exp[-C^2 \epsilon_n^2 / c_n],$$

where $c_n = \|E(Z^2(\cdot) | D_n)\|_{\infty}$ is bounded by $(\log n)n^{-\alpha/(1+\alpha)}$ in view of Part (1) result in the proof of Proposition 2.3.1. Therefore, the above posterior probability tends to zero. Finally,

$$P_0(\mathcal{L}(f^*) \in \mathcal{C}_L) = P_0(\|\mathcal{f} - f^*\|_{\infty} \leq \rho_n R_{n,\gamma}) = P_0(\|E\mathcal{f} - f^* + \mathcal{f} - E\mathcal{f}\|_{\infty} \leq \rho_n R_{n,\gamma}) \to 1,$$

by the Parts (2) and (3) results in the proof of Proposition 2.3.1, establishing the coverage of $C_L$.

To see $C_L \subset \mathcal{C}_L$, for any $\mathcal{L} \in C_L$, it is induced by some $f$ such that $\|f - \mathcal{f}\|_{\infty} \leq \rho_n R_{n,\gamma}$. In view of the Part (1) result in the proof of Proposition 2.3.1, the claim immediately follows.

**Proof of Theorem 2.3.14.** We take $J_k \asymp J_k^n = (n/\log n)^{1/(2\alpha+d)}$, $k = 1, \ldots, d$. For $0 < \gamma < 1/2$, by argument in the proof of Theorem of 5.3 of Yoo and Ghosal (2016), one can establish
that

\[
R_{n,\gamma} \asymp E(\|f - \tilde{f}\|_\infty | \mathcal{D}_n, \sigma^2),
\]

\[
R^2_{n,\gamma} \asymp \frac{\log n}{n} (J^*_n)^d = (\log n/n)^{2\alpha/(2\alpha + d)}.
\]

Recall from (2.8), \(\Pi(f|\mathcal{D}_n, \hat{\sigma}^2) \sim \text{GP}(\tilde{f}, \hat{\sigma}^2 \Sigma)\). Arguing similarly as in the proof of Theorem 2.3.13,

\[
\Pi(\mathcal{L} \notin C_\mathcal{L}|\mathcal{D}_n, \hat{\sigma}^2) \leq \exp\left[-C^2 R^2_{n,\gamma}/c_n\right],
\]

where \(c_n = \sup_x \text{var}(f - \tilde{f}|\mathcal{D}_n, \hat{\sigma}^2)\) is bounded by a constant multiple of

\[
\sup_x \Sigma(x, x) \lesssim \sup_x \|b(x)\|^2 \|\left(B^T B + \Lambda^0\right)^{-1}\|(2, 2) \lesssim (J^*_n)^d/n.
\]

Therefore, the above posterior probability tends to zero. In addition,

\[
P_0(\mathcal{L}(f^*) \in C_\mathcal{L}) = P_0(\|\tilde{f} - f^*\|_\infty \leq \rho R_{n,\gamma}) \to 1,
\]

by the third assertion of Proposition 2.3.5, establishing the coverage of \(C_\mathcal{L}\). To see \(C_\mathcal{L} \subset \bar{C}_\mathcal{L}\), for any \(\mathcal{L} \in C_\mathcal{L}\), it is induced by some \(f\) such that \(\|f - \tilde{f}\|_\infty \leq \rho R_{n,\gamma}\). In view of Lemma 2.2.2 and the second assertion of Proposition 2.3.5, the result immediately follows. ■
Chapter 3

Posterior Contraction and Credible Regions for Filaments of Regression Functions

3.1 Introduction

There is a large body of literature on the problem of estimating intrinsic lower dimensional structures of multivariate data. A filament or a ridge line is one of such geometric objects that draws a lot of attention in the recent years. Intuitively speaking, a filament consists of local maximizers of a smooth function $f$ (say a density or a regression function) when moving in a certain direction. Roughly speaking these are generalized modes that reside on hyperplanes that are normal to the steepest ascent direction. The Figure 3.1 are useful to convey the idea.

The filamentary structures (ridges) together with the valleys (i.e. local minimizers counterparts of ridges), critical points are the main features of the shapes of objects. They are common in medical images, satellite images and many other three dimensional objects. One important example comes from the study of the cosmic web — a large scale web structure of galaxies (clusters) connected by long threads composed of sparse hydrogen gas. These intergalactic connections are believed to trace the filaments of dark matters. The discovery and study of the dark matter is a key challenge of cosmology. For more references, see Novikov, Colombi and Doré (2006), Dietrich, Werner, Clowe, Finoguenov, Kitching, Miller and Simionescu (2012) and Chen, Ho, Brinkmann, Freeman, Genovese, Schneider and Wasserman (2015).
The filament estimation is closely related to manifold learning problem. Manifold learning problem assumes that the data points are generated from some a priori unknown lower dimensional structure with background noises. Arias-Castro, Donoho and Huo (2006) developed a test to detect if a dataset contains some small fraction of data points that are supported on a curve. Genovese, Perone-Pacifico, Verdinelli and Wasserman (2012) and Genovese, Perone-Pacifico, Verdinelli and Wasserman (2014) studied the problem of estimation of the manifold. They showed that the ridge of a density function can serve as a surrogate to the manifold and can be estimated with a better rate.

Recently, a promising algorithm called the Subspace Constrained Mean Shift (SCMS) algorithm, is proposed in Ozertem and Erdogmus (2011). It can be considered as an algorithm that generalizes the celebrated mean shift algorithm, which is designed for mode hunting (Fukunaga and Hostetler; 1975). The consistency properties of SCMS has also been a subject of some recent research (Arias-Castro, Mason and Pelletier; 2016). As the filament is a feature induced by some unknown function, say density, nonparametric estimation is often applied. The statistical properties of the estimated filaments, like convergence rates, limiting distribution and the construction of valid confidence regions are crucial for the assessment of the proposed method. There are a few recent papers on these issues. Genovese et al. (2014) established the convergence rates of the filament obtained from the kernel density estimation. Chen, Genovese and Wasserman (2015) provided Berry-Esseen type results for the limiting distribution and also developed a bootstrap-based method for uncertainty quantification of filaments. With a different approach, Qiao and Polonik (2016) also established the convergence rates and the extreme value type results for limiting distribution.

Since the filaments can be considered as generalized modes, relevant literature includes those on mode (and maximum) estimation, for instance, Facer and Müller (2003); Shoung and Zhang.
(2001) and a recent paper by Yoo and Ghosal (2017) in the Bayesian framework.

So far the study of filaments is limited only to densities only using the kernel approach. Our study supplements the current literature in two ways. First, we provide a Bayesian approach to the filament estimation in the regression context and study the contraction rates using a finite random B-splines series prior. This has theoretical advantages as the bias can be better controlled when the function is more than “minimally smooth” (i.e. more than four times differentiable). Secondly, we provide a way to construct credible set with sufficient frequentist coverage for the filaments. Different from the bootstrap-based confidence region proposed by Chen, Genovese and Wasserman (2015) which gives a band-shape region, our valid credible region consists of posterior filaments that have frequentist interpretation. Another difference is that the inferential target in this paper is the true quantity itself, while in Chen, Genovese and Wasserman (2015) the inference is targeted towards the debiased quantity.

Before we move on to the formal definition, it is worth to point out that some other possible definitions of ridges or filaments have also been discussed and studied in mathematics and computer sciences literature; see Eberly (1996) for more details. We study the filament as introduced in Eberly (1996), Chen, Genovese and Wasserman (2015) and Qiao and Polonik (2016).

This chapter is organized as follows. Some preliminary materials are given in Section 3.2. The model, prior and posterior are introduced in Section 3.3. Technical assumptions are given in Section 3.7. The main results in posterior contraction and credible region are presented in Section 3.5. Simulation results and an application for earthquake data are presented in Section 3.6 and 3.7 respectively. All proofs are given in Section 3.8.

3.2 Preliminaries

The filament or the ridge line of a smooth function defined on $\mathbb{R}^2$ is a collection of points at which the gradient of the function is orthogonal to the eigenvector of its Hessian whose corresponding eigenvalue is negative. The filament point (that is the point on the filament) is a generalization of mode of the function. To see this connection, recall a well-known result that tests for local maximum point (mode).

Let $f: \mathbb{R}^2 \mapsto \mathbb{R}$ be a smooth function, $\nabla f = (f^{(1,0)}, f^{(0,1)})^T$ be the gradient and $Hf$ be the Hessian. Recall a test for a local maximum point is the following

$$a^T\nabla f(x) = 0, \quad a^T Hf(x)a < 0,$$

for all nonzero vector $a$. Let $V(x)$ be the eigenvector of $Hf(x)$ that corresponds to the smallest
eigenvalues $\lambda(x)$. A point $x \in \mathbb{R}^2$ is called a ridge point if
\[ V^T(x)\nabla f(x) = 0, \quad V^T(x)Hf(x)V(x) < 0. \]
Therefore, a ridge point is a point at which the function has a local maximum along the direction given by $V$. Notice that $V^T(x)Hf(x)V(x) < 0$ is equivalent to $\lambda(x) < 0$.

More generally, for $f$ defined on $\mathbb{R}^d$, $0 \leq s \leq d-1$, the eigenvectors of $Hf(x)$ can be used to define two orthogonal spaces, namely, a $(d-s)$-dimensional normal space (corresponding to $(d-s)$-eigenvectors with smallest eigenvalues) and an $s$-dimensional tangent space (corresponding to the rest eigenvectors). An $s$-dimensional ridge point on $\mathbb{R}^d$ is a point where the gradient of $f$ is orthogonal to the normal space and the eigenvalues associated with the normal space are all (strongly) negative. Alternatively, such a point $x$ can be regarded as a point where $f$ attains the local maximum in the affine space spanned by the normal space translated by $x$. The modes are then simply 0-dimensional filaments. The 1-dimensional filament on $\mathbb{R}^2$ is of primary interest in our discussion here.

From now on, $f$ is assumed to be some smooth regression function. Suppose that the Hessian matrix $Hf(x)$ of $f$ at $x$ has eigenvector $V(x)$ corresponding to the smaller eigenvalue $\lambda(x)$. The filament $\mathcal{R}(f)$ of the regression function $f$ is formally defined as
\[ \mathcal{R}(f) = \mathcal{L} := \{ x : \langle \nabla f(x), V(x) \rangle = 0 \text{ and } \lambda(x) < 0 \}. \]

We also introduce an integral curve, which is the solution of the following differential equation
\[ \frac{d\Upsilon_{x_0}(t)}{dt} = V(\Upsilon_{x_0}(t)), \quad \Upsilon_{x_0}(0) = x_0. \]
Since $\Upsilon_{x_0}(-t) = \int_0^t (-V(\Upsilon_{x_0}(-s)))ds$, with negative time it can be interpreted as a curve tracing in the reverse direction, i.e., $-V$. Since the direction of $V$ does not play a role in the theoretical study, without loss of generality, $t$ is restricted on the $[0, T_{\text{max}}]$. Define the “hitting time” of the filament by the integral curve starting at point $x_0$,
\[ t_{x_0} = \arg\min\{ t \geq 0 : \langle \nabla f(\Upsilon_{x_0}(t)), V(\Upsilon_{x_0}(t)) \rangle = 0, \lambda(\Upsilon_{x_0}(t)) < 0 \}. \]

3.3 Model, prior and posterior

Throughout this chapter, let $d = 2$, thus $x = (x_1, x_2)$. We will use superscript * to emphasize the true values. Consider the nonparametric regression model
\[ Y_i = f(X_i) + \varepsilon_i, \]
where \( \varepsilon \sim \text{i.i.d.} \, \text{N}(0, \sigma^2) \) for \( i = 1, \ldots, n \). Both \( f \) and \( \sigma \) are unknown parameters. Let \( Y = (Y_1, \ldots, Y_n)^T, X = (X_1^T, \ldots, X_n^T)^T, F = (f(X_1), \ldots, f(X_n))^T \) and \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \), then we can write \( Y = F + \varepsilon \). We model \( f = b^T \theta \). The rest setup is the same as that in the Section 2.3.2.

Recall the vector

\[
b^{(r)}_{j_1,j_2}(x) = \left( \frac{\partial r_1}{\partial x_1} B_{j_1,q_1}(x_1) \frac{\partial r_2}{\partial x_2} B_{j_2,q_2}(x_2) : 1 \leq j_k \leq J_k, k = 1, 2 \right)^T.
\]

Also from Section 2.3.2, we have

\[
\Pi(D^T f | \mathbb{D}_n, \sigma^2) \sim \text{GP}(A_r Y + C_r \theta_0, \sigma^2 \Sigma_r),
\]
\[
A_r(x) = b^{(r)}_{j_1,j_2}(x)^T (B^T B + \Lambda_0^{-1})^{-1} B^T,
\]
\[
C_r(x) = b^{(r)}_{j_1,j_2}(x)^T (B^T B + \Lambda_0^{-1})^{-1} \Lambda_0^{-1},
\]
\[
\Sigma_r(x,y) = b^{(r)}_{j_1,j_2}(x)^T (B^T B + \Lambda_0^{-1})^{-1} b^{(r)}_{j_1,j_2}(y),
\]

where we use \( \text{GP} \) to denote a Gaussian process. Notice that under \( P_0 \), \( A_r(x)\varepsilon/\sigma_0 \) is a mean zero process with sub-Gaussian tail. As before, we take an empirical Bayes approach for \( \sigma^2 \). The empirical Bayes has the following posterior distribution

\[
\Pi(D^T f | \mathbb{D}_n, \sigma^2)_{| \sigma^2 = \hat{\sigma}^2} \sim \text{GP}(A_r Y + C_r \theta_0, \hat{\sigma}^2 \Sigma_r), \tag{3.1}
\]

where

\[
\hat{\sigma}^2_n = n^{-1} (Y - B \theta_0)^T (B \Lambda_0 B^T + I_n)^{-1} (Y - B \theta_0).
\]

### 3.4 Assumptions

We follow the standard assumptions in Qiao and Polonik (2016). For convenience, we write \( D^T f = f^{(r)} \) and let \( d^2 f(x) = (f^{(2,0)}(x), f^{(1,1)}(x), f^{(0,2)}(x))^T, \nabla f = (f^{(1,0)}, f^{(0,1)})^T \) and \( H f(x) \) be the Hessian of \( f \). We assume that the two distinct eigenvalues of \( H f(x) \) are distinct. Then \( V(x) \) and \( \lambda(x) \) take the following forms \( V(x) = G(d^2 f(x)) \) and \( \lambda(x) = J(d^2 f(x)) \) for some function \( G = (G_1, G_2)^T: \mathbb{R}^3 \to \mathbb{R}^2 \) and \( J: \mathbb{R}^3 \to \mathbb{R} \) given by

\[
G(u,v,w) = \left( \begin{array}{c}
2u - 2w + 2v - 2\sqrt{(w-u)^2 + 4v^2} \\
w - u + 4v - \sqrt{(w-u)^2 + 4v^2}
\end{array} \right),
\]
\[
J(u,v,w) = \frac{1}{2} \left( u + w - \sqrt{(u-w)^2 + 4v^2} \right).
\]

59
Throughout the proofs, we may take the normalized version of the eigenvector $V$, that is, $\|V\| = 1$. This is not necessary but it simplifies discussions.

We choose some compact subset $\mathcal{G}$ of $[0, 1]^2$ so that $t_{x_0}^*$ is unique for any $x_0 \in \mathcal{G}$. Define $L \oplus \delta = \cup_{x \in \mathcal{L}} B(x, \delta)$, where $B(x, \delta)$ is an open ball around $x$ of radius $\delta$. The following assumptions will be needed for the theory.

(A1) The truth $f^*$ belongs to a H"older Space $\mathcal{H}^\alpha([0, 1]^2)$ with $\alpha \geq 4$.

(A2) There is some $\delta > 0$ small, such that for all $x \in L^* \oplus \delta$, $Hf^*(x)$ has two distinct eigenvalues, with smaller eigenvalue $\lambda^*(x) \leq -\eta$ for some small positive value $\eta$.

(A3) For the $\delta > 0$ in (A2), for all $x \in L^* \oplus \delta$, $|\langle \nabla f^*(x), V^*(x) \rangle| \geq \eta$ for the same positive value $\eta$ in (A2).

(A4) The filament $L^*$ is a compact set such that $L^* = \{ \mathcal{Y}_{x_0}^*(t_{x_0}^*) : x_0 \in \mathcal{G} \}$.

(A5) Assume that there exits $a^* > 0$ such that for some $C_G > 0$, for any $x_0 \in \mathcal{G}$,

$$\inf_{x_0 \in \mathcal{G}} \inf_{0 \leq s < u < t_{x_0}^* + a^*} \left\| \mathcal{Y}_{x_0}(u) - \mathcal{Y}_{x_0}(s) \right\| u - s \geq C_G.$$ 

Assumption (A2) is important for our analysis as it guarantees the smoothness of $V^*$ (Qiao and Polonik; 2016) and that the normal vector of the filament $\nabla \langle \nabla f^*(x), V^*(x) \rangle$ is well-defined. Similar assumptions are needed in Genovese et al. (2014) and Chen, Genovese and Wasserman (2015).

Assumption (A3) says that the normal vector of the filament $\nabla \langle \nabla f^*(x), V^*(x) \rangle$ is not orthogonal to $V^*(x)$. In addition, it implies that $\nabla \langle \nabla f^*(x), V^*(x) \rangle \neq 0$, i.e., the set $\{ x : \langle \nabla f^*(x), V^*(x) \rangle = 0 \}$ is not “thick”. This means that a small change of $x \in L$ necessarily changes the sign of $\langle \nabla f^*(x), V^*(x) \rangle$. If one restricts attention to the locus defined by $\langle \nabla f^*(x), V^*(x) \rangle = 0$, noting rank($\nabla \langle \nabla f^*(x), V^*(x) \rangle$) = 1, the implicit function theorem says that $L$ is a one-dimensional manifold in $\mathbb{R}^2$. If for $x \in L^*$, $\nabla f^*(x) = 0$, Assumption (A3) should be interpreted as $|\langle (V^*(x))^T Hf^*(x)V^*(x) \rangle| > \eta$ which is then implied by Assumption (A2). To see this, note that

$$\langle \nabla \langle \nabla f^*(x), V^*(x) \rangle, V^*(x) \rangle = \nabla f^*(x)^T \nabla V^*(x)V^*(x) + V^*(x)^T Hf^*(x)V^*(x),$$

but $V^*(x)^T Hf^*(x)V^*(x) = \lambda^*(x)$. Assumption (A3) parallels the assumption (A2) in Genovese et al. (2014), where they assumed some upper bounds on the quantity related to the third derivative of the density function; see also the assumption (P1) in Chen, Genovese and Wasserman (2015). Assumption (A5) is common in the literature; see Koltchinskii, Sakhanenko and Cai (2007); Qiao and Polonik (2016). Several useful consequences of these assumptions are summarized in Remark 3.8.1.
3.5 Posterior contraction and credible sets for filaments

In this section, we provide the main theoretical results. Here is a brief outline of these results. Theorem 5.1 provides posterior contraction rates for the integral curve. Proposition 5.2 gives posterior contraction rates for the hitting time. Theorem 5.3 gives the posterior contraction rates for the filament along the integral curve. Theorem 5.8 establishes the posterior contraction rates for the filament around the truth, posterior rates for deviations between the posterior filament and the filament induced by the posterior mean, together with the convergence rates for the filament induced by posterior mean to that of the truth, all in Hausdorff distance. Theorem 5.9 provides a valid credible set with sufficiently high frequentist coverage.

