ABSTRACT

YOO, WILLIAM WEIMIN. Sup-norm Posterior Convergence Rates for Regression Models with Application to Estimating the Location of Function Maximum. (Under the direction of Subhashis Ghoshal.)

In the setting of nonparametric multivariate regression with unknown error variance σ^2, we propose a Bayesian method to estimate the regression function f and its characteristics, with particular interest in estimating its mixed partial derivatives, location of function maximum $\boldsymbol{\mu}$ and its value maximum M. Our prior consists of representing f using tensor-product B-splines with normal basis coefficients, where σ^2 is either estimated using empirical Bayes or endowed with an inverse-gamma prior. The frequentist properties of the resulting posterior distributions and credible sets are studied extensively. In particular, we establish pointwise, L_2-and sup-norm posterior convergence rates for f and its mixed partial derivatives, and show that they coincide with the minimax rates. Also, pointwise, L_2-and sup-norm credible sets are constructed to reflect their estimation uncertainties. Under appropriate conditions, we show that they have guaranteed frequentist coverage with optimal size up to a logarithmic factor. We extent our results on posterior convergence rates and credible sets to anisotropic f, such that f has different smoothness in each dimension. In addition, we introduce new results on tensor-product B-splines.

We derive an inequality that relates sup-norm distance of partial derivatives of f with Euclidean distance of $\boldsymbol{\mu}$, and also prove a similar inequality between f and M. Hence, posterior convergence rates and credible sets established for f and its mixed partial derivatives translate directly to $\boldsymbol{\mu}$ and M. Moreover, we propose a two-stage Bayesian procedure to estimate these two quantities. In the first stage, we endow f with the tensor-product B-splines prior as described. We then construct credible region for $\boldsymbol{\mu}$ based on the posterior distributions of f and its mixed partial derivatives. We sample from this region and represent f as a multivariate polynomial with normal coefficients. The corresponding induced posterior distributions of $\boldsymbol{\mu}$ and M will have improved posterior convergence rates. In particular, for α-smooth f, the optimal single-stage posterior convergence rates $(\log n/n)^{(\alpha-1)/(2\alpha+d)}$ and $(\log n/n)^{\alpha/(2\alpha+d)}$ for $\boldsymbol{\mu}$ and M improve to $n^{-(\alpha-1)/(2\alpha)}$ and $n^{-1/2}$ respectively when $\alpha > 1 + \sqrt{1 + d/2}$.

Sup-norm Posterior Convergence Rates for Regression Models with Application to Estimating the Location of Function Maximum

by
William Weimin Yoo

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

2014

APPROVED BY:

Brian Reich	Donald Martin
Arnab Maity	Subhashis Ghoshal Chair of Advisory Committee

DEDICATION

To my parents, my brothers and my friends. Without them, I would not have survived this long and ardours journey.

BIOGRAPHY

The author was born in Georgetown, Penang, Malaysia on 3rd February, 1986. After finishing high school, he enrolled in the American Degree Transfer program to pursue a Bachelor's degree in Actuarial Science. As a result, he spent two years at INTI International College, Nilai, and two years abroad at Drake University, Des Moines, Iowa. After graduating, he worked briefly as an Actuarial Executive at Allianz Life Insurance Malaysia. While trying to introduce stochastic models for future reserve computation, he realized that his current statistical knowledge is limited and needed more advanced methods. Therefore, he decided to do a Master's in statistics at the University of Waterloo, Canada. His passion for research and his inclination toward academia lead him to further pursue a PhD in statistics at the North Carolina State University in the U.S.A.

The author is a foodie and enjoys Penang hawker food and Durians. In addition, he likes traveling and reading comics.

ACKNOWLEDGEMENTS

I would like to thank Dr. Subhashis Ghoshal for being my dissertation advisor and also a mentor. I would also like to thank my committee members Dr. Brian Reich, Dr. Donald Martin and Dr. Arnab Maity for their suggestions and encouragements. Lastly, I thank the Department of Statistics and the Graduate School of North Carolina State University for funding me throughout my PhD studies and research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Bayesian nonparametrics

Bayesian nonparametrics is the application of Bayesian principles to nonparametric models for parameter estimation and inference. These models contain infinite-dimensional parameters and do not assume specific functional form for the true distribution. Bayesian inference on these models consists of two main steps: specifying prior distribution on these infinite-dimensional parameters, and combine the prior with the likelihood through Bayes formula to produce the corresponding posterior distribution.

Unlike parametric models, which have overly restrictive and often unjustifiable assumptions on the data generating mechanism, these models are flexible and can be applied to a wide variety of statistical problems. A canonical example is density estimation where the parameter is the density function itself. Another example, which will be the main focus of this dissertation is regression, where the regression function is unknown and assumed to belong to a function class. Additional examples include estimation of spectral density in time series, hazard rate function in survival analysis, transition density of Markov chain, nonparametric hypothesis tests and others. The Bayesian paradigm provides a unified and consistent framework for parameter estimation and inference, since all quantities of interest such as point estimator and credible regions can be derived solely from the posterior distribution. In addition, the development of Markov Chain Monte Carlo (MCMC) sampling methodologies has provided a tremendous boost to the use of Bayesian nonparametrics in practical applications. For example in genetics, machine learning, neuroscience, engineering and many more.

However, the myriad of opportunities offered by Bayesian nonparametrics come with several caveats. Unlike parametric Bayesian procedures, eliciting priors on infinite-dimensional spaces

poses unique conceptual and technical challenges. Natural choices of prior for certain statistical problem might lead to an inconsistent posteriors, i.e., a posterior distribution that does not concentrate at the true value of the parameter asymptotically. The inclusion of the true parameter in a prior's topological support does not by itself guarantee posterior consistency. This stems from the fact that it is possible to construct priors on infinite-dimensional spaces that will not be overwhelmed by the data even if we had indefinitely large sample sizes.

To avoid these difficulties, a list of default priors are introduced and catalogued for different problems. Such priors are constructed through some automatic mechanism and their mass is spread across the parameter space. Also, they are impartial with respect to any particular parameter, and they have low information content compared to the data. They are analogues to improper priors in parametric Bayesian inference, but these default priors are invariantly proper on infinite-dimensional spaces. Examples of default priors include Dirichlet process (Ferguson, 1973), Dirichlet mixtures (Ferguson, 1983, Lo, 1984), Gaussian process (Leonard, 1978, Lenk, 1988) and random series.

Sampling from the posterior distribution of infinite-dimensional parameters requires innovative and novel sampling strategies. The basic idea is to decompose the posterior into more elementary components, and reduce the problem into sampling from these finite-dimensional pieces. Often, the sampling problem is cast into a hierarchical framework by introducing latent variables. Examples include Gibbs sampling based on Pólya urn (MacEachern, 1994, Escobar and West, 1995) with its variants such as stick breaking Gibbs (Ishwaran and James, 2001) and no-gaps algorithm (MacEachern and Müller, 1998), Metropolis-Hastings with Gibbs sampling (Neal, 2000), slice sampling (Walker, 2007), non-MCMC sampling based on random series (Shen and Ghosal, 2014), variational methods (Blei and Jordan, 2006), and direct sampling if the posterior distribution has explicit expression due to conjugacy.

In the next section, we discuss Bayesian nonparametric regression, which is one of the main applications of Bayesian nonparametrics.

1.2 Bayesian nonparametric regression

The Bayesian nonparametric regression problem is as follows: suppose we have noisy observations from some unknown function $f : \mathbb{R}^d \to \mathbb{R}$ such that

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i, \tag{1.1}$$

where the covariates $\boldsymbol{X}_i \in U$ with U some bounded open set in $\mathbb{R}^d$, and ε_i are independent and identically distributed (i.i.d.) errors. Our aim is to estimate f and perhaps some of its other characteristics. Since f is unknown, we will typically assume that f belongs to some known function space $\mathcal{F}$. A Bayesian will then put a prior on f, compute its posterior, use the posterior mean as estimate for f and quantifies its uncertainty by constructing credible sets for f based on posterior variance.

There is a close connection between regression with spline theory since f can be modeled as splines in some applications. In this framework, Kimeldorf and Wahba (1970) and Wahba (1978) appear to be the first to use Gaussian process as prior for f, where they used what is called the integrated Brownian motion polynomially released at zero. Cox (1993) expanded f using random series with eigenfunctions as basis. de Jonge and van Zanten (2012) represented f using tensor-product B-splines and studied its asymptotic properties.

Certain characteristics of f are also of interest. For example the derivatives of the regression function assuming that the function is smooth enough. The location of the function maximum and its maximum value are useful characteristics that we will fully investigate in this dissertation (see Figure 1.1). Other characteristics include location of change-points or jump discontinuities and estimating the inverse of a regression function evaluated at a point in the function's range.