The following result gives a Bayesian counterpart of Theorem 3.3 of Qiao and Polonik (2016). Our proof is similar to theirs, but some technical details are different.

**Theorem 3.5.1** Under Assumptions (A1), (A2) and (A5), for \( J_1 \asymp J_2 \asymp (n/\log n)^{1/(1+2\alpha)} \) and \( T_{x_0}^* = t_{x_0}^* + a^* \), for \( \epsilon_n = (n/\log n)^{(2-\alpha)/(1+2\alpha)} \) and any \( M_n \to \infty \),

\[
\Pi \left( \sup_{x_0 \in \mathcal{G}} \sup_{t \in [0,T_{x_0}^*]} \| \Upsilon_{x_0}(t) - \Upsilon_{x_0}^*(t) \| > M_n \epsilon_n \|D_n\rangle \right) \xrightarrow{P_0} 0.
\]

**Remark 3.5.1** With the choice \( J_1 = J_2 = J \asymp (n/\log n)^{1/(1+2\alpha)} \), we obtain the same rate as in the Theorem 3.3 of Qiao and Polonik (2016) up to a logarithmic term. For \( \alpha = 4 \), the rate reduces to \( (n/\log n)^{-2/9} \). Note that if we choose \( J \asymp (n/\log n)^{1/(2(1+\alpha))} \) as it is the optimal choice for estimation of the function, it is easy to see from the proof that now \( \sup_x \| d^2 f(x) - d^2 f^*(x) \| \) is of the order \( (n/\log n)^{(2-\alpha)/(1+\alpha)} \) and the term (3.2) in Section 3.8 has posterior contraction rate of the order \( (n/\log n)^{(2-\alpha)/(1+\alpha)} \). Thus the contraction rate will then be \( \epsilon_n = (n/\log n)^{(2-\alpha)/(2(1+\alpha))} \), which is “suboptimal” in the present context.

The following proposition is a Bayesian analog of Proposition A.1 and Proposition 5.1 of Qiao and Polonik (2016). We can obtain better rate by using different magnitude of the tuning parameter; see the remark after the proposition.

**Proposition 3.5.2** Under Assumptions (A1)–(A5), for \( J_1 \asymp J_2 \asymp (n/\log n)^{1/(1+2\alpha)} \), for \( \epsilon_n = (n/\log n)^{(5-2\alpha)/(2(1+2\alpha))} \) and any \( M_n \to \infty \),

\[
\Pi \left( \sup_{x_0 \in \mathcal{G}} |t_{x_0}^* - t_{x_0}| > M_n \epsilon_n \|D_n\rangle \right) \xrightarrow{P_0} 0.
\]

If in addition, \( \nabla f^*(x) = 0 \) for all \( x \in \mathcal{L}^* \), then the rates improve to \( \epsilon_n = (n/\log n)^{(2-\alpha)/(1+2\alpha)} \).

**Remark 3.5.2** A better rate can be obtained if we choose \( J_1 = J_2 = J \asymp (n/\log n)^{1/(2(1+\alpha))} \). By Lemma 3.8.1, \( \sup_x \| \nabla f(x) - \nabla f^*(x) \| \) has posterior contraction rate \( (n/\log n)^{(1-\alpha)/(2(1+\alpha))} \).
while \( \sup_x \|V(x) - V^*(x)\| \) has posterior contraction rate \((n/ \log n)^{2-\alpha}/(2(1+\alpha))\). Recalling Remark 3.5.1, the posterior contraction rate will then be \( \epsilon_n = (n/ \log n)^{2-\alpha}/(2(1+\alpha)) \). If in addition, \( \nabla f^*(x) = 0 \) for all \( x \in \mathcal{L}^* \), the function then is a plateau without any rise or fall along the direction of filament. In this case, the rate improves to \( \epsilon_n = (n/ \log n)^{(1-\alpha)/(2(1+\alpha))} \).

Theorem 3.5.3 below is a Bayesian analog of the convergence of filaments points on the integral curves starting from the same points; see Qiao and Polonik (2016), Section 3.4, for similar results. Again a better rate is possible; see the remark following the theorem.

**Theorem 3.5.3** Under Assumptions (A1)–(A5), for \( J_1 \gg J_2 \gg (n/ \log n)^{1/(1+2\alpha)} \), for \( \epsilon_n = (n/ \log n)^{(5-2\alpha)/(2(1+2\alpha))} \) and any \( M_n \to \infty \),

\[
\Pi(\sup_{x_0 \in \mathcal{G}} \|\bar{Y}_{x_0}(t_{x_0}) - \hat{Y}_{x_0}(t_{x_0})\| > M_n \epsilon_n | \mathcal{D}_n) \to 0.
\]

**Remark 3.5.3** A better rate can be obtained with the choice \( J_1 = J_2 \gg (n/ \log n)^{1/(2(1+\alpha))} \), giving \( \epsilon_n = (n/ \log n)^{(2-\alpha)/(2(1+\alpha))} \) in view of Remark 3.5.1 and Remark 3.5.2.

Recall the definition of the Hausdorff distance in (1.8). Given two sets \( A \) and \( B \) under Euclidean metric, let \( d(A|B) := \sup_{x \in A} \inf_{y \in B} \|x - y\| \). The Hausdorff distance between \( A \) and \( B \) is defined as

\[
\text{Haus}(A, B) = \max\{d(A|B), d(B|A)\}.
\]

In what follows, we provide a refined bound for the Hausdorff distance and construct credible sets for the filaments. In fact, Theorem 3.5.3 gives an upper bound for the Hausdorff distance. However, for the purpose of constructing credible sets with sufficient frequentist coverage, we need to have the upper bound in terms of more primitive quantities such as the derivatives of underlying function. In view of Remark 3.5.3, in this section we will restrict to the choice \( J_1 \gg J_2 \gg J \gg (n/ \log n)^{1/(2(1+\alpha))} \).

Recall that \( \hat{f} = A\bar{Y} + C\theta_0 \) is the posterior mean of \( f \) conditional on \( \mathcal{D}_n \) and that \( \hat{V}, \hat{\bar{Y}}_{x_0}, \hat{\bar{L}} \) are the corresponding eigenvector, integral curve and filament induced by \( \hat{f} \). In view of Lemma 3.8.2 and Lemma 3.8.3, following the proofs of Theorem 3.5.1, Proposition 3.5.2, Theorem 3.5.3, it is straightforward to show that \( \sup_{x \in \mathcal{G}} \sup_t \|Y_{x_0}(t) - \bar{Y}_{x_0}(t)\| \), \( \sup_{x \in \mathcal{G}} \|t_{x_0} - \bar{t}_{x_0}\| \) and \( \sup_{x \in \mathcal{G}} \|Y_{x_0}(t_{x_0}) - \bar{Y}_{x_0}(\bar{t}_{x_0})\| \) are all small with high posterior probability in \( P_0 \)-probability. Likewise, it can be shown that the quantities induced by \( \hat{f} \) converge to the corresponding true quantities induced by \( f^* \). For instance \( \hat{Y}_{x_0}(t) \) converges to \( Y^*_{x_0}(t) \) uniformly in \( x_0 \in \mathcal{G} \) and \( \bar{t}_{x_0} \) converges to \( t^*_{x_0} \) uniformly in \( x_0 \in \mathcal{G} \) and \( \bar{Y}_{x_0}(\bar{t}_{x_0}) \) converges to \( Y^*_{x_0}(t^*_{x_0}) \) uniformly in \( x_0 \in \mathcal{G} \) in \( P_0 \)-probability.

The following two theorems summarize above argument. Theorem 3.5.4 is on the convergence rates of the Bayesian estimates of filaments to the true filaments. Theorem 3.5.5 is on the
posterior contraction rates of filaments around filaments induced by the posterior mean.

**Theorem 3.5.4** Under Assumptions (A1), (A2) and (A5), for $J_1 \asymp J_2 \asymp (n/\log n)^{1/(2(1+\alpha))}$, $T^*_{\theta_0} = t^*_{\theta_0} + a^*$, we have the following convergence rates:

$$\sup_{x_0 \in G} \sup_{t \in [0, T^*_{\theta_0}]} \| \tilde{\Upsilon}_{x_0}(t) - \Upsilon^*_{x_0}(t) \| = O_p \left( \left( n/\log n \right)^{2-\alpha/(2(1+\alpha))} \right).$$

If in addition, (A3) and (A4) hold, then

$$\sup_{x_0 \in G} | \tilde{t}_{x_0} - t^*_{x_0} | = O_p \left( \left( n/\log n \right)^{2-\alpha/(2(1+\alpha))} \right),$$

and

$$\sup_{x_0 \in G} \| \tilde{\Upsilon}_{x_0}(\tilde{t}_{x_0}) - \Upsilon^*_{x_0}(t^*_{x_0}) \| = O_p \left( \left( n/\log n \right)^{2-\alpha/(2(1+\alpha))} \right).$$

**Theorem 3.5.5** Under Assumptions (A1), (A2) and (A5), for $J_1 \asymp J_2 \asymp (n/\log n)^{1/(2(1+\alpha))}$, $T^*_{\theta_0} = t^*_{\theta_0} + a^*$, we have the following posterior contraction rates: for any $M_n \to \infty$,

$$\Pi(\sup_{x_0 \in G} \sup_{t \in [0, T^*_{\theta_0}]} \| \Upsilon_{x_0}(t) - \tilde{\Upsilon}_{x_0}(t) \| > M_n(n/\log n)^{(3-2\alpha)/(4(1+\alpha))} \| \mathbb{D}_n \|) \xrightarrow{P_0} 0.$$

If in addition, (A3) and (A4) hold, then

$$\Pi(\sup_{x_0 \in G} | t_{x_0} - \tilde{t}_{x_0} | > M_n(n/\log n)^{(2-\alpha)/(2(1+\alpha))} \| \mathbb{D}_n \|) \xrightarrow{P_0} 0,$$

and

$$\Pi(\sup_{x_0 \in G} \| \Upsilon_{x_0}(t_{x_0}) - \tilde{\Upsilon}_{x_0}(\tilde{t}_{x_0}) \| > M_n(n/\log n)^{(2-\alpha)/(2(1+\alpha))} \| \mathbb{D}_n \|) \xrightarrow{P_0} 0.$$

The following proposition says that with high posterior probability the induced filament from the posterior satisfies similar properties the true filament has, so does the induced filament by $\tilde{f}$, with $P_0$-probability tending to one.

**Proposition 3.5.6** Suppose that $f$ has a tensor-product B-splines prior with order $q_1 = q_2 \geq \alpha$ and $J_1 \asymp J_2 \asymp (n/\log n)^{1/2(\alpha+1)}$, then the following assertions hold.

(i) The filament $\mathcal{L}$ of $f$ drawn from the posterior distribution satisfies assumptions (A2)–(A5) with posterior probability tending to 1 under $P_0$-probability.

(ii) The induced filament $\tilde{\mathcal{L}}$ of the posterior mean $\tilde{f}$ satisfies assumptions (A2)–(A5) with $P_0$-probability tending to 1.
Consider two filament curves $L$ and $\hat{L}$ that are close enough. This can be guaranteed in the probability limit in view of previous theorem if the inducing functions $f$ and $\hat{f}$ are close enough. Hence the second eigenvectors $V$ and $\hat{V}$ and the corresponding integral curves $\Upsilon$ and $\hat{\Upsilon}$ are close enough. The following lemma is inspired by Genovese et al. (2014) Theorem 4, where they relate the Hausdorff distance to $\|V\|$ but under a different set of assumptions.

**Lemma 3.5.7** Consider two regression functions $f$ and $\hat{f} : [0, 1]^2 \mapsto \mathbb{R}$ that are sufficiently close in supremum metric and both satisfy assumptions (A1)-(A4), then the Hausdorff distance between the two induced filaments satisfies, for some positive constants $C$,

$$\text{Haus}(L, \hat{L}) \leq \frac{C}{\eta} \left( \sup_{x \in [0,1]^2} |f(2,0)(x) - \hat{f}(2,0)(x)| + \sup_{x \in [0,1]^2} |f(1,1)(x) - \hat{f}(1,1)(x)| \right.\left. + \sup_{x \in [0,1]^2} |f(0,2)(x) - \hat{f}(0,2)(x)| \right).$$

So far, filaments are defined as a set on the plane. But one can also think about filaments that are projected onto the surface of the functions, that is a set in the three-dimensional space. Considering the sets $A = \{(x, \hat{f}(x)) : x \in \hat{L}\}$, $B = \{(x, f(x)) : x \in L\}$, we can obtain $\text{Haus}(A, B) \approx \text{Haus}(L, \hat{L})$. To see this, one first notices that the extra term $\sup_{x \in \hat{L}} \inf_{y \in L} |\hat{f}(x) - f(y)|$ shows up in the upper bound of $d(A|B)$. Let $x_0 \in \hat{L}$ as in the proof and note that

$$\inf_{y \in L} |\hat{f}(x_0) - f(y)| \leq |\hat{f}(\Upsilon x_0(0)) - f(\Upsilon x_0(t))|$$

$$\leq |\hat{f}(\Upsilon x_0(0)) - f(\Upsilon x_0(0))| + |f(\Upsilon x_0(0)) - f(\Upsilon x_0(t))|,$$

but then the last line is bounded by $\|\hat{f} - f\|_\infty + (\sup_x \|\nabla f(x)\|)\|\Upsilon x_0(t) - \Upsilon x_0(0)\|$. This indicates $d(A|B) \approx d(\hat{L}|L)$ and hence the claim.

In view of Theorems 3.5.4, 3.5.5 and Proposition 3.5.6, we have the following result.

**Theorem 3.5.8** Under Assumptions (A1)-(A5), with $J_1 \asymp J_2 \asymp (n/\log n)^{1/(2(\alpha+1))}$, and for $M_n \to \infty$,

$$\Pi(\text{Haus}(L, L^*) > M_n(n/\log n)^{\frac{2-\alpha}{2(\alpha+1)}} |D_n) \overset{P_0}{\rightarrow} 0,$$

$$\Pi(\text{Haus}(L, \hat{L}) > M_n(n/\log n)^{\frac{2-\alpha}{2(\alpha+1)}} |D_n) \overset{P_0}{\rightarrow} 0,$$

and

$$\text{Haus}(\hat{L}, L^*) = O_p\left((n/\log n)^{\frac{2-\alpha}{2(\alpha+1)}}\right).$$
Remark 3.5.4 The third assertion of Theorem 3.5.8 for the convergence rate of the filament induced by the posterior mean is an improvement over the rate in Theorem 5 of Genovese et al. (2014) when $\alpha > 4$. By the argument following Lemma 3.5.7, these results are directly applicable to the filaments projected on the surface of the functions.

For the following result, we use $k \in \{1, 2, 3\}$ instead of $r$ as the index for the order of derivatives, that is $f^{(1)} = f^{(2,0)}$, $f^{(1)} = f^{(1,1)}$ and $f^{(1)} = f^{(0,2)}$. Similarly, we define $A^{(k)}, C^{(k)}, \Sigma^{(k)}$. Let $\tilde{f}^{(k)} := A^{(k)}Y + C^{(k)}\theta_0$ be the posterior mean of $f^{(k)}$ and $\tilde{L}$ be the induced filament. For some $0 < \gamma < 1/2$, let $R_{n,k,\gamma}$ denotes the $1-\gamma$ quantile of the posterior distribution of $\|f^{(k)} - \tilde{f}^{(k)}\|_\infty$.

Let $C_{f,k,\gamma} := \{f : \|f^{(k)} - \tilde{f}^{(k)}\|_\infty \leq \rho R_{n,k,\gamma}\}$ for some large $\rho > 0$.

The following theorem provides two valid credible sets with sufficiently high frequentist coverage as the sample size increases.

**Theorem 3.5.9** Assume (A1)–(A5), for the choice of $J_1 \asymp J_2 \asymp (n/\log n)^{1/(\alpha+1)}$ and some $\rho > 0$ sufficiently large, for the following two sets,

$$C_L = \{R(f) : f \in \cap_{k=1}^3 C_{f,k,\gamma}\},$$

$$\bar{C}_L = \{L : \text{Haus}(\mathcal{L}, \tilde{L}) \leq \frac{C}{\eta} \rho \max_{1\leq k \leq 3} R_{n,k,\gamma}\},$$

the credibility of $C_L$ and its coverage probability for $\mathcal{L}^*$ tend to 1 and $C_L \subset \bar{C}_L$ with high posterior probability with $P_0$-probability tending to 1.

### 3.6 Simulation

Many algorithms have been proposed to find filaments. We here use the Subspace Constrained Mean Shift (SCMS) algorithm proposed in Ozertem and Erdogmus (2011). The algorithm is also used in Genovese et al. (2014), Chen, Genovese and Wasserman (2015) and Chen, Ho, Brinkmann, Freeman, Genovese, Schneider and Wasserman (2015). Even though in the previous study, the algorithm is based on kernel density estimator, our study suggests that nothing hinders the efficacy of the algorithm when applied on a series based expansion in either regression framework or in density estimation. In the following, we give a description of Subspace Constrained Mean Shift Algorithm.

**Algorithm** (Subspace Constrained Mean Shift Algorithm)

Set $\epsilon > 0$, $\tau > 0$ and select a collection of points $\{x_1, \ldots, x_n\}$, compute $f(x_i)$ and keep only those points for which $f(x_i) > \tau$. Now for each $x_i$, let $x_i^{(1)} = x_i$. Now iterate through the following steps starting from $t = 1$: 

65
(1) evaluate $\nabla f(x^{(t)}_i)$;

(2) evaluate the Hessian $Hf(x^{(t)}_i)$ and perform spectral decomposition to get $V(x^{(t)}_i)$ the normalized eigenvector of $Hf(x^{(t)}_i)$ with the smallest eigenvalue;

(3) update $x^{(t+1)}_i = \bar{a}V(x^{(t)}_i)V^T(x^{(t)}_i)\nabla f(x^{(t)}_i) + x^{(t)}_i$;

(4) stop if $\|x^{(t+1)}_i - x^{(t)}_i\| < \epsilon$ or $|V^T(x^{(t)}_i)\nabla f(x^{(t)}_i)| < \epsilon$.

If in Step 3, one instead updates $x^{(t+1)}_i = \bar{a}\nabla f(x^{(t)}_i) + x^{(t)}_i$, then it is just the celebrated mean shift algorithm which can be used to find the mode of the function $f$ (the terminology “mean shift” can be better understood when $\hat{f}$ is estimated using a kernel density estimator, as then $\nabla f(x^{(t)}_i)$ is proportional to the “mean-shift vector”).

The simulation setup will be similar to the one given in Section 2.4.2. We consider the same function

$$f(x_1, x_2) = 1 + \left(\phi(\sqrt{x_1^2 + x_2^2})\right)^{1+\cos^2(\tan^{-1}(x_2/x_1))},$$

where $\phi(\cdot)$ is the normal density function with mean 0.5 and standard deviation 0.3. The sample size is 2000. The Figure 3.2 shows the function $f$ and its filament. We use fifth-order B-splines functions, that is $q_1 = q_2 = 5$. We choose the pair $(J_1, J_2) = (9, 9)$ by their posterior mode.

![Figure 3.2: The function $f$ and its filament.](image)

We use $\tau = 2$, $\bar{a} = 0.02$ and $\epsilon = 10^{-6}$. Different choices of $J_1$ and $J_2$ have been experimented. The general principle seems to agree with that of the level set estimation. That is, “oversmoothing” may distort the filaments, whereas “undersmoothing” seems to produce similar results as using $J_1 = J_2 = 9$, as can be seen in Figure 3.3. We also provide uncertainty quantification in Figure 3.4 with $\gamma = 0.1, \rho = 1.2$. Each graph shows 100 filaments drawn the induced posterior
distribution for $C_L$. To evaluate $R_{n,k,\gamma}$, we first draw 200 posterior samples $\theta$, and compute their posterior mean $\bar{\theta}$. Next we compute $\sup_x | b_{j_1,j_2}^{(k)}(x)^T (\theta - \bar{\theta}) |$ by searching on a crude grid and pick the maximum point on the grid and then starting from this maximum point apply the gradient ascent or descent method to check if nearby points can achieve greater (absolute) value. We keep the largest value as the supremum. The $(1-\gamma)$-empirical quantile over all these suprema will then be the cut-off $R_{n,k,\gamma}$. The posterior filaments that fall in the set $C_L$ can then be generated.

![Images](image1.png)

Figure 3.3: Filaments of a nonparametric regression function (effects of the smoothing parameter $J_1$ and $J_2$). The orange circle is the truth, the blue curve is the estimated filament induced by the posterior mean. Top left: $J_1 = J_2 = 7$; top right: $J_1 = J_2 = 8$; bottom left: $J_1 = J_2 = 9$; bottom right: $J_1 = J_2 = 15$.

To evaluate the coverage performance, we compute the Hausdorff distance between $L^*$ and $\tilde{L}$. From the definition of $C_L$, we set $C/\eta$ to different values. Simulation shows that $L^*$ belongs to $C_L$ for 91%, 94%, 98% time when $C/\eta$ takes value $7.3 \times 10^{-4}$, $7.5 \times 10^{-4}$ and $8 \times 10^{-4}$ respectively. In general, $C$ can be computed as $\sup_x \| \nabla \hat{f}(x) \|$ and $\eta$ as the smallest value of $-\lambda$ along the filament induced by the posterior mean. With this method, we obtain 100% coverage — high
coverage as the theory predicts.

Figure 3.4: Filaments of a nonparametric regression function (uncertainty Quantification). Left: $J_1 = J_2 = 7$. Right: $J_1 = J_2 = 9$.

3.7 Application

For application, we use an earthquake dataset for California and its vicinity from January 1st of 2013 to December 31th of 2017 with magnitude 3.0 and above on the Richter scale \(^1\). The dataset consists of 3772 observations, among which 3383 observations have magnitude between 3 and 4; 355 observations between 4 and 5; 34 observations above 5. The average magnitude is 3.439. The left panel in Figure 3.5 shows the data scatter plot. The sizes of circles are proportional to the magnitudes of the earthquakes.

In SCMS algorithm, we use $\bar{a} = 5 \times 10^{-6}$ and $\tau = 3$ and $\epsilon = 10^{-6}$. We use $q_1 = q_2 = 4$ and $J_1 = J_2 = 32$. We draw 200 posterior samples to compute the posterior mean. The filament induced by the posterior mean is plotted as the blue curve in Figure 3.5. The same filament is overlayed on the magnitude surface as given in Figure 3.6. To obtain uncertainty quantification using filaments from posterior samples, we use $\gamma = 0.1$, $\rho = 1.2$. The results showed that 91% of posterior samples fall into $C_L$, only slightly higher than the nominal credibility level. We randomly pick 100 of them to plot the uncertainty quantification as given in the right panel of Figure 3.5.

The filaments hence obtained provide useful characterization of the features of earthquake magnitude. Geographically, these filaments pass through the most populous coastal urban and suburban areas in California, for instance, Eureka city, San Francisco and Los Angeles. Since this application has utilized very small portion of earthquake data, we believe that a large scale

\(^1\)The data is publicly available from https://earthquake.usgs.gov/earthquakes/.
study for different periods of historical times will be useful for the study of the dynamics of the earthquakes. Uncertainty quantification regions provide statistical understanding of what might be a reasonable shift of these filaments through spatial and temporal domain and are also helpful for discovering newly emerging crustal activities.

Figure 3.5: Left: Earthquake data points (in red). Right: 100 filaments in the posterior constructed with high frequentist coverage. The thick blue curves are the filaments induced by the posterior mean.

Figure 3.6: The magnitude surface and filaments induced by the posterior mean.
3.8 Technical Proofs

In this section, we will provide lemmas and formal proofs of the results stated in the main text. To focus on discussion, we will draw on several useful results directly given in the following remark.

**Remark 3.8.1** Under the assumptions (A1) to (A5), the following assertions hold.

1. Integral curves are dense and non-overlapping.
2. \( x_0 \mapsto t^*_{x_0} \) is continuous.
3. \( \| \nabla \nabla f(x) - \nabla f^*(x) \| \leq C \exp(t) \| x_0 - x_0' \| \).
4. \( G = (G_1, G_2)^T, \nabla G \) are Lipschitz continuous and each element of Hessian \( HG_1, HG_2 \) is bounded on some open set \( Q_\delta \subset \mathbb{R}^3 \) such that \( \{ d^2 f^*(x) : x \in [0,1]^2 \} \subset Q_\delta \). Thus \( V^* \) is Lipschitz continuous.

There results can be proved using the arguments similar to Qiao and Polonik (2016). Original argument and proofs appear in that paper in various places. To save space, we do not provide the details here, but point to these following specific places in their paper. Result (1) is given in the discussion in page 10, result (2) and (3) are discussed in pages 22 and 48, and result (4) is discussed in pages 53–55.

3.8.1 Some lemmas

**Lemma 3.8.1** (*Posterior contraction around truth*) Under above assumptions, with \( J_1 \asymp J_2 \asymp J \) and \( J^2 \leq n \), chosen such that \( \delta_{n,k,J} \) vanishes to zero, for any \( M_n \to \infty \),

\[
\Pi\left( \sup_{x \in [0,1]^2} |D^r f(x) - D^r f^*(x)| > M_n \delta_{n,r,J} |D_n| \right) \xrightarrow{P_0} 0,
\]

\[
\Pi\left( \sup_{x \in [0,1]^2} \| \nabla f(x) - \nabla f^*(x) \| > M_n \delta_{n,1,J} |D_n| \right) \xrightarrow{P_0} 0,
\]

\[
\Pi\left( \sup_{x \in [0,1]^2} \| d^2 f(x) - d^2 f^*(x) \| > M_n \delta_{n,2,J} |D_n| \right) \xrightarrow{P_0} 0,
\]

\[
\Pi\left( \sup_{x \in [0,1]^2} \| \nabla f(x) - \nabla f^*(x) \| \xrightarrow{F} > M_n \delta_{n,2,J} |D_n| \right) \xrightarrow{P_0} 0,
\]

70
\[ \Pi( \sup_{x \in [0,1]^2} \| \nabla d^2 f(x) - \nabla d^2 f^*(x) \|_F > M_n \delta_{n,3,J} \| \mathbb{D}_n \|) \xrightarrow{P_0} 0, \]
\[ \Pi( \sup_{x \in [0,1]^2} \| V(x) - V^*(x) \|_F > M_n \delta_{n,2,J} \| \mathbb{D}_n \|) \xrightarrow{P_0} 0, \]
\[ \Pi( \sup_{x \in [0,1]^2} \| \nabla V(x) - \nabla V^*(x) \|_F > M_n \delta_{n,3,J} \| \mathbb{D}_n \|) \xrightarrow{P_0} 0, \]

where \( \delta_{n,k,J} = J^k ((\log n/n) J^2 + J^{-2a})^{1/2}. \)

**Proof.** These results can be directly adapted from Yoo and Ghosal (2016) by noting the highest degree of derivatives in each expression. We shall show the rates for \( \sup_x \| H f(x) - H f^*(x) \|_F, \) \( \sup_x \| \nabla d^2 f(x) - \nabla d^2 f^*(x) \|_F \) and \( \sup_x \| \nabla V(x) - \nabla V^*(x) \|_F. \) Notice that

\[
\begin{align*}
\| H f(x) - H f^*(x) \|_F &= \left( (f^{(2,0)}(x) - f^{*(2,0)}(x))^2 + 2(f^{(1,1)}(x) - f^{*(1,1)}(x))^2 + (f^{(0,2)}(x) - f^{*(0,2)}(x))^2 \right)^{1/2}, \\
&\leq 4^{-1/2} \left( |f^{(2,0)}(x) - f^{*(2,0)}(x)| + 2|f^{(1,1)}(x) - f^{*(1,1)}(x)| + |f^{(0,2)}(x) - f^{*(0,2)}(x)| \right).
\end{align*}
\]

The rate for \( \sup_x \| H f(x) - H f^*(x) \|_F \) then follows easily.

Also, the contraction rate for \( \| \nabla d^2 f(x) - \nabla d^2 f^*(x) \|_F \) follows from the inequality \( \| \nabla d^2 f(x) - \nabla d^2 f^*(x) \|_F \leq 6^{-1/2} \sum_{k=1}^r |D^r f(x) - D^r f^*(x)|. \)

Lastly, since \( \nabla V(x) - \nabla V^*(x) = \nabla G(d^2 f(x)) - \nabla G(d^2 f^*(x)) \) \( \nabla d^2 f(x) - \nabla d^2 f^*(x), \) which is a 2 \times 2 matrix. Straightforward calculation gives its (1,1)-th element

\[
\begin{align*}
\left( G^{(1,0,0)}_1(d^2 f(x)) f^{(3,0)}(x) - G^{(1,0,0)}_1(d^2 f^*(x)) f^{*(3,0)}(x) \right), \\
+ \left( G^{(0,1,0)}_1(d^2 f(x)) f^{(2,1)}(x) - G^{(0,1,0)}_1(d^2 f^*(x)) f^{*(2,1)}(x) \right), \\
+ \left( G^{(0,0,1)}_1(d^2 f(x)) f^{(1,2)}(x) - G^{(0,0,1)}_1(d^2 f^*(x)) f^{*(1,2)}(x) \right).
\end{align*}
\]

The absolute value of the first summand is bounded by the sum of

\[
\left| G^{(1,0,0)}_1(d^2 f(x)) f^{(3,0)}(x) - G^{(1,0,0)}_1(d^2 f(x)) f^{*(3,0)}(x) \right|,
\]

and

\[
\left| G^{(1,0,0)}_1(d^2 f(x)) f^{*(3,0)}(x) - G^{(1,0,0)}_1(d^2 f^*(x)) f^{*(3,0)}(x) \right|.
\]

Noting that \( d^2 f(x) \) contracts to \( d^2 f^*(x) \) uniformly in \( x, \) hence \( \{d^2 f(x) : x \in [0,1]^2\} \subset Q_\delta \) with posterior probability tending to 1. The first term is bounded by a constant multiple of \( |f^{(3,0)}(x) - f^{*(3,0)}(x)|, \) in view of the result 4 of Remark 3.8.1. By the same remark and the
assumption \( \| f^* \|_{\alpha, \infty} < \infty \), the second term is bounded by \( \| d^2 f(x) - d^2 f^*(x) \| f^{*,(3,0)}(x) \). Using similar arguments for the second and third summand, one can see that the absolute value of the (1,1) element of \( \nabla V(x) - \nabla V^*(x) \) is bounded by
\[
|f^{(3,0)}(x) - f^{*,(3,0)}(x)| + |f^{(2,1)}(x) - f^{*,(2,1)}(x)| + |f^{(1,2)}(x) - f^{*,(1,2)}(x)|.
\]
Dealing the rest elements of \( \nabla V(x) - \nabla V^*(x) \) similarly, we can see that \( \| \nabla V(x) - \nabla V^*(x) \|_F \lesssim \sum_{r, |r| = 3} |D^r f - D^r f^*| \). □

In above lemma, the optimal rates are obtained when \( J \asymp (n/\log n)^{1/(2(\alpha+1))} \), which then yields
\[
\delta_{n,k,J} = \epsilon_n \asymp (\log n/n)^{(\alpha-k)/(2\alpha+2)}.
\]

In addition, we have the following two lemmas whose proofs closely follow that of Lemma 3.8.1 and thus are omitted. Denote the posterior mean of \( f \) by \( \tilde{f} := AY + C\theta_0 \) and similarly define the quantities it induces, for instance, \( \tilde{V}, \tilde{\gamma}_x \) and \( \tilde{L} \). We then have the following lemma.

**Lemma 3.8.2 (Posterior contraction around the posterior mean)** Under above assumptions and \( J_1 \asymp J_2 \asymp J \) and \( J^2 \leq n \), with \( J \) is chosen such that \( \eta_{n,k,J} \) vanishes to zero, for any \( M_n \to \infty \),
\[
\Pi( \sup_{x \in [0,1]^2} |D^r f(x) - D^r \tilde{f}(x)| > M_n \eta_{n,J} |D_n| \mathbb{P}_0 \to 0, \quad \Pi( \sup_{x \in [0,1]^2} \|Df(x) - D\tilde{f}(x)\| > M_n \eta_{n,J} |D_n| \mathbb{P}_0 \to 0, \quad \Pi( \sup_{x \in [0,1]^2} \|d^2 f(x) - d^2 \tilde{f}(x)\| > M_n \eta_{n,J} |D_n| \mathbb{P}_0 \to 0, \quad \Pi( \sup_{x \in [0,1]^2} \|Hf(x) - H\tilde{f}(x)\|_F > M_n \eta_{n,J} |D_n| \mathbb{P}_0 \to 0, \quad \Pi( \sup_{x \in [0,1]^2} \|d^2 f(x) - d^2 \tilde{f}(x)\|_F > M_n \eta_{n,J} |D_n| \mathbb{P}_0 \to 0, \quad \Pi( \sup_{x \in [0,1]^2} \|V(x) - \tilde{V}(x)\|_F > M_n \eta_{n,J} |D_n| \mathbb{P}_0 \to 0, \quad \Pi( \sup_{x \in [0,1]^2} \|\nabla V(x) - \nabla \tilde{V}(x)\|_F > M_n \eta_{n,J} |D_n| \mathbb{P}_0 \to 0, \quad \Pi( \sup_{x \in [0,1]^2} \|\nabla V(x) - \nabla \tilde{V}(x)\|_F > M_n \eta_{n,J} |D_n| \mathbb{P}_0 \to 0,
\]
where \( \eta_{n,k,J} = J^{k+1}(\log n/n)^{1/2} \).

One can also readily obtain the following convergence rates for the Bayesian estimators induced by \( \tilde{f} \).
Lemma 3.8.3 (Convergence of posterior mean) Under the above assumptions and $J_1 \asymp J_2 \asymp J$ and $J^2 \leq n$, with $J$ is chosen such that $\delta_{n,k,J}$ vanishes to zero, then we have

\[
\sup_{x \in [0,1]^2} |D^r \hat{f}(x) - D^r f^*(x)| = O_p(\delta_{n,|r|,J}),
\]
\[
\sup_{x \in [0,1]^2} \|\nabla \hat{f}(x) - \nabla f^*(x)\| = O_p(\delta_{n,1,J}),
\]
\[
\sup_{x \in [0,1]^2} \|d^2 \hat{f}(x) - d^2 f^*(x)\| = O_p(\delta_{n,2,J}),
\]
\[
\sup_{x \in [0,1]^2} \|H \hat{f}(x) - H f^*(x)\|_F = O_p(\delta_{n,2,J}),
\]
\[
\sup_{x \in [0,1]^2} \|\nabla d^2 \hat{f}(x) - \nabla d^2 f^*(x)\|_F = O_p(\delta_{n,3,J}),
\]
\[
\sup_{x \in [0,1]^2} \|\tilde{V}(x) - V^*(x)\|_F = O_p(\delta_{n,2,J}),
\]
\[
\sup_{x \in [0,1]^2} \|\nabla \tilde{V}(x) - \nabla V^*(x)\|_F = O_p(\delta_{n,3,J}),
\]

where $\delta_{n,k,J} = J^k \left((\log n/n)J^2 + J^{-2\alpha}\right)^{1/2}$.

3.8.2 Proofs of the main results

Proof of Theorem 3.5.1. We first sketch a proof for the following result as first proved in Koltchinskii et al. (2007):

\[
\sup_{x_0 \in \mathcal{G}} \sup_{t \in [0,T_{x_0}]} \|\Psi_{x_0}(t) - \Psi_{x_0}^*(t)\| \lesssim \sup_{x_0 \in \mathcal{G}} \sup_{t \in [0,T_{x_0}]} \left\| \int_0^t (V - V^*)(\Psi_{x_0}^*(s))ds \right\|.
\]

To see this, let

\[
y_{x_0}(t) := \Psi_{x_0}(t) - \Psi_{x_0}^*(t) = \int_0^t (V(\Psi_{x_0}(s)) - V^*(\Psi_{x_0}^*(s)))ds,
\]
\[
z_{x_0}(t) := \int_0^t (V - V^*)\Psi_{x_0}^*(s)ds + \int_0^t \nabla V^*(\Psi_{x_0}^*(s))z_{x_0}(s)ds.
\]

Note that $\delta_{x_0}(t) := y_{x_0}(t) - z_{x_0}(t) = \int_0^t \nabla V^*(\Psi_{x_0}^*(s))\delta(s)ds + R_{x_0}(t)$ for some reminder term $R_{x_0}(t)$. So $\|\delta_{x_0}(t)\| \leq \|R_{x_0}(t)\| + \int_0^t \|\nabla V^*(\Psi_{x_0}^*(s))\|_F\|\delta_{x_0}(s)\|ds$ and by the Gronwall-Bellman inequality (Kim; 2009),

\[
\|\delta_{x_0}(t)\| \leq \sup_{t \in [0,T_{x_0}]} \left(\|R_{x_0}(t)\| \exp \left( \int_0^t \|\nabla V^*(\Psi_{x_0}^*(s))\|_Fds \right) \right).
\]
Hence \( \sup_{t \in [0, T^*_{x_0}]} \| \delta_{x_0}(t) \| \) \( \lesssim \) \( \sup_{t \in [0, T^*_{x_0}]} \| R_{x_0}(t) \| \). It can be shown that with high posterior probability \( \sup_{t \in [0, T^*_{x_0}]} \| R_{x_0}(t) \| \ll \int_{0}^{T^*_{x_0}} \| y_{x_0}(s) \| ds \) in \( P_0 \)-probability tending to 1 following the argument in page 1586 of Koltchinskii et al. (2007). Therefore, \( \sup_{t \in [0, T^*_{x_0}]} \| \delta_{x_0}(t) \| \ll \int_{0}^{T^*_{x_0}} \| y_{x_0}(s) \| ds \). Since \( y_{x_0}(t) = \delta_{x_0}(t) + z_{x_0}(t) \), it follows that

\[
\sup_{t \in [0, T^*_{x_0}]} \| \delta_{x_0}(t) \| \ll \int_{0}^{T^*_{x_0}} \| z_{x_0}(s) \| ds \lesssim \sup_{t \in [0, T^*_{x_0}]} \| z_{x_0}(t) \|.
\]

Then

\[
\sup_{t \in [0, T^*_{x_0}]} \| y_{x_0}(t) \| = \sup_{t \in [0, T^*_{x_0}]} \| \delta_{x_0}(t) + z_{x_0}(t) \| \lesssim \sup_{t \in [0, T^*_{x_0}]} \| z_{x_0}(t) \|.
\]

But

\[
\| z_{x_0}(t) \| \leq \left\| \int_{0}^{t} (V - V^*) \mathcal{Y}_{x_0}^*(s)ds \right\| + \int_{0}^{t} \| \nabla V^*(\mathcal{Y}_{x_0}^*(s)) \| F \| z_{x_0}(s) \| ds.
\]

One more application of the Gronwall-Bellman inequality and then taking the supremum on the left hand side yields

\[
\sup_{t \in [0, T^*_{x_0}]} \| z_{x_0}(t) \| \leq \sup_{t \in [0, T^*_{x_0}]} \left( \left\| \int_{0}^{t} (V - V^*) \mathcal{Y}_{x_0}^*(s)ds \right\| \exp \left( \int_{0}^{t} \| \nabla V^*(\mathcal{Y}_{x_0}^*(s)) \| F ds \right) \right)
\]

\[
\lesssim \sup_{t \in [0, T^*_{x_0}]} \left\| \int_{0}^{t} (V - V^*) \mathcal{Y}_{x_0}^*(s)ds \right\|.
\]

Taking another supremum over all \( x_0 \) gives us the result.