The regression function and its associated characteristics can be estimated in one or multi-stages. In this context, single-stage estimation means we use a fixed design to collect all samples for estimation in one go. For multi-stage estimation, we employ instead sequential design to collect samples, i.e., one collects samples based on information gathered from previous samples, and estimation is conducted through stages. It has been proven by many research papers that multi-stage estimation improves the accuracy of the estimate, and it is possible to construct estimators in sequential designs that have faster convergence rates than the minimax rates for single-stage procedures. Examples in the literature include two-stage estimation of location of function maximum (Belitser et al., 2012), change-points (Lan et al., 2009) and inverse regression (Tang et al., 2011).

To start, we describe in detail some default priors for regression functions that are used in this dissertation.

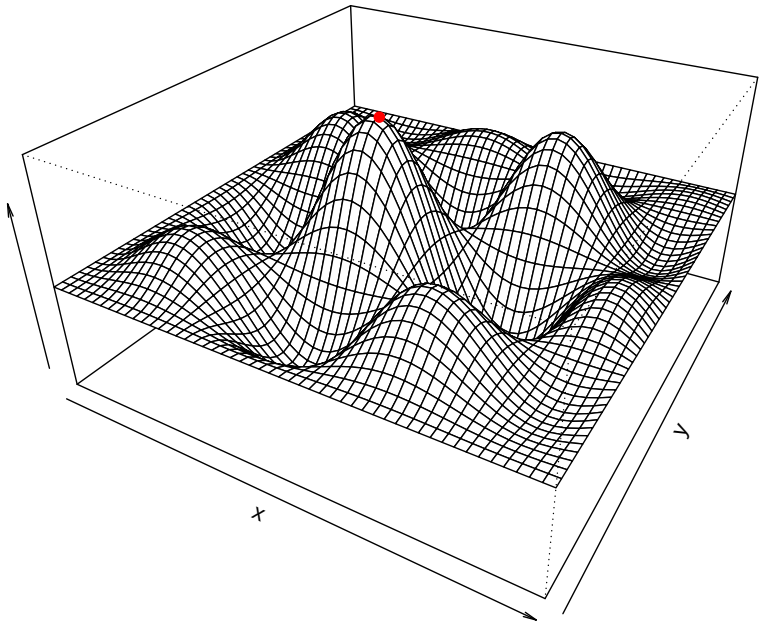


Figure 1.1: The red dot is the point of global maximum, we are interested in determining the location of this point in the x, y plane below and its maximum value.

1.3 Priors on regression functions

1.3.1 Gaussian process

Let $X = \{X(\boldsymbol{t}) : \boldsymbol{t} \in U\}$ be a stochastic process indexed by $U \subseteq \mathbb{R}^d$. Define its mean function as $\mathrm{E}[X(\boldsymbol{t})] = \mu(\boldsymbol{t})$ and covariance function as $\mathrm{Cov}[X(\boldsymbol{t}), X(\boldsymbol{s})] = V(\boldsymbol{t}, \boldsymbol{s})$ for any $\boldsymbol{t}, \boldsymbol{s} \in \mathbb{R}^d$. We say X is a Gaussian process if for any n finite collection of points $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$,

$$[X(\boldsymbol{t}_1), \ldots, X(\boldsymbol{t}_n)] \sim \mathrm{N}_n(\boldsymbol{\mu}, \boldsymbol{V}), \tag{1.2}$$

where $\boldsymbol{\mu} = (\mu(\boldsymbol{t}_1), \ldots, \mu(\boldsymbol{t}_n))^T$ and $\boldsymbol{V}$ is the Gram matrix of the covariance function $V(\cdot, \cdot)$, i.e., $v_{ij} = V(\boldsymbol{t}_i, \boldsymbol{t}_j)$ for $1 \leq i, j \leq n$.

Since any realization of X is a function with domain U, X is a default prior on the function space $\mathcal{F}$. If additional functional properties are desired such as continuity and differentiability, we modify the prior by imposing certain conditions on its covariance function. It turns out that mixed partial derivatives of X are also Gaussian processes. The support of X is given by the completion of its reproducing kernel Hilbert space (for an introduction see van der Vaart

and van Zanten, 2008a). Specific instances of Gaussian processes include the Brownian motion or Wiener process (Figure 1.2), integrated Brownian motion, fractional Brownian motion, Riemann-Liouville process (see Section 4.2 of van der Vaart and van Zanten, 2008b).