Now coming back to the main proof, by an integral form Taylor expansion of \( V(x) (= G(d^2 f(x)) \) around \( G(d^2 f^*(x)) \), by the uniform boundedness of each element of second derivative \( G \), one can get

\[
\sup_{x_0 \in \mathcal{G}} \left\| \int_{0}^{t} [(V - V^*)(\mathcal{Y}_{x_0}^*(s)) - \nabla G(d^2 f^*(\mathcal{Y}_{x_0}^*(s)))d^2(f - f^*)(\mathcal{Y}_{x_0}^*(s))] ds \right\|
\]

is bounded by a constant multiple of \( \sup_{x_0} \| d^2 f(x) - d^2 f^*(x) \|^2 \). In view of Lemma 3.8.1, with the choice of \( J \) in the assumption, \( \sup_x \| d^2 f(x) - d^2 f^*(x) \|^2 \) is of the order \( (n/ \log n)^{(5 - \alpha)/(1 + 2\alpha)} \). Now it suffices to prove that

\[
\sup_{x_0 \in \mathcal{G}} \left\| \int_{0}^{t} \nabla G(d^2 f^*(\mathcal{Y}_{x_0}^*(s)))d^2(f - f^*)(\mathcal{Y}_{x_0}^*(s))ds \right\|
\]
has posterior contraction rate \( (n/ \log n)^{(2-\alpha)/(1+2\alpha)} \). To this end, we demonstrate that

\[
\sup_{x_0 \in \mathcal{G}} \left| \int_0^t \frac{\partial G_1}{\partial x_1} \left( \frac{d^2 f^* (\mathcal{T}_{x_0}^* (s))}{d^2} \right) (f^{(2,0)} - f^{(2,0)*}) (\mathcal{T}_{x_0}^* (s)) ds \right|.
\]

has this posterior contraction rate. The rate for the other components can be derived similarly. Let \( \mathcal{U}_n \) be a shrinking neighborhood of \( \sigma_0^2 \) such that \( \hat{\sigma}_n^2 \in \mathcal{U}_n \) and \( \Pi(\sigma^2 \in \mathcal{U}_n | \mathbb{D}_n) \to 1 \) with probability 1.

Let \( \tilde{G}(\mathcal{T}_{x_0}^* (s)) = \frac{\partial G_1}{\partial x_1} (d^2 f^* (\mathcal{T}_{x_0}^* (s))) \). To derive the rate, it suffices to show the rate for

\[
E_0 \left[ \sup_{\sigma^2 \in \mathcal{U}_n} \mathbb{E} \left( \sup_{x_0, t} \left| \int_0^t \tilde{G}(\mathcal{T}_{x_0}^* (s)) (f^{(2,0)} - f^{(2,0)*}) (\mathcal{T}_{x_0}^* (s)) ds \right|^2 | \mathbb{D}_n, \sigma^2 \right) \right] \tag{3.2}
\]

decays as \( (n/ \log n)^{(4-2\alpha)/(1+2\alpha)} \), which then by Markov’s inequality yields the desired posterior contraction rate. Notice that the posterior variance of \( f^{(2,0)}(x) \) does not depend on \( \mathbb{D}_n \), while its posterior mean \( A_{(2,0)} Y + C_{(2,0)} \theta_0 \) does not depend on \( \sigma^2 \). Also, in the posterior distribution conditional on \( \sigma^2 \), we have \( f^{(2,0)}(x) - A_{(2,0)} (x) Y - C_{(2,0)} (x) \theta_0 = \sigma b_{j_1, j_2}^{(2,0)} (x) (B^T B + \Lambda_0^{-1})^{-1/2} Z \), where \( Z \) is a \( J^2 \times 1 \) vector of standard normal random variables. Thus we can write

\[
E_0 \left[ \sup_{\sigma^2 \in \mathcal{U}_n} \mathbb{E} \left( \sup_{x_0, t} \left| \int_0^t \tilde{G}(\mathcal{T}_{x_0}^* (s)) (A_{(2,0)} (x) Y + C_{(2,0)} \theta_0 - f^{(2,0)*}) (\mathcal{T}_{x_0}^* (s)) ds \right|^2 | \mathbb{D}_n, \sigma^2 \right) \right] \tag{3.3}
\]

\[
\leq E_0 \left[ \sup_{\sigma^2 \in \mathcal{U}_n} \mathbb{E} \left( \sup_{x_0, t} \left| \int_0^t \tilde{G}(\mathcal{T}_{x_0}^* (s)) b_{j_1, j_2}^{(2,0)} (\mathcal{T}_{x_0}^* (s)) ds (B^T B + \Lambda_0^{-1})^{-1/2} Z \right|^2 \right) \right] \tag{3.4}
\]

\[
+ E_0 \left( \sup_{x_0, t} \left| \int_0^t \tilde{G}(\mathcal{T}_{x_0}^* (s)) A_{(2,0)} (\mathcal{T}_{x_0}^* (s)) ds \right|^2 \right) \tag{3.5}
\]

We shall tackle each of above three terms one by one. Let

\[
a(x_0, t) = \int_0^t \tilde{G}(\mathcal{T}_{x_0}^* (s)) b_{j_1, j_2}^{(2,0)} (\mathcal{T}_{x_0}^* (s)) ds,
\]

where the integral is taken elementwise. Now consider the first term (3.3). Define \( U_{1,n}(x_0, t) = \)
\(a(x_0, t)(B^T B + \Lambda_0^{-1})^{-1/2}Z\) which is Gaussian. For fixed \(x_0, t\), we have

\[
E[U_{1,n}(x_0, t)]^2 = a(x_0, t)^T (B^T B + \Lambda_0^{-1})a(x_0, t) \lesssim \frac{J_2}{n} \|a(x_0, t)\|^2 \lesssim \frac{J_5}{n},
\]

where last step follows from Lemma 3.8.4. Consider the two points \((x_0, t)\) and \((\tilde{x}_0, \tilde{t})\), where the distance

\[
d((x_0, t), (\tilde{x}_0, \tilde{t})) := \sqrt{\text{Var}(U_{1,n}(x_0, t) - U_{1,n}(\tilde{x}_0, \tilde{t}))} \]

\[
\lesssim \left(\frac{J_2}{n}\right)^{1/2} \|a(x_0, t) - a(\tilde{x}_0, \tilde{t})\| \]

\[
\lesssim \left(\frac{J_7}{n} (|t - \tilde{t}| + \|x_0 - \tilde{x}_0\|^2)\right)^{1/2};
\]

here the last line follows from Lemma 3.8.4. Let \(\tilde{d}((x_0, t), (\tilde{x}_0, \tilde{t})) := (|t - \tilde{t}| + \|x_0 - \tilde{x}_0\|^2)^{1/2}.\) Thus

\[
d((x_0, t), (\tilde{x}_0, \tilde{t})) \lesssim \rho_n \tilde{d}((x_0, t), (\tilde{x}_0, \tilde{t})) \text{ for } \rho_n = \frac{J_7/2}{n^{1/2}}.\] By Lemma A.11 of Yoo and Ghosal (2016), setting \(\delta_n \asymp 1/J\) there, we obtain \(E\left(\sup_{x_0, t}[U_{1,n}(x_0, t)]^2\right) = O(J_5 \log n / n).\) Therefore, (3.3) is of the order \((\log n) J_5 / n.\)

For the term (3.5), let

\[
U_{2,n}(x_0, t) = \int_0^t \tilde{G}(Y_{x_0}(s))A_{(2,0)}(Y_{x_0}^*(s))\varepsilon ds = a(x_0, t)(B^T B + \Lambda_0^{-1})^{-1}B^T \varepsilon.
\]

Observe that for fix \(x_0, t\), we have

\[
E[U_{2,n}(x_0, t)]^2 = \sigma_0^2 a(x_0, t)^T (B^T B + \Lambda_0^{-1})^{-1}B^T B(B^T B + \Lambda_0^{-1})^{-1}a(x_0, t)
\]

\[
\lesssim \sigma_0^2 \frac{J_2}{n} \frac{J^2}{n} a(x_0, t)^T a(x_0, t)
\]

\[
\lesssim \sigma_0^2 \frac{J^5}{n}.
\]

Since \(\varepsilon\) is sub-Gaussian, \(U_{2,n}(x_0, t)\) is sub-Gaussian with mean 0 and variance \(E[U_{2,n}(x_0, t)]^2\). By the same argument, \(E\left(\sup_{x_0, t}[U_{2,n}(x_0, t)]^2\right) = O(J_5 \log n / n).\) Therefore, (3.5) is of the order \((\log n) J_5 / n.\)
Finally, for (3.4), write

\[
\left| \int_0^t \tilde{G}(Y_{x_0}^*(s)) \left( A_{(2,0)}(Y_{x_0}^*(s)) + C_{(2,0)}(Y_{x_0}^*(s)) \theta_0 - f^{(2,0)*}(Y_{x_0}^*(s)) \right) ds \right|
\]

or

\[
= \left| \int_0^t \tilde{G}(Y_{x_0}^*(s)) b^{(2,0)}_{j_1,j_2} (T_{x_0}^*(s)) (B^T B + \Lambda_0^{-1})^{-1} \left( B^T (F^* - B\theta_\infty) + \Lambda_0^{-1}(\theta_0 - \theta_\infty) \right) ds \right|
\]

The first term on the right hand side of above equation is just

\[
\left| a(x_0, t)^T (B^T B + \Lambda_0^{-1})^{-1} \left( B^T (F^* - B\theta_\infty) + \Lambda_0^{-1}(\theta_0 - \theta_\infty) \right) \right|
\]

which is bounded by

\[
\|a(x_0, t)\|_1 \|(B^T B + \Lambda_0^{-1})^{-1}\|_{\infty,\infty} \left( \|B^T (F^* - B\theta_\infty)\|_{\infty} + \|\Lambda_0^{-1}\|_{\infty,\infty} (\|	heta_0\|_{\infty} + \|	heta_\infty\|_{\infty}) \right).
\]

Note that \(\|a(x_0, t)\|_1 \lesssim J^2\) by Lemma 3.8.4. Since \(\|(B^T B + \Lambda_0^{-1})^{-1}\|_{\infty,\infty} = O(J^2/n)\) and \(\|B^T (F^* - B\theta_\infty)\|_{\infty} = O(n J^{-2-\alpha})\), this term is of order \((J^4/n) + J^{2-\alpha}\). The second term of the right hand side of above equation can be bounded by the supremum of the approximate error which is of the order \(J^{2-\alpha}\). Therefore, (3.4) is of the order \(J^4((J^4/n^2) + J^{-2\alpha})\).

Putting all the above terms together, (3.2) is of order \((\log n/n) J^5 + J^4((J^4/n^2) + J^{-2\alpha})\). For \(J \asymp (n/\log n)^{1/(1+2\alpha)}\), (3.2) is of order \((n/\log n)^{(4-2\alpha)/(1+2\alpha)}\), which then completes the proof. \(\blacksquare\)

**Lemma 3.8.4** For any \(x_0, \tilde{x}_0 \in G\), \(t \in [0, T_{x_0}]\) and \(\tilde{t} \in [0, T_{x_0}\tilde{t}]\), let \(r = (2,0), (1,1)\) or \((0,2), a(x_0, t) := \int_0^t \tilde{G}(Y_{x_0}^*(s)) b^{(r)}_{j_1,j_2} (Y_{x_0}^*(s)) ds\), where the integral is taken elementwise, under Assumption (A1), (A2) and (A5),

\[
\|a(x_0, t)\|_1 \lesssim J^2,
\|a(x_0, t)\|_2 \lesssim J^3,
\|a(x_0, t) - a(\tilde{x}_0, \tilde{t})\|_2 \lesssim J^3|t - \tilde{t}| + J^5\|x_0 - \tilde{x}_0\|^2.
\]

**Proof.** We shall consider only two separate cases (i) \(r = (2,0)\) and (ii) \(r = (1,1)\). In the proof, we may write \(\sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \) as \(\sum_{j_1,j_2}\).
(i). For \( r = (2, 0) \) (similarly for \( r = (0, 2) \)),

\[
\|a(x_0, t)\|_1 = \sum_{j_1, j_2} \left| \int_0^t \tilde{G}_1(s) B''_{j_1}(\Upsilon^*_{1,x_0}(s)) B_{j_2}(\Upsilon^*_{2,x_0}(s)) ds \right|
\]

\[
\leq \sum_{j_1, j_2} \int_0^t |B''_{j_1}(\Upsilon^*_{1,x_0}(s))| B_{j_2}(\Upsilon^*_{2,x_0}(s)) ds,
\]

the last line follows by \( \sum_{j_2=1}^{J_2} B_{j_2}(x_2) = 1 \) and noticing that

\[
B''_{j_1,q_1}(x) = \frac{(q_1 - 1)(q_1 - 2) B_{j_1,q_1-2}(x)}{(t_{j_1} - t_{j_1,q_1+1})(t_{j_1} - t_{j_1,q_1+2})} + \frac{(q_1 - 1)(q_1 - 2) B_{j_1-1,q_1-2}(x)}{(t_{j_1-1} - t_{j_1,q_1})(t_{j_1-1} - t_{j_1,q_1+1})}
- \frac{(q_1 - 1)(q_1 - 2) B_{j_1-1,q_1-2}(x)}{(t_{j_1-1} - t_{j_1,q_1+1})(t_{j_1-1} - t_{j_1,q_1+2})} + \frac{(q_1 - 1)(q_1 - 2) B_{j_1-2,q_1-2}(x)}{(t_{j_1-2} - t_{j_1-1,q_1})(t_{j_1-2} - t_{j_1-1,q_1+1})},
\]

which implies \( \sum_{j_1=1}^{J_1} |B''_{j_1}(\Upsilon^*_{1,x_0}(s))| \lesssim 4J^2 \). Therefore, \( \|a(x_0, t)\|_1 \lesssim J^2 \).

Let \( S_{j_1} = [t_{1,j_1-1}, t_{1,j_1}], S_{j_2} = [t_{2,j_2-1}, t_{2,j_2}] \) and \( 1_{j_1,j_2}(s) := 1 \{ s : \Upsilon^*_{x_0}(s) \in S_{j_1} \times S_{j_2} \} \).

Turn to \( \|a(x_0, t)\|^2 \), which is

\[
\sum_{j_1,j_2} \left( \int_0^t \tilde{G}_1(s) B''_{j_1}(\Upsilon^*_{1,x_0}(s)) B_{j_2}(\Upsilon^*_{2,x_0}(s)) ds \right)^2
\]

\[
= \int_0^t \int_0^t \tilde{G}_1(s) \tilde{G}_1(s') \sum_{j_1,j_2} \left( B''_{j_1}(\Upsilon^*_{1,x_0}(s)) B_{j_2}(\Upsilon^*_{2,x_0}(s)) B''_{j_1}(\Upsilon^*_{1,x_0}(s')) B_{j_2}(\Upsilon^*_{2,x_0}(s')) \right) ds ds'
\]

\[
\lesssim \int_0^t \int_0^t \sum_{j_1,j_2} \left( |B''_{j_1}(\Upsilon^*_{1,x_0}(s))| B_{j_2}(\Upsilon^*_{2,x_0}(s)) |B''_{j_1}(\Upsilon^*_{1,x_0}(s'))| B_{j_2}(\Upsilon^*_{2,x_0}(s')) \right) ds ds'
\]

\[
= \int_0^t \int_0^t \sum_{j_1,j_2} 1_{j_1,j_2}(s) 1_{j_1,j_2}(s') \left( |B''_{j_1}(\Upsilon^*_{1,x_0}(s))| B_{j_2}(\Upsilon^*_{2,x_0}(s)) |B''_{j_1}(\Upsilon^*_{1,x_0}(s'))| B_{j_2}(\Upsilon^*_{2,x_0}(s')) \right) ds ds'.
\]

The last equality is obtained as follows. Since for any fix \( j_1, j_2, B''_{j_1}(\cdot) \) is supported on \( S_{j_1} \) and \( B_{j_2}(\cdot) \) is supported on \( S_{j_2} \). So \( 1_{j_1,j_2}(s) 1_{j_1,j_2}(s') = 1 \{ (s, s') : \Upsilon^*_{x_0}(s) \in S_{j_1} \times S_{j_2} \text{ and } \Upsilon^*_{x_0}(s') \in S_{j_1} \times S_{j_2} \} \). Notice that for \( n \) large, \( |S_{j_1}| \asymp |S_{j_2}| \asymp J^{-1} \) and \( \Upsilon^*_{x_0}(s) \in S_{j_1} \times S_{j_2} \) and \( \Upsilon^*_{x_0}(s') \in S_{j_1} \times S_{j_2} \) implies that \( \|\Upsilon^*_{x_0}(s) - \Upsilon^*_{x_0}(s')\| \lesssim J^{-1} \). By Assumption (A5),

\[
C_g |s - s'| \leq |s - s'|((\Upsilon^*_{x_0}(s) - \Upsilon^*_{x_0}(s'))/(s - s')) \lesssim J^{-1},
\]

and hence \( |s - s'| \lesssim J^{-1} \). Therefore, for \( n \) large enough, above quantity can be further bounded
by a constant multiple of

$$\int_0^t \int_0^t \sum_{j_1, j_2} \mathbb{1}\{|s - s'| < CJ^{-1}\} \left( |B''_{j_1}(\mathbf{Y}^*_{1,x_0}(s))|B_{j_2}(\mathbf{Y}^*_2(s))|B''_{j_1}(\mathbf{Y}^*_{1,x_0}(s'))|B_{j_2}(\mathbf{Y}^*_2(s')) \right) ds ds'.$$

Noting $B_{j_2}(x_2) \leq 1$ and $\sum_{j_2=1}^{j_2} B_{j_2} = 1$, it can be further bounded by

$$\int_0^t \int_0^t \sum_{j_1, j_2} \mathbb{1}\{|s - s'| < CJ^{-1}\} \left( |B''_{j_1}(\mathbf{Y}^*_{1,x_0}(s))|B_{j_2}(\mathbf{Y}^*_2(s))|B''_{j_1}(\mathbf{Y}^*_{1,x_0}(s'))|B_{j_2}(\mathbf{Y}^*_2(s')) \right) ds ds' \lesssim \int_0^t \int_0^t \sum_{j_1} \mathbb{1}\{|s - s'| < CJ^{-1}\} \left( |B''_{j_1}(\mathbf{Y}^*_{1,x_0}(s))|B''_{j_1}(\mathbf{Y}^*_{1,x_0}(s')) \right) ds ds' \lesssim J^2J^{-1} \int_0^t \sum_{j_1=1}^{j_1} |B''_{j_1}(\mathbf{Y}^*_{1,x_0}(s))| ds.$$

From argument used in bounding $\|a(x_0, t)\|_1$, we have $\sum_{j_1=1}^{j_1} |B''_{j_1}(\mathbf{Y}^*_{1,x_0}(s))| \lesssim J^2$. This completes the proof for $\|a(x_0, t)\|_2 \lesssim J^3$.

For the third result, we write

$$\|a(x_0, t) - a(\bar{x}_0, \bar{t})\|^2 \lesssim \|a_1(x_0, t, \bar{t})\|^2 + \|a_2(\bar{t}, x_0, \bar{x}_0)\|^2,$$

where

$$a_1(x_0, t, \bar{t}) := \int_0^t \tilde{G}(\mathbf{Y}^*_x(s))b_{j_1,j_2}^{(2,0)}(\mathbf{Y}^*_x(s))ds - \int_0^{\bar{t}} \tilde{G}(\mathbf{Y}^*_x(s))b_{j_1,j_2}^{(2,0)}(\mathbf{Y}^*_x(s))ds,$$

and

$$a_2(\bar{t}, x_0, \bar{x}_0) := \int_0^{\bar{t}} \tilde{G}(\mathbf{Y}^*_x(s))b_{j_1,j_2}^{(2,0)}(\mathbf{Y}^*_x(s))ds - \int_0^t \tilde{G}(\mathbf{Y}^*_x(s))b_{j_1,j_2}^{(2,0)}(\mathbf{Y}^*_x(s))ds.$$ 

First, note that

$$\|a_1(x_0, t, \bar{t})\|^2 = \sum_{j_1,j_2} \left( \int_0^t \tilde{G}(\mathbf{Y}^*_x(s))B''_{j_1}(\mathbf{Y}^*_x(s))B_{j_2}(\mathbf{Y}^*_2(s))ds \right)^2 \lesssim J^2 J^{-1} \int_0^{\bar{t}} (4J^2)ds = J^3|t - \bar{t}|,$$

where the second line follows by a similar argument used to bound $\|a(x_0, t)\|^2$. 

79
Next, $\|a_2(\tilde{t}, x_0, \tilde{x}_0)\|^2$ is given by

$$\sum_{j_1,j_2} \left[ \int_0^t \left( \tilde{G}(\Psi_x^+(s))B_{j_1}''(\Psi_{1,x_0}^+(s))B_{j_2}(\Psi_{2,x_0}^+(s)) - \tilde{G}(\Psi_x^+(s))B_{j_1}''(\Psi_{1,\tilde{x}_0}^+(s))B_{j_2}(\Psi_{2,\tilde{x}_0}^+(s))ds \right) \right]^2,$$

which is bounded (up to a constant multiple) by

$$\sum_{j_1,j_2} \left( \int_0^t \tilde{G}(\Psi_x^+(s))(B_{j_1}''(\Psi_{1,x_0}^+(s)) - B_{j_1}''(\Psi_{1,\tilde{x}_0}^+(s)))B_{j_2}(\Psi_{2,x_0}^+(s))ds \right)^2$$

$$+ \sum_{j_1,j_2} \left( \int_0^t \tilde{G}(\Psi_x^+(s))(B_{j_2}(\Psi_{2,x_0}^+(s)) - B_{j_2}(\Psi_{2,\tilde{x}_0}^+(s)))B_{j_1}''(\Psi_{1,x_0}^+(s))ds \right)^2$$

$$+ \sum_{j_1,j_2} \left( \int_0^t \left( \tilde{G}(\Psi_x^+(s)) - \tilde{G}(\Psi_x^+(s)) \right)B_{j_1}''(\Psi_{1,x_0}^+(s))B_{j_2}(\Psi_{2,x_0}^+(s))ds \right)^2.$$

Bound the first term in the right hand side of above expression as

$$\sum_{j_1,j_2} \left( \int_0^t \tilde{G}(\Psi_x^+(s))(B_{j_1}''(\Psi_{1,x_0}^+(s)) - B_{j_1}''(\Psi_{1,\tilde{x}_0}^+(s)))B_{j_2}(\Psi_{2,x_0}^+(s))ds \right)^2$$

$$\lesssim \sum_{j_1,j_2} \left( \int_0^t |\tilde{G}(\Psi_x^+(s))||B''_{j_1}(\Psi_{1,x_0}^+(s))||\Psi_{1,x_0}^+(s) - \Psi_{1,\tilde{x}_0}^+(s)||B_{j_2}(\Psi_{2,x_0}^+(s))ds \right)^2$$

$$\lesssim \|x_0 - \tilde{x}_0\|^2 \sum_{j_1,j_2} \left( \int_0^t |\tilde{G}(\Psi_x^+(s))||B''_{j_1}(\Psi_{1,x_0}^+(s))||B_{j_2}(\Psi_{2,x_0}^+(s))ds \right)^2$$

$$\lesssim \|x_0 - \tilde{x}_0\|^2 \|x_0 - \tilde{x}_0\|^2 J^5.$$

The second term is bounded as

$$\sum_{j_1,j_2} \left( \int_0^t \tilde{G}(\Psi_x^+(s))(B_{j_2}(\Psi_{2,x_0}^+(s)) - B_{j_2}(\Psi_{2,\tilde{x}_0}^+(s)))B_{j_1}''(\Psi_{1,\tilde{x}_0}^+(s))ds \right)^2$$

$$\lesssim \sum_{j_1,j_2} \left( \int_0^t |\tilde{G}(\Psi_x^+(s))||\Psi_{2,x_0}^+(s) - \Psi_{2,\tilde{x}_0}^+(s)||B'_{j_2}(\Psi_{2,x_0}^+(s))||B''_{j_1}(\Psi_{1,\tilde{x}_0}^+(s))ds \right)^2$$

$$\lesssim \|x - \tilde{x}_0\|^2 \sum_{j_1,j_2} \left( \int_0^t |\tilde{G}(\Psi_x^+(s))||B'_{j_2}(\Psi_{2,x_0}^+(s))||B''_{j_1}(\Psi_{1,x_0}^+(s))ds \right)^2.$$
\[ \lesssim \|x - \tilde{x}_0\|^2 \int_0^t 1\{\|s - s'| \leq CJ^{-1}\} \]
\[ \times \sum_{j_1,j_2} |B'_{j_2}(\Upsilon^*_x(x_0, s))| |B''_{j_1}(\Upsilon^*_x(x_0, s))| |B'_{j_2}(\Upsilon^*_x(x_0, s'))| |B''_{j_1}(\Upsilon^*_x(x_0, s'))| ds ds' \]
\[ \lesssim \|x - \tilde{x}_0\|^2 \int_0^t 1\{\|s - s'| \leq CJ^{-1}\} J^2 \sum_{j_1=1} J^2 |B''_{j_1}(\Upsilon^*_x(x_0, s))| |B'_{j_2}(\Upsilon^*_x(x_0, s'))| ds ds' \]
\[ \lesssim J^5 \|x_0 - \tilde{x}_0\|^2, \]

where the second line follows from the mean value theorem, the third line from the Lipschitz continuity of \( \Upsilon^*_x \) in \( x_0 \) (Remark 3.8.1) whereas fourth line follows by a similar argument used to bound \( \|a(x, t)\|^2 \). The third term
\[ \sum_{j_1,j_2} \left( \int_0^t (\tilde{G}(\Upsilon^*_x(s)) - \tilde{G}(\Upsilon^*_x(\tilde{x}_0(s))) B''_{j_1}(\Upsilon^*_x(x_0, s)) B_{j_2}(\Upsilon^*_x(x_0, s)) ds \right)^2 \]
\[ \lesssim \sum_{j_1,j_2} \int_0^t \|\Upsilon^*_x(s) - \Upsilon^*_x(\tilde{x}_0(s))\| |B''_{j_1}(\Upsilon^*_x(x_0, s))| |B_{j_2}(\Upsilon^*_x(x_0, s))| ds \]
\[ \lesssim \|\tilde{x}_0 - \tilde{x}_0\|^2 \sum_{j_1,j_2} \left( \int_0^t |B''_{j_1}(\Upsilon^*_x(x_0, s))| |B_{j_2}(\Upsilon^*_x(x_0, s))| ds \right)^2 \]
\[ \lesssim \|\tilde{x}_0 - \tilde{x}_0\|^2 J^3, \]

where the second line holds by mean value theorem and the third line by the Lipschitz continuity of \( \Upsilon^*_x \) in \( x_0 \) (Lemma 3.8.1) and last line holds by similar argument for \( \|a(x_0, t)\|^2 \).

In summary, we have that \( \|a_2(\tilde{t}, x_0, \tilde{x}_0)\|^2 \lesssim J^5 \|x_0 - \tilde{x}_0\|^2 \) and \( \|a_1(x_0, t, \tilde{t})\|^2 \lesssim J^3 |t - \tilde{t}|. \)

(ii). Now turning to the case \( r = (1, 1) \). By similar argument we have
\[ \|a(x_0, t)\|_1 = \sum_{j_1,j_2} \left| \int_0^t \tilde{G}_1(s) B'_{j_1}(\Upsilon^*_x(x_0, s)) B'_{j_2}(\Upsilon^*_x(x_0, s)) ds \right| \]
\[ \lesssim \sum_{j_1,j_2} \int_0^t |B'_{j_1}(\Upsilon^*_x(x_0, s))| |B'_{j_2}(\Upsilon^*_x(x_0, s))| ds \]
\[ \leq J \sum_{j_1=1} J_j \int_0^t |B'_{j_1}(\Upsilon^*_x(x_0, s))| ds \lesssim J^2 \]
Likewise, \( \|a(x_0,t)\|^2 = \sum_{j_1,j_2} \left( \int_0^t \tilde{G}_j(s)B_{j_1}(\Upsilon_{1,x_0}(s))B_{j_2}(\Upsilon_{2,x_0}(s))ds \right)^2 \) can be bounded by

\[
\int_0^t \int_0^t \sum_{j_1,j_2} 1\{|s - s'| < CJ^{-1}\} \left| \left( B'_{j_1}(\Upsilon_{1,x_0}(s))B'_{j_2}(\Upsilon_{2,x_0}(s))B'_{j_1}(\Upsilon_{1,x_0}(s'))B'_{j_2}(\Upsilon_{2,x_0}(s')) \right) \right| dsds' 
\]  
\[
\lesssim J \int_0^t \int_0^t \sum_{j_1,j_2} 1\{|s - s'| < CJ^{-1}\} \left| B'_{j_1}(\Upsilon_{1,x_0}(s))B'_{j_2}(\Upsilon_{2,x_0}(s))B'_{j_1}(\Upsilon_{1,x_0}(s')) \right| dsds' 
\]  
\[
\lesssim J^2 \int_0^t \int_0^t \sum_{j_1} 1\{|s - s'| < CJ^{-1}\} \left| B'_{j_1}(\Upsilon_{1,x_0}(s)) \right| dsds' 
\]  
\[
\lesssim J^3 J^{-1} \int_0^t \sum_{j_1} |B'_{j_1}(\Upsilon_{1,x_0}(s))| ds 
\]  
\[
\lesssim J^3. 
\]  

The third result \( \|a_1(x_0,t,\tilde{t})\|^2 \lesssim J^3 |t - \tilde{t}| \) and \( \|a_2(\tilde{t},x,\tilde{x}_0)\|^2 \lesssim J^5 \|x_0 - \tilde{x}_0\|^2 \) can be derived in a similar manner, since

\[
\|a_1(x_0,t,\tilde{t})\|^2 = \sum_{j_1,j_2} \left( \int_{\tilde{t}}^t \tilde{G}_j(\Upsilon_{x_0}(s))B'_{j_1}(\Upsilon_{1,x_0}(s))B'_{j_2}(\Upsilon_{2,x_0}(s))ds \right)^2 
\]  
\[
\lesssim J^2 J^{-1} \int_{\tilde{t}}^t (4J^2)ds 
\]  
\[
= J^3 |t - \tilde{t}|, 
\]  

where the second line follows by a similar argument used in bounding \( \|a(x_0,t)\|^2 \). Next, to bound \( \|a_2(\tilde{t},x_0,\tilde{x}_0)\|^2 \), we need to estimate the following three terms

\[
\sum_{j_1,j_2} \left( \int_0^\tilde{t} \tilde{G}_j(\Upsilon_{x_0}(s))(B'_{j_1}(\Upsilon_{1,x_0}(s)) - B'_{j_1}(\Upsilon_{1,x_0}(s'))B'_{j_2}(\Upsilon_{2,x_0}(s'))ds \right)^2, 
\]  
\[
\sum_{j_1,j_2} \left( \int_0^\tilde{t} \tilde{G}_j(\Upsilon_{x_0}(s))(B'_{j_2}(\Upsilon_{2,x_0}(s)) - B'_{j_2}(\Upsilon_{2,\tilde{x}_0}(s'))B'_{j_1}(\Upsilon_{1,\tilde{x}_0}(s'))ds \right)^2, 
\]  
\[
\sum_{j_1,j_2} \left( \int_0^\tilde{t} \tilde{G}(\Upsilon_{x_0}(s)) - \tilde{G}(\Upsilon_{\tilde{x}_0}(s))B'_{j_1}(\Upsilon_{1,\tilde{x}_0}(s))B'_{j_2}(\Upsilon_{2,\tilde{x}_0}(s))ds \right)^2. 
\]  

This can be done by similar argument used for the case \( r = (2,0) \) and we omit the details. ■
Proof of Proposition 3.5.2. We first sketch a proof for that for any \( \epsilon > 0, \)

\[
\Pi(\sup_{x_0 \in \mathcal{G}} |t^*_x - t_{x_0}| > \epsilon \| \mathbb{D}_n \) \overset{P_0}{\to} 0. 
\]

Recall that \( t^*_x = \arg\min_t \{ t \geq 0 : \langle \nabla f^*(\mathcal{Y}^*_0(t)), V^*(\mathcal{Y}^*_0(t)) \rangle = 0, \lambda^*(\mathcal{Y}^*_0(t)) < 0 \} \) Let \( C_{x_0} = \{ t \in [0, t^*_x - \epsilon], \langle \nabla f^*(\mathcal{Y}^*_0(t)), V^*(\mathcal{Y}^*_0(t)) \rangle = 0 \}. \) Notice that the existence of unique \( t^* \) is equivalent to the following set of assertions,

1. \( \inf_{x_0 \in \mathcal{G}, t \in C_{x_0}} \lambda^*(\mathcal{Y}^*_0(t)) > 0, \)

2. \( \inf_{x_0 \in \mathcal{G}, t \in [0, t^*_x - \epsilon) \setminus C_{x_0}} |\nabla f^*(\mathcal{Y}^*_0(t)), V^*(\mathcal{Y}^*_0(t))| > 0, \)

3. \( \sup_{x_0 \in \mathcal{G}, t \in [t^*_x - \epsilon, t^*_x + \epsilon]} \lambda^*(\mathcal{Y}^*_0(t)) < 0, \)

4. for all \( x_0 \in \mathcal{G}, \) there exists \( t_{x_0} \in [t^*_x - \epsilon, t^*_x + \epsilon] \) such that \( \nabla f^*(\mathcal{Y}^*_0(t_{x_0})), V^*(\mathcal{Y}^*_0(t_{x_0})) = 0. \)

By invoking the arguments similar to that in the proof of Proposition A.1 of Qiao and Polonik (2016), with \( P_0 \)-probability tending to 1, the above assertions hold with high posterior probability.

Now we are ready to establish the contraction rate. The proof of the bounds needed to establish this one is similar to that of Proposition 5.1 of Qiao and Polonik (2016). Let \( D_{\mathcal{Y}_x, V} f(t) := \nabla f(\mathcal{Y}_x(t))^T V(\mathcal{Y}_x(t)) \) and

\[
D^2_{\mathcal{Y}_x, V} f(t) := \langle \nabla \langle \nabla f(\mathcal{Y}_x(t)), V(\mathcal{Y}_x(t)) \rangle, V(\mathcal{Y}_x(t)) \rangle = \nabla f(\mathcal{Y}_x(t))^T \nabla V(\mathcal{Y}_x(t)) V(\mathcal{Y}_x(t)) + V(\mathcal{Y}_x(t))^T H f(\mathcal{Y}_x(t)) V(\mathcal{Y}_x(t)).
\]

We have for some \( \tilde{t} \) between \( t_{x_0} \) and \( t^*_x, \)

\[
0 = D_{\mathcal{Y}_x, V} f(t_{x_0}) - D_{\mathcal{Y}_x, V} f(t^*_x) + D^2_{\mathcal{Y}_x, V} f(\tilde{t})(\tilde{t} - t^*_x).
\]

We claim that

\[
\Pi(\inf_{x_0 \in \mathcal{G}} |D^2_{\mathcal{Y}_x, V} f(\tilde{t})| > \eta \| \mathbb{D}_n \) \overset{P_0}{\to} 0. 
\]

Since

\[
\sup_{x \in \mathcal{G}} |D^2_{\mathcal{Y}_x, V} f(\tilde{t}) - D^2_{\mathcal{Y}_x, V^*} f^*(t^*_x)| \geq \inf_{x \in \mathcal{G}} |D^2_{\mathcal{Y}_x, V} f(\tilde{t})| - \inf_{x \in \mathcal{G}} |D^2_{\mathcal{Y}_x, V^*} f^*(t^*_x)|
\]

and by Assumption (A3), \( \inf_{x \in \mathcal{G}} |D^2_{\mathcal{Y}_x, V^*} f^*(t^*_x)| \geq \eta, \) it suffices to show that with \( P_0- \)
probability tending to 1, \( \sup_{x \in \mathcal{G}} \left| D^2_{T\to x_0} f^*(\tilde{t}) - D^2_{T\to x_0} f^*(t^*_x) \right| \) is small with high posterior probability. To see this, we write

\[
\sup_x \left| D^2_{T\to x_0} f^*(\tilde{t}) - D^2_{T\to x_0} f^*(t^*_x) \right| \leq \sup_x \left| D^2_{T\to x_0} f^*(\tilde{t}) - D^2_{T\to x_0} f^*(\hat{t}) \right| + \sup_x \left| D^2_{T\to x_0} f^*(\hat{t}) - D^2_{T\to x_0} f^*(t^*_x) \right|.
\]

The first term contracts to zero, simply due to the uniform contraction results for \( \nabla f(x) \), \( V(x) \), \( \nabla V(x) \), \( Hf(x) \) (Lemma 3.8.1) and for \( \Upsilon_{x_0} \) (Theorem 3.5.1). The second term contracts to zero, since \( T_{x_0}^*(t) \) is continuous in \( t \) and \( \nabla f^*(x) \), \( V^*(x) \), \( \nabla V^*(x) \), \( Hf^*(x) \) are uniform continuous.