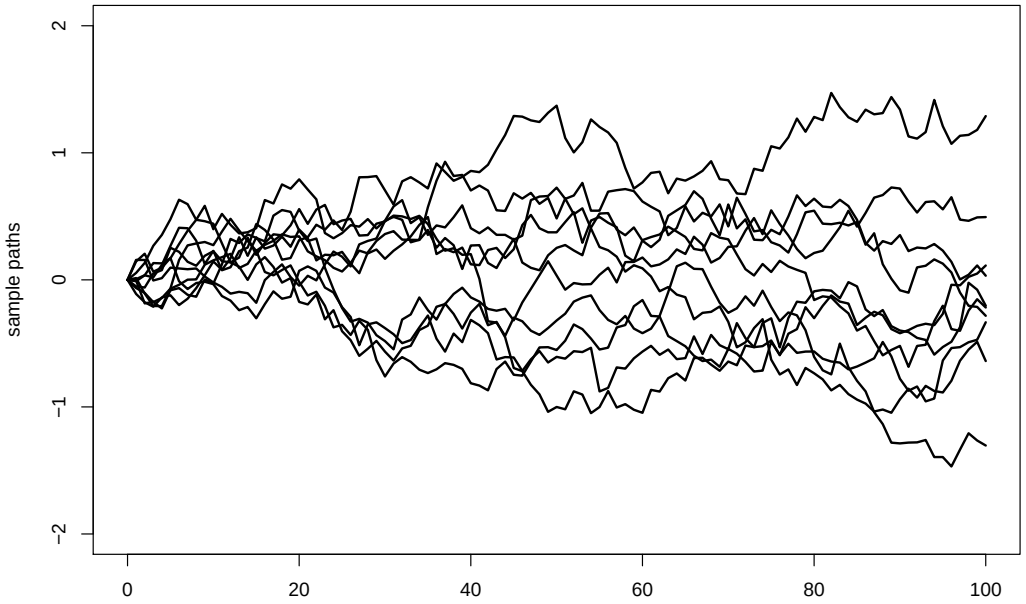


Figure 1.2: Sample paths generated from a Brownian motion, each of this path is a continuous function in $\mathbb{R}$.

1.3.2 Random series

Let $f : \mathbb{R} \to \mathbb{R}$ be the regression function in (1.1) such that $f \in \mathcal{F}$. Suppose $\{\phi_j : j \geq 1\}$ is a set of basis for $\mathcal{F}$, then we can represent f truncated at level J as

$$f(x) = \sum_{j=1}^{J} \theta_j \phi_j(x). \tag{1.3}$$

The above series representation is a random function and hence its name random series. Depending on the underlying function space $\mathcal{F}$, commonly used basis functions include the polynomials, Fourier basis, B-splines and wavelets. Prior elicitation can be accomplished by putting separate independent priors on $(\theta_1, \ldots, \theta_J)^T$ and J respectively. This will then induce a prior on f. For multivariate f, we replace ϕ_j by its tensor-product and sum over the different J's in all dimensions. It turns out that by the Karhunen-Loève theorem, any Gaussian process can

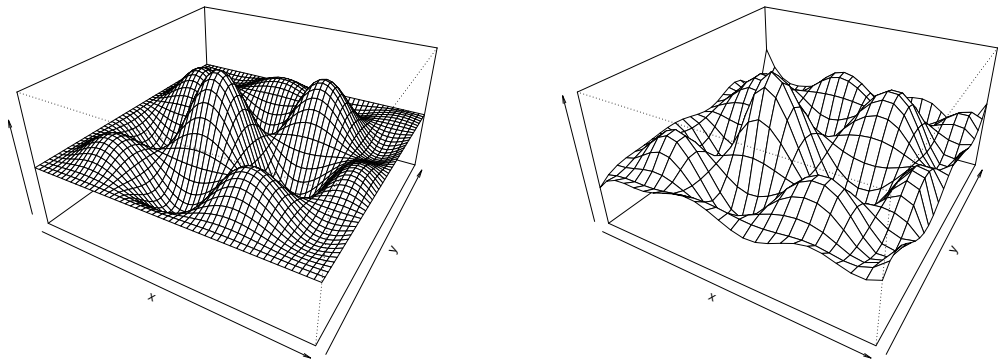


Figure 1.3: The figure on the left is the true f, while on the right is the posterior mean of f using bivariate cubic B-splines random series.

be represented as a infinite random series with normal coefficients, and the eigenfunctions of its covariance function as basis. Random series is a powerful tool in Bayesian nonparametrics. With appropriate choices of basis and priors on the coefficients, random series can approximate a wide variety of functions well (see Figure 1.3).