Also, \( \sup_{x_0} |D_{T\to x_0} f(t^*_x)| = \sup_{x_0} |D_{T\to x_0} f(t^*_x) - D_{T\to x} f(t^*_x)| \) is given by

\[
\sup_x \left| \nabla f(\Upsilon_{x_0}(t^*_x)) - \nabla f(\Upsilon^*(t^*_x)), V(\Upsilon_{x_0}(t^*_x)) - V^*(\Upsilon^*(t^*_x)) \right| + \sup_x \left| \nabla f^*(\Upsilon_{x_0}(t^*_x)) - \nabla f^*(\Upsilon^*(t^*_x)), V^*(\Upsilon_{x_0}(t^*_x)) - V^*(\Upsilon^*(t^*_x)) \right|.
\]

Now since

\[
\sup_{x_0} |t^*_x - t_{x_0}| \leq \frac{1}{\inf_{x_0} D^2_{T\to x_0} f(t)} \sup_{x_0} |D_{T\to x_0} f(t^*_x)|,
\]

we have

\[
\sup \left| t^*_x - t_{x_0} \right| \leq \sup_{x_0 \in \mathcal{G}} \left( \| \nabla f(\Upsilon_{x_0}(t^*_x)) - \nabla f(\Upsilon^*(t^*_x)) \| \sup_{x_0 \in \mathcal{G}} \| V^*(\Upsilon_{x_0}(t^*_x)) \| \right)
+ \sup_{x_0 \in \mathcal{G}} \left( \| V(\Upsilon_{x_0}(t^*_x)) - V^*(\Upsilon^*(t^*_x)) \| \sup_{x_0 \in \mathcal{G}} \| \nabla f^*(\Upsilon_{x_0}(t^*_x)) \| \right).
\]

It is easy to see that

\[
\sup_{x_0 \in \mathcal{G}} \| \nabla f(\Upsilon_{x_0}(t^*_x)) - \nabla f^*(\Upsilon^*(t^*_x)) \| \lesssim \sup_{x_0 \in \mathcal{G}} \| \nabla f(x) - \nabla f^*(x) \| + \sup_{x_0 \in \mathcal{G}} \sup_{t \in [0, T_{x_0}]} \| \Upsilon_{x_0}(t) - \Upsilon^*(t) \|,
\]

\[
\sup_{x_0 \in \mathcal{G}} \| V(\Upsilon_{x_0}(t^*_x)) - V^*(\Upsilon^*(t^*_x)) \| \lesssim \sup_{x_0 \in \mathcal{G}} \| V(x) - V^*(x) \| + \sup_{x_0 \in \mathcal{G}} \sup_{t \in [0, T_{x_0}]} \| \Upsilon_{x_0}(t) - \Upsilon^*(t) \|.
\]

With the choice of \( J \sim (n/\log n)^{1/(+2\alpha)} \), by Lemma 3.8.1, \( \sup_x \| \nabla f(x) - \nabla f^*(x) \| \) has posterior contraction rate of the order \( (n/\log n)^{(3-2\alpha)/(2(1+2\alpha))} \), while \( \sup_x \| V(x) - V^*(x) \| \) has that of the order \( (n/\log n)^{(5-2\alpha)/(2(1+2\alpha))} \). Therefore, considering the rate from Theorem 3.5.1, we have the desired result. ■
Proof of Theorem 3.5.3.} One can write
\[
\lambda_{x_0}(t_{x_0}) - \lambda_{x_0}(t_{x_0}^*) = \lambda_{x_0}(t_{x_0}) - \lambda_{x_0}(t_{x_0}) + \lambda_{x_0}(t_{x_0}) - \lambda_{x_0}(t_{x_0}^*)
\]
\[
= V(\lambda_{x_0}(t_{x_0}))(t_{x_0} - t_{x_0}^*) + \lambda_{x_0}(t_{x_0}^*) - \lambda_{x_0}(t_{x_0})
\]
\[
= V^*(\lambda_{x_0}(t_{x_0}))(t_{x_0} - t_{x_0}^*) + (V(\lambda_{x_0}(t_{x_0}^*)) - V^*(\lambda_{x_0}(t_{x_0}^*)))|t_{x_0} - t_{x_0}^*|
\]
\[
+ \lambda_{x_0}(t_{x_0}^*) - \lambda_{x_0}(t_{x_0})
\]
for $t_{x_0}$ between $t_{x_0}$ and $t_{x_0}^*$. Therefore,
\[
\sup_{x_0 \in G} \|\lambda_{x_0}(t_{x_0}) - \lambda_{x_0}(t_{x_0}^*)\| \leq \sup_{x_0 \in G} |t_{x_0} - t_{x_0}^*| \left(1 + \sup_{x_0 \in G} \|V(\lambda_{x_0}(t_{x_0}^*)) - V^*(\lambda_{x_0}(t_{x_0}^*))\| \right)
\]
\[
+ \sup_{x_0 \in G} \|\lambda_{x_0}(t_{x_0}^*) - \lambda_{x_0}(t_{x_0})\|.
\]
In view of the posterior contraction of $V$ (Lemma 3.8.1), Theorem 3.5.1 and Proposition 3.5.2, the posterior of $\sup_{x_0 \in G} \|V(\lambda_{x_0}(t_{x_0}^*)) - V^*(\lambda_{x_0}(t_{x_0}^*))\|$ contracts to zero.

Since
\[
\sup_{x_0 \in G} \|\lambda_{x_0}(t_{x_0}) - \lambda_{x_0}(t_{x_0}^*)\| \leq \sup_{x_0 \in G} |t_{x_0} - t_{x_0}^*| + \sup_{x_0 \in G} \|\lambda_{x_0}(t_{x_0}^*) - \lambda_{x_0}(t_{x_0})\|,
\]
by Theorem 3.5.1 and Proposition 3.5.2, we establish the rate for $\sup_{x_0 \in G} \|\lambda_{x_0}(t_{x_0}) - \lambda_{x_0}(t_{x_0}^*)\|$.

Proof of Proposition 3.5.6.} Consider (A2) for (i). Note that
\[
\lambda(\lambda_{x_0}(t_{x_0})) - \lambda^*(\lambda_{x_0}(t_{x_0}))
\]
\[
= V(\lambda_{x_0}(t_{x_0}))^THf(\lambda_{x_0}(t_{x_0}))V(\lambda_{x_0}(t_{x_0})) - V^*(\lambda_{x_0}(t_{x_0}))^THf^*(\lambda_{x_0}(t_{x_0}))V^*(\lambda_{x_0}(t_{x_0})).
\]
Thus by the posterior uniform contraction and uniform continuity of $V$, $Hf$ (Lemma 3.8.1) and the uniform contraction of $\lambda_{x_0}(t_{x_0})$ (Theorem 3.5.3) over $x_0$, with $P_0$-probability tending to 1, clearly condition (A2) holds with high posterior. Condition (A3) trivially holds by the same uniform contraction results and noting that
\[
\langle \nabla f(\lambda_{x_0}(t_{x_0})), V(\lambda_{x_0}(t_{x_0})), V(\lambda_{x_0}(t_{x_0}))\rangle
\]
\[
= \nabla f(\lambda_{x_0}(t_{x_0}))^T\nabla V(\lambda_{x_0}(t_{x_0}))V(\lambda_{x_0}(t_{x_0})) + V(\lambda_{x_0}(t_{x_0}))^THf(\lambda_{x_0}(t_{x_0}))V(\lambda_{x_0}(t_{x_0}))V(\lambda_{x_0}(t_{x_0})).
\]
The compactness condition follows trivially by the continuity of $\lambda(x)$ and $\langle \nabla f(x), V(x) \rangle$. Condition (A5) holds also due to Theorem 3.5.3. In view of Lemma 3.8.3 and Theorem 3.5.4, (ii)
Proof of Lemma 3.5.7. Let \( x_0 \) be an arbitrary point on \( \hat{\mathcal{L}} \). Thus \( \Upsilon_{x_0}(t_0) = x_0 \in \mathcal{L} \) and \( \Upsilon_{x_0}(t) \in \mathcal{L} \) for some \( t_0 > 0 \). Note that \( \inf_{y \in \mathcal{L}} ||x_0 - y|| \leq ||\Upsilon_{x_0}(0) - \Upsilon_{x_0}(t_0)|| \leq t_0 \) as \( ||V(x)|| = 1 \). Let \( D_{\Upsilon_{x_0}, V} f(t) := \frac{d}{dt} f(\Upsilon_{x_0}(t)) = \nabla f(\Upsilon_{x_0}(t))^T V(\Upsilon_{x_0}(t)) \) and

\[
D_{\Upsilon_{x_0}, V}^2 f(t) := \langle \nabla (\nabla f(\Upsilon_{x_0}(t)), V(\Upsilon_{x_0}(t))), V(\Upsilon_{x_0}(t)) \rangle
\]

\[= \nabla f(\Upsilon_{x_0}(t))^T \nabla V(\Upsilon_{x_0}(t)) V(\Upsilon_{x_0}(t)) + V(\Upsilon_{x_0}(t))^T H f(\Upsilon_{x_0}(t)) V(\Upsilon_{x_0}(t)),\]

where the second equality is due to the chain rule. A Taylor expansion yields that

\[D_{\Upsilon_{x_0}, V} f(t) - D_{\Upsilon_{x_0}, V} f(t_0) = (t - t_0) D_{\Upsilon_{x_0}, V}^2 f(\tilde{t})\]

for some \( \tilde{t} \) between \( 0 \) and \( t_0 \). In particular, since \( D_{\Upsilon_{x_0}, V} f(t_0) = 0 \), letting \( t = 0 \), we obtain

\[D_{\Upsilon_{x_0}, V} f(0) = -t_0 D_{\Upsilon_{x_0}, V}^2 f(\tilde{t}).\]

Furthermore,

\[|D_{\Upsilon_{x_0}, V} f(0)| = |D_{\Upsilon_{x_0}, V} f(0) - D_{\Upsilon_{x_0}, \hat{V}} f(0)|\]

\[= |\nabla f(\Upsilon_{x_0}(0))^T V(\Upsilon_{x_0}(0)) - \nabla \hat{f}(\Upsilon_{x_0}(0))^T \hat{V}(\Upsilon_{x_0}(0))|\]

\[\leq \sup_x |\nabla f(x)^T V(x) - \nabla \hat{f}(x)^T \hat{V}(x)|\]

\[\leq \sup_x |\nabla f(x)^T (V(x) - \hat{V}(x))| + \sup_x |(\nabla f(x) - \nabla \hat{f}(x))^T \hat{V}(x)|\]

\[\leq C \sup_x \|V(x) - \hat{V}(x)\| + \sup_x \|\nabla f(x) - \nabla \hat{f}(x)\|\]

\[\leq C \sup_x \|V(x) - \hat{V}(x)\|.
\]

By the uniform continuity of \( \nabla f(x) \), \( \nabla V(x) \), \( V(x) \) and \( H f(x) \) and the continuity of \( \Upsilon_{x_0}(t) \) in \( t \), without loss of generality, we can make \( t_0 \) small enough (see the comments after the proof). Thus we have

\[|D_{\Upsilon_{x_0}, V}^2 f(\tilde{t})| < \eta/2.
\]

By Assumption (A3), \(|D_{\Upsilon_{x_0}, V} f(t_0)| > \eta\), and hence

\[|D_{\Upsilon_{x_0}, V}^2 f(\tilde{t})| \geq |D_{\Upsilon_{x_0}, V} f(t_0)| - |D_{\Upsilon_{x_0}, V} f(t_0) - D_{\Upsilon_{x_0}, V} f(\tilde{t})| > \eta - \eta/2 = \eta/2.
\]
Therefore, \( \inf_{y \in L} \| x_0 - y \| \leq t_{x_0} \leq \frac{C}{\eta} \sup_x \| V(x) - \hat{V}(x) \| \). Thus
\[
d(\hat{L}|L) \leq \frac{C}{\eta} \sup_x \| V(x) - \hat{V}(x) \|.
\]
Similarly, \( d(L|\hat{L}) \leq \frac{C}{\eta} \sup_x \| V(x) - \hat{V}(x) \| \). Therefore,
\[
\text{Haus}(L, \hat{L}) \leq \frac{C}{\eta} \sup_{x \in [0,1]} \| V(x) - \hat{V}(x) \|.
\]
Recall that \( V(x) = G(df^2(x)) \). Now since \( G \) is a fixed Lipschitz continuous function, it is easy to get the upper bound for \( \sup_{x \in [0,1]} \| V(x) - \hat{V}(x) \| \) in terms of the supremum distance of the derivatives of \( f(x) - \hat{f}(x) \).

In above proof, \( t_{x_0} \) can be made arbitrarily small in the limit. To see this, if Assumption (A5) holds for the \( f \) (or more precisely, for \( \Upsilon_{x_0} \), then
\[
\| \Upsilon_{x_0}(t_{x_0} - \hat{\Upsilon}_{x_0}(\hat{t}_{x_0})) \| = \| \Upsilon_{x_0}(t_{x_0}) - x_0 \| = \| \Upsilon_{x_0}(t_{x_0}) - \Upsilon_{x_0}(0) \| > C_G t_{x_0}.
\]
Since \( \| \hat{\Upsilon}_{x_0}(\hat{t}_{x_0}) - \Upsilon_{x_0}(t_{x_0}) \| \) can be made arbitrarily small due to the closedness in supremum norm (see previous theorems), \( t_{x_0} \) can be made arbitrarily small.

**Proof of Theorem 3.5.9.** For \( \gamma < 1/2 \), by argument in the proof of Theorem 5.3 of Yoo and Ghosal (2016), one can establish that
\[
R_{n,k,\gamma} \asymp \mathbb{E} \left( \| f^{(k)} - \tilde{f}^{(k)} \|_{\infty} \right),
\]
\[
R_{n,k,\gamma}^2 \asymp (\log n/n) J^0 \asymp (\log n/n)^{(\alpha-2)/(\alpha+1)}.
\]
Recall from (3.1) that \( \Pi(f^{(k)}|D_n, \sigma^2) |_{\sigma^2=\hat{\sigma}^2} \sim \text{GP}(\tilde{f}^{(k)}, \hat{\sigma}^2 \Sigma^{(k)}) \). By Borell’s inequality (see Proposition A.2.1 of Van der Vaart and Wellner (1996)),
\[
\Pi(L \notin L|D_n, \hat{\sigma}^2) \leq \Pi(L \text{ corresponds to } f \notin C_{f,k,\gamma} \text{ for some } k \in \{1, 2, 3\}|D_n, \hat{\sigma}^2)
\]
\[
\leq \sum_{k=1}^{3} \Pi(\| f^{(k)} - \tilde{f}^{(k)} \|_{\infty} > \rho R_{n,k,\gamma}|D_n, \hat{\sigma}^2)
\]
\[
\leq 3 \max_{1 \leq k \leq 3} \exp[-c^2 R_{n,k,\gamma}^2/c_{n,k}],
\]
for some large constant \( c > 0 \) and \( c_{n,k} := \sup_x \text{var}(f^{(k)} - \tilde{f}^{(k)}|D_n, \hat{\sigma}^2) \) is bounded by a constant

87
\[
\sup_x \Sigma^{(k)}(x, x) \lesssim \sup_x \|L^{(k)}_{J_1, J_2}(x)\|^2 \|/(B^T B + \Lambda_0^{-1})^{-1}\|_{(2,2)} \lesssim (\log n)^{-3/(\alpha+1)} n^{(2-\alpha)/(\alpha+1)}.
\]

Therefore, the above posterior probability tends to zero. Finally,

\[
P_0(L^* \in C_L) = P_0(\|f^{(k)} - \tilde{f}^{(k)}\|_{\infty} \leq \rho R_{n,k,\gamma}, \forall k \in \{1, 2, 3\}) \to 1,
\]

by Lemma 3.8.3, establishing the coverage of \( C_L \).

To see \( C_L \subset \bar{C}_L \), for any \( L \in C_L \), it is induced by some \( f \) such that \( \|f^{(k)} - \tilde{f}^{(k)}\|_{\infty} \leq \rho \max_{1 \leq k \leq 3} R_{n,k,\gamma}, \forall k \in \{1, 2, 3\} \). In view of the discussion in the beginning of this section and Proposition 3.5.6, since Lemma 3.5.7 holds with \( P_0 \)-probability tending to 1 with \( L \) being the filament in posterior and \( \hat{L} \) being the filament induced by the posterior mean \( \hat{f} \), the result immediately follows. \( \blacksquare \)
Chapter 4

Direct Bayesian Learning of Level Curves

4.1 Introduction

In Chapter 2, we studied some Bayesian properties of estimation for level curves using the plug-in approach. In this chapter, we shall provide a Bayesian method where the level curve is directly modeled and the squared exponential periodic (SEP) Gaussian process prior is used for the curve. Let \( f : \mathbb{R}^2 \rightarrow \mathbb{R} \). The level set is given by \( \mathcal{L} := \{ x : f(x) = c \} \). We assume that this set is a simple closed smooth curve denoted by \( \gamma \).

4.2 Model, prior and posterior

Assume we have i.i.d. sample \((Y_i, X_i)_{i=1}^n\) from

\[ Y_i = f(X_i) + \varepsilon_i, \]

where \( \varepsilon_i \) is i.i.d Gaussian with mean 0 and variance \( \sigma_0^2 \) for \( i = 1, \ldots, n \). For convenience, we may write \( X \) as \((\omega, r)\) — the corresponding polar coordinates of \( X \). Therefore the data can be expressed as \( \{(\omega_i, r_i), Y_i\}_{i=1}^n \). As in Li and Ghosal (2017), we assume that the curve \( \gamma \) is a smooth function of \( \omega \) that maps from the unit sphere in \( \mathbb{R}^2 \) to \( \mathbb{R}^+ \). The region enclosed by this curve is given by \( \Gamma \).

One can use a rescaled squared exponential Gaussian process to induce a prior on \( \gamma \). Li and Ghosal (2017) showed that the squared exponential (SE) Gaussian process on the unit sphere in \( \mathbb{R}^2 \) is equivalent to the squared exponential periodic (SEP) Gaussian process on \([0, 1]\). They then obtained the analytical eigen-decomposition of the SEP kernel. Such decomposition allows
one to write down a series representation of the curve. Let \( \{ \psi_k(\omega), k = 1, 2, \ldots \} \) be the Fourier basis functions:

\[
\left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \cos(\omega), \frac{1}{\sqrt{\pi}} \sin(\omega), \frac{1}{\sqrt{\pi}} \cos(2\omega), \frac{1}{\sqrt{\pi}} \sin(2\omega), \ldots \right\}.
\]

The curve \( \gamma \) can be written as

\[
\gamma(\omega) = \eta(\omega) + \sum_{k=1}^{\infty} z_k \psi_k(\omega),
\]

where \( z_k \sim N(0, \nu_k(a)/\tau^2) \) independently for \( k = 1, 2, \ldots \); here \( \nu_1(a) = e^{-2a^2} I_0(2a^2) \), \( \nu_2(j(a) = \nu_{2j+1}(a) = e^{-2a^2} I_j(2a^2) \) and \( I_j \) is the modified Bessel function of the first kind of order \( j \).

Let \( \gamma = (\gamma(\omega_1), \ldots, \gamma(\omega_n))^T \). We can write using up to \( L \) terms in the series,

\[
\gamma = \Psi z + \eta, \quad z \sim N(0, \Sigma_a/\tau^2),
\]

where \( \Sigma_a = \text{diag}(\nu_1(a), \ldots, \nu_L(a)) \) and \( \Psi = [\psi_j(\omega_i)]_{i=1,j=1}^{n_L} \). For simplicity, we shall take \( \eta \) to be some fixed constant. For some choices of constants \( \epsilon_0, c_1, c_2 > 0 \), we shall also assume that the joint density of \( \mu \) is proportional to

\[
\exp \left\{ -\frac{c_2}{2} \sum_{(i,j):i \neq j, \|X_i - X_j\| < \epsilon_0} \frac{\|\mu_i - \mu_j\|^2}{\|X_i - X_j\|^2} - \frac{c_1}{2} \sum_i \|\mu_i\|^2 \right\}.
\]

Therefore, \( \mu = (\mu_1, \ldots, \mu_n)^T \sim N(0, (\Gamma + c_1 I)^{-1}) \) and the elements of \( \Gamma \) are given by \( \Gamma_{ii} = \sum_{k \neq i} 2c_2/\|X_i - X_k\|^2 \) and \( \Gamma_{ij} = -2c_2/\|X_i - X_j\|^2 \) for \( i \neq j \), with summations only taking place for those pairs with distance less than \( \epsilon_0 \). Let \( R(z) = \prod_{i=1}^N R_i(z) \), where \( R_i = [c, \infty) \) if \( X_i \in \Gamma \) and \( R_i = (-\infty, c] \) if \( X_i \notin \Gamma \). Let \( P(R|\mu, \Sigma) \) denote the probability that a multivariate normal random vector \( N(\mu, \Sigma) \) belongs to a region \( R \). We have

\[
f(\mu|z, X) = \frac{\exp\left( -\frac{1}{2} \mu^T (\Gamma + c_1 I) \mu \right) \prod_{i=1}^n 1_i(\mu_i)}{\int_{R_1(z)} \cdots \int_{R_n(z)} \exp\left( -\frac{1}{2} \mu^T (\Gamma + c_1 I) \mu \right) d\mu},
\]

where \( 1_i = 1_{R_i(z)} \). We can write the denominator of (4.2) as

\[
(2\pi)^{n/2} |\Gamma + c_1 I|^{-1/2} P(R(z)|0, (\Gamma + c_1 I)^{-1}).
\]
Suppose also \( Y | \mu, X, \sigma \sim N(\mu, \sigma^2 I) \), and we put priors

\[
a \sim \text{gamma}(\alpha_a, \beta_a),
\]

\[
\tau^2 \sim \text{gamma}(\alpha_{\tau^2}, \beta_{\tau^2}),
\]

\[
\sigma^{-2} \sim \text{gamma}(\alpha_{\sigma^2}, \beta_{\sigma^2}),
\]

for some positive constants \( \alpha_a, \alpha_{\tau^2}, \beta_{\tau^2}, \beta_{\sigma^2} \) and \( \beta_{\sigma^2} \).

The parameters are given by \((z, \tau^2, a, \mu, \sigma^2)\). The posterior sampling is then given in the following:

\[
\mu, z | \tau^2, a, \sigma^2, Y \propto f(Y | \mu, \sigma^2) f(\mu | z) f(z | \tau^2, a),
\]

\[
\tau^2 | z, a \propto f(\tau^2) f(z | \tau^2, a),
\]

\[
a | z, \tau^2 \propto f(a) f(z | \tau^2, a),
\]

\[
\sigma^{-2} | \mu, Y \propto f(Y | \mu, \sigma^2) f(\sigma^{-2}).
\]

More specifically, the posteriors are given by the following:

\[
(\mu, z) | \tau^2, a, \sigma^2, Y \propto \exp \left\{ -\frac{1}{2\sigma^2} (Y - \mu)^T (Y - \mu) - \frac{1}{2} \mu^T (\Gamma + c_1 I) \mu \\
- \frac{1}{2} \tau^2 z^T \Sigma^{-1} z \right\} \prod_{i=1}^n \frac{1}{P(R(z)|0, (\Gamma + c_1 I)^{-1})},
\]

(4.3)

\[
\tau^2 | z, a \sim \text{gamma} \left( \alpha_{\tau^2} + \frac{L}{2}, \beta_{\tau^2} + \frac{z^T \Sigma^{-1} z}{2} \right),
\]

(4.4)

\[
a | z, \tau^2 \propto \exp \left\{ -\log \frac{\left| \Sigma_a \right|}{2} - \frac{\tau^2 z^T \Sigma^{-1} z}{2} + (\alpha_a - 1) \log a - \beta_a a \right\},
\]

(4.5)

\[
\sigma^{-2} | \mu, Y \sim \text{gamma} \left( \alpha_{\sigma^2} + \frac{n}{2}, \beta_{\sigma^2} + \frac{(Y - \mu)^T (Y - \mu)}{2} \right).
\]

(4.6)

We can use Metropolis-Hastings algorithm to update \( \mu, z \). Using some transition probability function \( J \) (e.g., multivariate normal), for a pair of proposed values \((\mu^*, z^*)\) and current values \((\mu^{(s)}, z^{(s)})\) at the \( s \)-th stage, we can compute

\[
r = \exp \{ \log f(\mu^*, z^*) - \log f(\mu^s, z^s) + \log J(\mu^s, z^s | \mu^*, z^*) - \log J(\mu^*, z^s | \mu^s, z^s) \},
\]

and set \((\mu^{(s+1)}, z^{(s+1)})\) to \((\mu^*, z^*)\) with probability of \( \min(1, r) \) and to \((\mu^{(s)}, z^{(s)})\) otherwise.
4.3 Discussion

The posterior sampling in this problem has a unique challenge. First, note that the joint sampling of \((\mu, z)\) using Metropolis-Hastings algorithm described above is motivated by the following observations. Given \(z, \mu\) follows a multivariate truncated normal distribution. A direct sampling from \(\mu\) given \(z\) is straightforward. However, it does not flip the directions of the elements of \(\mu\). Similarly, given \(\mu\), updating \(z\) alone will not flip the directions of the elements of \(\mu\), as they need to be compliant with the directions already given. It is expected that the joint sampling of \((\mu, z)\) should be a right way to do. However, it is observed that the flips rarely happen.

As an alternative approach, one may want to obtain the posterior distribution of \(z\) with \(\mu\) integrated out. To do this, first note that

\[
f(Y|\sigma^2, z) = \int f(Y|z, \mu, \sigma^2) f(\mu|z) d\mu
\]

\[
= \int_{R(z)} N(Y|\mu, \sigma^2 I) \cdot N(\mu|0, (\Gamma + c_1 I)^{-1}) d\mu
\]

\[
= \frac{c(Y)}{\text{P}(R(z)|0, (\Gamma + c_1 I)^{-1})} \int_{R(z)} N(\mu|\mu_0, \Sigma_0) d\mu,
\]

By conjugacy, the last equation further simplifies to

\[
\frac{c(Y)}{\text{P}(R(z)|0, (\Gamma + c_1 I)^{-1})} \int_{R(z)} N(\mu|\mu_0, \Sigma_0) d\mu,
\]

where

\[
c(Y) = \int N(Y|\mu, \sigma^2 I) \cdot N(\mu|0, (\Gamma + c_1 I)^{-1}) d\mu,
\]

\[
\mu_0 = ((c_1 + \sigma^{-2})I + \Gamma)^{-1} \sigma^{-2}Y,
\]

\[
\Sigma_0 = ((c_1 + \sigma^{-2})I + \Gamma)^{-1}.
\]

Then we can write

\[
f(Y|\sigma^2, z) = \frac{c(Y) \text{P}(R(z)|\mu_0, \Sigma_0)}{\text{P}(R(z)|0, (\Gamma + c_1 I)^{-1})}.
\]

Therefore,

\[
f(z|\tau^2, a, \sigma^2, Y) = \left( \int f(Y|z, \mu, \sigma^2) f(\mu|z) d\mu \right) f(z|\tau^2, a) = \frac{c(Y) \text{P}(R(z)|\mu_0, \Sigma_0) f(z|\tau^2, a)}{\text{P}(R(z)|0, (\Gamma + c_1 I)^{-1})}.
\]
In Metropolis-Hastings algorithm, given two values \( z_1 \) and \( z_2 \), we evaluate the ratio

\[
\frac{f(z_1 | \tau^2, a, \sigma^2, Y)}{f(z_2 | \tau^2, a, \sigma^2, Y)} = \frac{P(R(z_1)|\mu_0, \Sigma_0)}{P(R(z_2)|\mu_0, \Sigma_0)} \cdot \frac{P(R(z_2)|0, (\Gamma + c_0 I)^{-1})}{P(R(z_1)|0, (\Gamma + c_0 I)^{-1})} \cdot \frac{f(z_1 | \tau^2, a)}{f(z_2 | \tau^2, a)}. \tag{4.8}
\]

To update \( \mu \), note that

\[
\mu | \tau^2, a, \sigma^2, z, Y \propto \exp \left\{ -\frac{1}{2\sigma^2}(Y - \mu)^T(Y - \mu) - \frac{1}{2} \mu^T(\Gamma + c_1 I)\mu \right\} \prod_{i=1}^{n} 1_i(\mu_i). \tag{4.9}
\]

Therefore, the posterior distribution of \( \mu \) is a truncated normal with mean \( \mu_0 \) and covariance matrix \( \Sigma_0 \) on the region \( R(z) \).

### 4.4 Simulation

We investigate the second approach mentioned in Section 4.3 by simulation. We generate \( Y_i = f(X_i) + \varepsilon_i \), where \( \varepsilon_i \) is i.i.d. normally distributed with mean 0 and precision coefficient 10; and the function \( f \) is given by

\[
f(x_1, x_2) = g_1(x_1, x_2) + g_2(x_1, x_2),
\]

where \( g_1 \) and \( g_2 \) are the density functions of the following two variables respectively,

\[
N_1 \sim \text{Normal} \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.1 & 0.01 \\ 0.01 & 0.1 \end{pmatrix} \right],
\]

\[
N_2 \sim \text{Normal} \left[ \begin{pmatrix} 0.3 \\ 0.3 \end{pmatrix}, \begin{pmatrix} 0.05 & 0.01 \\ 0.01 & 0.1 \end{pmatrix} \right].
\]

The sample size is 300. The Figure 4.1 shows the true level curve for \( c = 0.8 \) and the initial curve with the directions of the elements of \( y \) and \( \mu \) respectively. We set \( c_1 = 0.1 \), \( c_2 = 0.06 \). We estimate \( \sigma^2 \) by some preliminary estimator—say the estimator given by the expression (2.9). We investigate three configurations by setting \( \epsilon_0 \) to 0.5, 1 or 5. When \( \epsilon_0 = 5 \), all pairs of distances will contribute to the summation in (4.1). All these simulation results are put in Figures 4.2, 4.3 and 4.4 at the end of this chapter. These figures show a clear shrinkage of the curves toward the correct center. At certain stages of the iterations, some curves can come close to the true
curve to some degree. It is worth to point out that in this simulation setting, the sample size 300 is chosen deliberately not too large so that efficient computation of orthant probabilities are possible. It is unclear how much effects sample size can have on the results, as Figure 4.1 shows that observations are sparse in some regions. A recent paper Azzimonti and Ginsbourger (2017) proposed a method to compute orthant probability for a multivariate normal random vector of up to 7000 dimensions. We are positive that this method will be useful for our purpose and further study will be conducted in the future.

![Figure 4.1: The left figure gives the true level curve and the red and blue dots indicate the directions of the elements of \( y \). The right figure gives the initial curve and the red and blue dots indicate the initial directions of the elements of \( \mu \).](image)

### 4.5 Future research

The simulation results for the approach discussed in last section suggests that the method is promising. In this section, we propose a third interesting approach where the flipping is enforced with some artificial probability. This allows \( \mu \) to flip regardless of the curve \( z \). In this approach, the boundary restriction induced by \( z \) may be considered indefinite albeit the curve may be given. The idea is the following. Even if \( X_i \) is outside the boundary, \( \mu_i \) may still have a small probability to flip above the level. Similarly, even if \( X_i \) is inside the boundary, \( \mu_i \) still has a small probability to flip below the level. To capture both scenarios, we introduce Bernoulli random variables \( \delta_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(p), i = 1, \ldots, n \) and \( \pi_0 \overset{\sim}{\sim} \text{Beta}(1, \beta) \). We modify our notations \( R_i \) and \( 1_i \)
via defining \( R(z, \delta) = \prod_{i=1}^{n} R_i(z, \delta_i) \), where

\[
R_i = \begin{cases} 
\,[c, \infty), & \text{if } X_i \in \Gamma \text{ and } \delta_i = 0, \\
\,(-\infty, c], & \text{if } X_i \in \Gamma \text{ and } \delta_i = 1, \\
\,(-\infty, c], & \text{if } X_i \not\in \Gamma \text{ and } \delta_i = 0, \\
\,[c, \infty), & \text{if } X_i \not\in \Gamma \text{ and } \delta_i = 1,
\end{cases}
\]

and letting \( 1_i = 1_{R_i(z, \delta_i)} \). To simplify notation, let \( c(\mu, z, \tau^2, \sigma^2, a) \) denote

\[
-\frac{1}{2\sigma^2} (Y - \mu)^T (Y - \mu) - \frac{1}{2} \mu^T (\Gamma + c_1 I) \mu - \frac{1}{2} \tau^2 z^T \Sigma_\alpha^{-1} z.
\]

The joint distribution of \((\mu, z, \delta)\) given all other parameters and data will be proportional to

\[
\exp \left\{ c(\mu, z, \tau^2, \sigma^2, a) \right\} \prod_{i=1}^{n} 1_i(\mu_i) \left( \frac{\pi_0^{\sum_i \delta_i} (1 - \pi_0)^{n - \sum_i \delta_i}}{\pi_0^{\sum_i \delta_i} (1 - \pi_0)^{n - \sum_i \delta_i}} \right) \frac{P(R(z, \delta) | 0, (\Gamma + c_1 I)^{-1})}{P(R(z, \delta) | 0, (\Gamma + c_1 I)^{-1})}.
\]

Integrate out \( \delta \), we further have the distribution of \((\mu, z)\) proportional to

\[
\sum_{\delta_1=0}^{1} \cdots \sum_{\delta_n=0}^{1} \exp \left\{ c(\mu, z, \tau^2, \sigma^2, a) \right\} \prod_{i=1}^{n} 1_i(\mu_i) \left( \frac{\pi_0^{\sum_i \delta_i} (1 - \pi_0)^{n - \sum_i \delta_i}}{\pi_0^{\sum_i \delta_i} (1 - \pi_0)^{n - \sum_i \delta_i}} \right) \frac{P(R(z, \delta) | 0, (\Gamma + c_1 I)^{-1})}{P(R(z, \delta) | 0, (\Gamma + c_1 I)^{-1})}.
\]

This expression has a unique advantage. For any given \((\mu, z)\), there is one and only one sequence of \((\delta_1, \ldots, \delta_n)\) that makes the distribution non-degenerate. Instead of evaluating all the summations, for any given \((\mu, z)\), we only need to find the sequence \((\delta_1, \ldots, \delta_n)\) that makes the distribution non-degenerate and then compute one summand only. This sequence makes \( \mu_i \in R_i \) for all \( i = 1, \ldots, n \) and is easy to identify. It is convenient when we use a normal transition probability for \((\mu, z)\) in the Metropolis-Hastings step. By introducing an artificial probability, it is expected that this can facilitate the flipping of \( \mu \) as there will be no loss of samples due to the restriction of the truncation.

Instead of updating \( \mu \) and \( z \) together. Here is an alternative approach to enforce artificial flipping of \( \mu \). Let \( c(\mu, \sigma^2) \) denote

\[
-\frac{1}{2\sigma^2} (Y - \mu)^T (Y - \mu) - \frac{1}{2} \mu^T (\Gamma + c_1 I) \mu.
\]

Note that the joint posterior distribution of \((\mu_i, \delta_i)\) given \( z, \mu_{\cdot i}, \delta_{\cdot i} \), where the subscript \(-i\)
denotes all elements except the $i$-th one, is proportional to
\[
\exp\left( c(\mu, \sigma^2) \right) \mathbb{1}_i(\mu_i) p_{0}^{1-\delta_i}(1-\pi_0)^{1-\delta_i} P(R(z, \delta)|0, (\Gamma + c_1 I)^{-1}).
\]\(4.10\)

Therefore, we can sample the posterior distribution $\mu_i$ given $z, \mu_{-i}, \delta$ from
\[
\exp\left( c(\mu, \sigma^2) \right) \mathbb{1}_i(\mu_i).
\]\(4.11\)

We recognize that this is a univariate truncated normal distribution for $\mu_i$ (up to some factor) with mean $\mu_0$ and covariance $\Sigma_0$ given all other variables. The presence of $\delta_i$ dictates the direction to sample $\mu_i$.