1.4 Consistency and convergence rates

Due to the sheer size of an infinite-dimensional parameter space, there is no guarantee that the posterior distribution will converge to the point mass of the truth, even if the truth is contained in the support of the prior. This property, which we call posterior consistency holds in finite-dimensional parametric models under relatively mild conditions, but a much more involved analysis is needed to establish consistency in nonparametric models. Freedman (1963) constructed an example using geometric distribution showing posterior inconsistency. Other examples can be found in Diaconis and Freedman (1986) for location models and de Blasi et al. (2013) for Gibbs-type priors. The celebrated Schwartz (1965) paper introduced the concept of using Kullback-Leibler neighborhoods and test with exponentially small errors to prove consistency for dominated statistical models. Extensions of Schwartz's result were provided by Barron et al. (1999) and Ghosal et al. (1999).

Since the definitions given below apply to more general parameter spaces, we will temporarily leave the regression framework. Let Θ be the parameter space endowed with a metric d and a Borel sigma-field $\mathscr{B}$. Let $\theta_0 \in \Theta$ be the true parameter. Suppose we observed $\boldsymbol{X} = (X_1, \ldots, X_n)^T \sim P_\theta^{(n)}$, where $P_\theta^{(n)}$ is the joint distribution of $\boldsymbol{X}$ indexed by $\theta \in \Theta$. Also,

let $P_{\theta_0}^{(n)}$ be the corresponding true joint distribution and define $\mathfrak{X}^{(n)}$ to be the sample space of $\boldsymbol{X}$. Note we do not assume independence or i.i.d. structure of the observations. Given a prior Π on the Borel sets of Θ, let $\Pi(\cdot|\boldsymbol{X})$ denote the corresponding posterior. We say that the posterior distribution $\Pi(\cdot|\boldsymbol{X})$ is consistent at θ_0 if for any $\epsilon > 0$,

$$\Pi(\theta : d(\theta, \theta_0) > \epsilon|\boldsymbol{X}) \to 0 \tag{1.4}$$

in $P_{\theta_0}^{(n)}$-probability as $n \to \infty$. Consistency is a frequentist concept. But applying it in the Bayesian context means that as we collect more and more samples, the posterior distribution will converge to the point mass of θ_0. Another implication of consistency is that with large enough samples, posterior distributions computed using different priors will agree (their distances in weak topology will go to zero). Hence Bayesians with different prior beliefs will have their opinions merged, and will ultimately reach a consensus iff consistency holds.

We can take this concept one step further by asking how fast ϵ can go to 0 with n maintaining the property that the posterior probability in (1.4) goes to zero. This rate of convergence can be succinctly described as follows: given a positive sequence ϵ_n such that $\epsilon_n \to 0$, we say that posterior distribution $\Pi(\cdot|\boldsymbol{X})$ converges to θ_0 at the rate ϵ_n with respect to the metric d if for any sequence $M_n \to \infty$,

$$\Pi(\theta : d(\theta, \theta_0) > M_n\epsilon_n|\boldsymbol{X}) \to 0 \tag{1.5}$$

in $P_{\theta_0}^{(n)}$-probability as $n \to \infty$. We are naturally interested in the fastest rate possible, i.e., smallest ϵ_n. Often the optimal rate will be equal to the minimax estimation rate in the frequentist sense up to a constant or logarithm factor. In our case, Θ is the function space $\mathcal{F}$ with true function f_0. We will study supremum norm (sup-norm for short) convergence rates of $f \in \mathcal{F}$ and its derivatives, where the metric used is the function supremum norm, i.e., $d(f, f_0) = \|f - f_0\|_\infty = \sup_{x \in U} |f(x) - f_0(x)|$. Note that estimating the derivatives of f can be viewed as an inverse problem.

If we had analytic expression for the posterior through conjugacy, then (1.5) can be established by Markov's inequality. If not, then more elaborate theory is needed. In this case, there are usually four conditions that we need to verify depending on the dependence structure of $\boldsymbol{X}$. Instead of providing the exact technical details of these conditions, we choose to provide heuristic descriptions of them. Interested readers can refer to Ghosal et al. (2000) or Ghosal and van der Vaart (2007) for a detailed discussion. The foremost condition needed is the existence of test function $\phi_n : \mathfrak{X}^{(n)} \to \{0, 1\}$ with exponentially small error for the following hypothesis