To update $\delta_i$ given $z, \mu_{-i}$, and $\delta_{-i}$, we integrate out $\mu_i$ in (4.10), which gives
\[
\frac{P(\mu_i \in R_i(z, \delta_i)|\mu_0, \Sigma_0) \pi_0^{1-\delta_i}}{P(\mu_i \in R_i(z, \delta_i), \mu_{-i} \in R_{-i}|0, (\Gamma + c_1 I)^{-1})}.
\]\(4.12\)

Define
\[
P_1(z, \mu_{-i}, \delta_{-i}) = \frac{P(\mu_i \in R_i(z, 1)|\mu_0, \Sigma_0) \pi_0}{P(\mu_i \in R_i(z, 1), \mu_{-i} \in R_{-i}|0, (\Gamma + c_1 I)^{-1})},
\]
\[
P_2(z, \mu_{-i}, \delta_{-i}) = \frac{P(\mu_i \in R_i(z, 0)|\mu_0, \Sigma_0)(1-\pi_0)}{P(\mu_i \in R_i(z, 0), \mu_{-i} \in R_{-i}|0, (\Gamma + c_1 I)^{-1})}.
\]

Then the posterior density of $\delta_i$ is given by
\[
P(\delta_i = 1|z, \mu_{-i}, \delta_{-i}) = \frac{\hat{P}_1(z, \mu_{-i}, \delta_{-i})}{\hat{P}_1(z, \mu_{-i}, \delta_{-i}) + \hat{P}_2(z, \mu_{-i}, \delta_{-i})}.
\]

At last, we update $\pi_0 \sim \text{Beta}(1 + \sum_{i=1}^n \delta_i, \beta + n - \sum_{i=1}^n \delta_i)$.

This new approach, together with the approach described in Section 4.3, will be subject to intensive study by the author in the future.
Figure 4.2: Direct Bayes estimate of level curves ($\epsilon_0 = 0.5$). Going from the left to the right and from the top to the bottom, these figures show the status in the 100-th, 200-th, 300-th, 400-th, 1000-th iteration and the posterior mean with one fifth burn-in.
Figure 4.3: Direct Bayes estimate of level curves ($\epsilon_0 = 1$). Going from the left to the right and from the top to the bottom, these figures show the status in the 100-th, 200-th, 300-th, 400-th, 1000-th iteration and the posterior mean with one fifth burn-in.
Figure 4.4: Direct Bayes estimate of level curves ($\epsilon_0 = 5$). Going from the left to the right and from the top to the bottom, these figures show the status in the 100-th, 200-th, 300-th, 400-th, 1000-th iteration and the posterior mean with one fifth burn-in.
REFERENCES


Appendix A

Some auxiliary lemmas

**Lemma A.0.1** For a sequence of i.i.d random variables $Z_j \sim N(0, \sigma^2)$, for some universal constant $C_1$ and $C_2$,

$$C_1 \sigma \sqrt{\log n} \leq E \max_{1 \leq j \leq n} Z_j \leq E \max_{1 \leq j \leq n} |Z_j| \leq C_2 \sigma \sqrt{\log n}$$

**Proof.** For proof, see Van der Vaart and Wellner (1996).

**Lemma A.0.2** The following bounds for integral of product of B-splines hold for any $j_k \in \{1, \ldots, J_k\}$ and $j'_k \in \{1, \ldots, J_k\}$ and $k = 1, \ldots, d$:

$$\int_0^1 B_{j_k,q_k}(x)dx_k \gtrsim J_k^{-1}$$

$$\int_0^1 B_{j_k,q_k}(x)B_{j'_k,q_k}(x)dx_k \lesssim J_k^{-1}.$$

**Proof.** Consider the second expression first. In view of the result in p.128 of De Boor (1978),

$$\int_0^1 B_{j_k,q_k}(s)ds = \sum_{i \geq j_k+1} B_{i,q_k+1}(1)$$

thus

$$\int_0^1 B_{j_k,q_k}(s)ds \lesssim J_k^{-1} \text{ as } \sum_{i \geq j_k+1} B_{i,q_k+1}(1) \leq 1.\text{ Since } 0 \leq B_{j'_k,q_k} \leq 1,$$

$$\int_0^1 B_{j_k,q_k}(x_k)B_{j'_k,q_k}(x_k)dx_k \leq \int_0^1 B_{j_k,q_k}(s)ds \lesssim J_k^{-1}.$$

Now turning to first expression. Suppose $x_k \in (t_{j_k-1,j_k})$, by the derivative formula of
B-splines, it can be shown that

\[ B_{jk}(x_k) = (t_{k,j} - x_k)^{q_k - 1} \prod_{i_k=1}^{q_k-1} (t_{k,j} - t_{k,j-1} - q_k + i_k)^{-1} \]

Since \( t_{k,j} - t_{k,j-1} - q_k + i_k \lesssim \Delta_k(q_k - i_k) \), the right hand side of the above expression is lower bounded by \( (t_{k,j} - x_k)^{q_k - 1} \prod_{i_k=1}^{q_k-1} (\Delta_k^{-1}(q_k - i_k)^{-1}) \). Therefore, \( B_{jk}(x_k) \gtrsim \Delta_k - \Delta_k^{-1}(q_k - i_k) \). Now

\[
\int_0^1 B_{jk}(x_k) dx_k \geq \int_{t_{k,j-1}}^{t_{k,j}} B_{jk}(x_k) dx_k \\
= \frac{\Delta_k^{2(q_k-1)}}{[(q_k - 1)!]^2} \int_{t_{k,j-1}}^{t_{k,j}} (t_{k,j} - x_k)^{2(q_k-1)} dx_k \\
\geq \frac{\Delta_k^{2(q_k-1)}}{[(q_k - 1)!]^2} \frac{(t_{k,j} - t_{k,j-1})^{2q_k-1}}{2q_k - 1}.
\]

Due to the assumption that \( \min_{0 \leq \ell \leq N_k} \delta_{k,\ell} \gtrsim \Delta_k/C \), the last expression is lower bounded by \( \Delta_k \), which is of the order \( J_{k}^{-1} \).

Lemma A.0.3 For \( f \in H^\alpha([0, 1]^d) \) with \( \alpha \leq \min(q_1, \ldots, q_d) \). Then for \( J_k \asymp J_{n}^* \) for \( k = 1, \ldots, d \) where \( J_{n}^* \) is sufficiently large, there exists a function \( f_\infty := b_{J_{1}, \ldots, J_{d}}^T \theta_\infty \) for some \( \theta_\infty \) such that

\[
\|f_\infty - f\|_\infty \leq C \sum_{k=1}^{d} J_{k}^{\alpha - \alpha}, \\
\|D_r f_\infty - D_r f\|_\infty \leq C \sum_{k=1}^{d} J_{k}^{\alpha - \alpha},
\]

for some positive constant \( C \) depending on \( \alpha \) and \( q \) and for every integer vector \( r \) satisfying \( |r| < \alpha \). In addition, the following holds:

1. If \( f > 0 \), each element of \( \theta_\infty \) can be restricted to positive;
2. If \( 0 < f < 1 \), then each element of \( \theta_\infty \) can be restricted to \((0, 1)\);
3. If \( f \) is a density function such that \( \sqrt{T} \) belongs to \( H^\alpha([0, 1]^d) \), then there exists some \( \theta_\infty \) such that \( (\theta_\infty)^T \Gamma^{-1/2} \Xi \Gamma^{-1/2} \theta_\infty = 1 \) and the same rates described above hold for \( \|\sqrt{T} - b^T \theta_\infty\|_\infty \), where \( b \) is defined in (2.10).

Proof. See Schumaker (2007) or Lemma E.7 of Ghosal and van der Vaart (2017) for the proof of the main result and the claims (1)–(2). We shall prove claim (3) here for the univariate
case. Multivariate case follows similarly. Suppose \( \sqrt{\mathcal{J}} \in \mathcal{H}^\alpha \). By Claim (1), there exists some \( \eta_1 \) such that \( \| \sqrt{\mathcal{J}} - b^T \eta_1 \|_\infty \lesssim J^{-\alpha} \), for some \( \eta_1 \in (0, \infty)^d \) and \( \| b^T \eta_1 \|_\infty < \infty \). We define \( \eta_{2,j} = C_j \eta_{1,j} \) for \( 1 \leq j \leq J \), and \( C_j \) is the normalization factor in the definition of \( \tilde{B}_{j,q} \). Thus \( \| \sqrt{\mathcal{J}} - \tilde{b}^T \eta_2 \|_\infty \lesssim J^{-\alpha} \) and \( \| \tilde{b}^T \eta_2 \|_\infty = \| b^T \eta_1 \|_\infty \). Let \( \theta = \eta_2/(\eta_2^T \Gamma^{-1/2} \Xi \Gamma^{-1/2} \eta_2)^{1/2} \). Notice that \( \theta^T \Gamma^{-1/2} \Xi \Gamma^{-1/2} \theta = 1 \). We have

\[
\| \sqrt{\mathcal{J}} - \tilde{b}^T \theta \|_\infty = \| \sqrt{\mathcal{J}} - \tilde{b}^T \eta_2 + \tilde{b}^T \eta_2 - \tilde{b}^T \theta \|_\infty \\
\leq \| \sqrt{\mathcal{J}} - \tilde{b}^T \eta_2 \|_\infty + \left| \tilde{b}^T \eta_2 \left( 1 - \frac{1}{(\eta_2^T \Gamma^{-1/2} \Xi \Gamma^{-1/2} \eta_2)^{1/2}} \right) \right|_\infty \\
\lesssim J^{-\alpha} + \| \tilde{b}^T \eta_2 \|_\infty \left| 1 - \frac{1}{(\eta_2^T \Gamma^{-1/2} \Xi \Gamma^{-1/2} \eta_2)^{1/2}} \right|. 
\]

Now also \( \int (f - (\tilde{b}^T \eta_2)^2) = \int (\sqrt{\mathcal{J}} + \tilde{b}^T \eta_2)(\sqrt{\mathcal{J}} - \tilde{b}^T \eta_2) \leq \| \sqrt{\mathcal{J}} - \tilde{b}^T \eta_2 \|_\infty \int (\sqrt{\mathcal{J}} + \tilde{b}^T \eta_2) \). Since \( \int \tilde{b}^T \eta_2 = \int b^T \eta_1 \leq \| b^T \eta_1 \|_\infty \leq \| \eta_1 \|_\infty < \infty \) by fact that \( \sum_{j=1}^J B_{j,q} = 1 \). Also, \( \int \sqrt{\mathcal{J}} \leq (\int f)^{1/2} = 1 \). We have \( \int (f - (\tilde{b}^T \eta_2)^2) \lesssim J^{-\alpha} \). Thus \( | 1 - \eta_2^T \Gamma^{-1/2} \Xi \Gamma^{-1/2} \eta_2 | \lesssim J^{-\alpha} \), which implies that \( \eta_2^T \Gamma^{-1/2} \Xi \Gamma^{-1/2} \eta_2 > 1/2 \) for \( J \) large. Therefore

\[
\left| 1 - \frac{1}{(\eta_2^T \Gamma^{-1/2} \Xi \Gamma^{-1/2} \eta_2)^{1/2}} \right| \leq \sqrt{2} | \eta_2^T \Gamma^{-1/2} \Xi \Gamma^{-1/2} \eta_2 - 1 | \lesssim J^{-\alpha}, 
\]

using the fact that \( | 1 - \sqrt{x} | \leq | (1 - \sqrt{x})(1 + \sqrt{x}) | = | 1 - x | \) for \( x \geq 0 \). This completes the proof \( \| \sqrt{\mathcal{J}} - \tilde{b}^T \theta \|_\infty \lesssim J^{-\alpha} \). ■

Lemma A.0.4 For regression applications with observations \( (Y_i, X_i)_{i=1}^n \), the following convention is assumed. Let \( Y = (Y_1, \ldots, Y_n)^T \) and \( X = (X_1^T, \ldots, X_n^T)^T \) denote the response variables and the covariates respectively. For the fixed design cases, \( X \) is considered fixed and \( Y_i \) is independently distributed according to some distribution indexed by \( X_i \). For the random design case, \( X_i \) is further assumed to be i.i.d. For both cases, we also assume that there exists some cumulative distribution function \( \mathbb{G} \) with positive and continuous density

\[
\sup_x | \mathbb{G}_n(x) - \mathbb{G}(x) | = o \left( \prod_{k=1}^d J_k^{-1} \right), \tag{A.1} 
\]

where \( \mathbb{G}_n \) is the empirical distribution of \( \{ X_i : i = 1, \ldots, n \} \). If the Condition (A.1 holds), then we have

\[
C_1 n \left( \prod_{k=1}^d J_k^{-1} \right) I \leq B^T B \leq C_2 n \left( \prod_{k=1}^d J_k^{-1} \right) I, 
\]
for some constants $C_1, C_2 > 0$. If in addition, for some constants $0 < c_1 \leq c_2 < \infty$ such that

$$c_1 I \leq \Lambda \leq c_2 I,$$

we have

$$\left( C_1 n \prod_{k=1}^{d} J_k^{-1} + c_2^{-1} \right) I \leq B^T B + \Lambda^{-1} \leq \left( C_2 n \prod_{k=1}^{d} J_k^{-1} + c_1^{-1} \right) I.$$

Appendix B

Posterior contraction rates using finite random series priors

Shen and Ghosal (2015) first provided a master theorem for posterior contraction rates using finite random series priors with adaptation. See also Section 10.4 of Ghosal and van der Vaart (2017). Here we adapt Lemma 10.20 and Theorem 10.21 of Ghosal and van der Vaart (2017) for our purpose.

Recall that the observations \( X = (X_i)_{i=1}^n \) have joint distribution given by \( P_f \) and density function given by \( p_f \). Here \( f \) belongs to some parameter space \( F \) which is endowed with some semimetric \( d_n \). This metric \( d_n \) will be the one with respect to which the contraction rates are described in the master theorems. As in the basic testing condition (1.6), we may use another semimetric \( e_n \). For convenience, the basic testing condition (1.6) is reproduced here:

\[
P_f^{(n)} \phi_n \leq \exp(-C_1 n \epsilon^2), \quad \sup_{f \in F_n} P_f^{(n)}(1 - \phi_n) \leq \exp(-C_2 n \epsilon^2).
\]

We model \( f \) as \( f = \sum_{j=1}^J \theta_j \psi_{J,j} \) for some basis functions \( \psi_J = \{\psi_{J,1}, \ldots, \psi_{J,J} \} \) and impose some prior on the coefficients \( \theta \). Let \( S(M_n) \subset [-M_n, M_n] \) be certain sets to be specified in the context and \( F_n := \{\theta^T \psi_{J,n} : \theta \in \mathbb{R}^J, \theta \in S(M_n)^J \} \). First we state the following two conditions (P1) and (P2) for the prior:

(P1) \( \Pi(||\theta_{J,n} - \theta_0|| \leq \epsilon) \geq \exp(-c_1 J_n \log(\sqrt{J_n}/\epsilon)) \) for every \( \theta_0 \in \mathbb{R}^J \) with \( ||\theta_0||_\infty \leq H \) for some positive constants \( c_1 \) and \( H \), and every sufficiently small \( \epsilon > 0 \).

(P2) There exists some \( M_0 > 0 \) such that \( \Pi(\theta_{J,n} \notin S(M_n)^J) \leq J_n \exp(-c_2 M_n^\tau) \) for some positive constant \( c_2 \) and all \( M_n \geq M_0 > 0 \) for some large constant \( \tau > 0 \).

We let \( d \) to be some natural metric on \( F \), for instance \( L^\infty \)-metric. We would also need the following condition for the series representation.
For every $\theta_1, \theta_2 \in \mathbb{R}^J$, $d(\theta_1^T \psi, \theta_2^T \psi) \leq \rho(J) \| \theta_1 - \theta_2 \|$ for some positive increasing function $\rho$ which is some multiple of polynomial in $J$.

A sufficient condition for Condition (P1) is that each element of $\theta_j$ follows some distributions which has continuous positive density function in its support, e.g., gamma distribution, exponential distribution. It also holds for multivariate normal distribution and Dirichlet distributions if the parameters lie in a fixed compact set. A sufficient condition for (P2) is that $
abla(\theta_j \in S(M_n)) \leq \exp(-c_2 M_n^r)$ for every $j = 1, \ldots, J$.

Theorem B.0.1 Suppose the Conditions (P1), (P2) and (C1) hold and the testing condition (1.6) holds with some semi-metric $d_n = e_n$ on $F_n$. Assume that for a positive sequence of numbers $\epsilon_n \to 0$ with $n\epsilon_n^2 \to \infty$, for $J_n > 3, M_n > 0$ and some sufficiently large constant $b > 0$, the following four conditions hold:

\begin{align*}
\inf_{\theta \in [-H,H]^J} d(f^*, \theta^T \psi_{J_n}) &\leq \epsilon_n, \quad \text{(B.1)} \\
J_n \log(J_n) + J_n \log \left( \frac{1}{\epsilon_n} \right) &\lesssim n\epsilon_n^2, \quad \text{(B.2)} \\
J_n \log \left( \frac{nJ_nM_n}{\epsilon_n} \right) &\lesssim n\epsilon_n^2, \quad \text{(B.3)} \\
\log J_n + bn\epsilon_n^2 &\leq c_2 M_n^r. \quad \text{(B.4)}
\end{align*}

Suppose also that there exist constants $a, c > 0$ such that for every $f_1, f_2 \in F_n$,

\begin{align*}
d_n(f_1, f_2) &\lesssim n^d d(f_1, f_2)^a, \quad \text{(B.5)}
\end{align*}

and for every $f$ such that $d(f^*, f)$ is sufficiently small,

\begin{align*}
K(p_f^{(n)}, p_f^{(n)}) &\lesssim nd^2(f^*, f). \quad \text{(B.6)}
\end{align*}

Then $\Pi(d_n(f, f^*) > K_n \epsilon_n | Y) \to 0$, for every $K_n \to \infty$.

The following theorem is more general as it states the condition (1.6) using a fairly arbitrary class of sets.

Theorem B.0.2 Suppose that for a sequence of positive numbers $\epsilon_n$ such that $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$, there exist a sequence of sets $B_{n,k}$ such that $\{ F_n \cap \{ f : d_n(f^*, f) > \epsilon_n \} \} \subset \bigcup_{k=1}^{N_n} B_{n,k}$, and a sequence of tests $\{ \phi_{n,k} : k = 1, \ldots, N_n \}$ for some positive constants $C_1, C_2$, the following two conditions hold:

\begin{align*}
P_f^{(n)} \phi_{n,k} &\leq \exp(-C_1 n\epsilon_n^2), \quad \sup_{f \in B_{n,k}} P_f^{(n)}(1 - \phi_{n,k}) \leq \exp(-C_2 n\epsilon_n^2), \quad \text{(B.7)}
\end{align*}
and
\[ C_1 n \epsilon_n^2 - \log N_n \to \infty. \quad (B.8) \]

Suppose that in addition, for \( J_n > 3, M_n > 0 \) and some sufficiently large constant \( b > 0 \), the following hold:

\[
\begin{align*}
\inf_{\theta \in [-H,H]^J_n} d(f^*, \theta^T \psi J_n) &\leq \epsilon_n, \quad (B.9) \\
J_n \log(J_n) + J_n \log \left( \frac{1}{\epsilon_n} \right) &\lesssim n \epsilon_n^2, \quad (B.10) \\
\log J_n + b n \epsilon_n^2 &\leq M_n^\tau. \quad (B.11)
\end{align*}
\]

Assume also that for every \( f \) with \( d(f^*, f) \) sufficiently small,
\[ K(p_{\frac{f^*}{n}}^{(n)}, p_f^{(n)}) \lesssim nd^2(f^*, f). \quad (B.12) \]

Then \( \Pi(d_n(f, f^*) > K_n \epsilon_n | Y) \to 0 \), for every \( K_n \to \infty \).