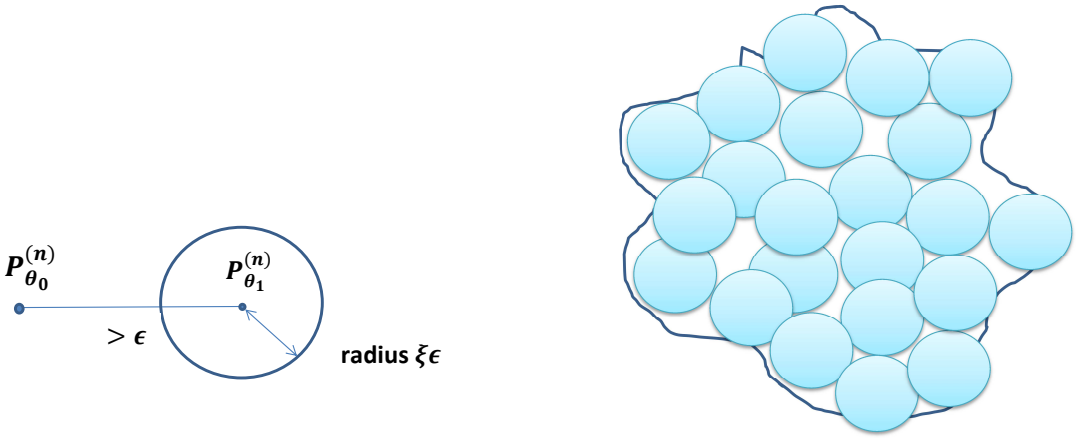


Figure 1.4: Schematic diagrams: testing problem in (1.6) on the left while the right shows the number of balls needed to cover Θ_n (covering number).

testing problem. Let $\xi, \epsilon_n > 0$ and some $\theta_1 \in \Theta$. Suppose $\boldsymbol{X} \sim P$ for some joint probability distribution P and we wish to test the following (see left panel of Figure 1.4):

$$H_0 : P = P_{\theta_0}^{(n)}, \qquad H_1 : P = P_{\theta}^{(n)} \quad \text{for some } \theta : d(\theta, \theta_1) > \xi\epsilon_n. \tag{1.6}$$

A test function ϕ_n for this problem is such that it rejects the null if $\phi_n(\boldsymbol{X}) = 1$. We assume that there exists ϕ_n such that its Type I and II errors are exponentially small. Construction of such test function typically begins by considering a series of hypothesis tests where the alternative is a small ball separated from the null value of the parameter. One then combine these tests to obtain the desired ϕ_n. Generally, imposing this condition all over the parameter space is impossible for non-compact parameter spaces, because the number of small balls needed to cover the alternative parameter space in (1.6) is infinite. A useful technique is then to form such exponentially powerful tests only within a sieve, which is a sequence of approximating compact subsets $\Theta_n \subseteq \Theta$ such that $\cup_{n=1}^{\infty} \Theta_n = \Theta$. Then the testing condition can be satisfied on Θ_n by controlling the covering numbers of Θ_n, i.e., the number of small balls required to cover Θ_n. This then effectively reduces the testing requirement to a condition on the entropy (logarithm of covering number) of the sieve. The second condition requires the prior to put sufficient mass near the true θ_0, where distances are measured using Kullback-Leibler divergence. The last condition needs the complement of Θ_n to receive small prior mass, so that we can substitute Θ with Θ_n in the first testing or entropy condition. If the prior is a Gaussian process, then more explicit sufficient conditions are given by van der Vaart and van Zanten (2008b).

1.5 Coverage and credible region

Given parameter space Θ. Let $\theta_0 \in \Theta$ be the true parameter. We say that a subset $C(\boldsymbol{X})$ of Θ is a $1-\gamma$ credible region or set if the posterior mass of $C(\boldsymbol{X})$ is $1-\gamma$ for $\gamma \in [0,1]$. That is,

$$\Pi(\theta : \theta \in C(\boldsymbol{X})|\boldsymbol{X}) = 1-\gamma. \tag{1.7}$$

Credible regions are used to quantify the uncertainty of estimating θ, with the posterior spread indicating the margin of error in estimation. Unlike frequentist's confidence regions, (1.7) is a true probability statement and we can say the posterior probability of θ falling in $C(\boldsymbol{X})$ is $1-\gamma$. We note that $C(\boldsymbol{X})$ depends on the data $\boldsymbol{X}$. In parametric Bayesian inference, various criteria were proposed as to what constitute the best $1-\gamma$ credible region, and methods were developed to construct them. For example, constructing the highest posterior density (HPD) credible region, where we try to find the smallest $C(\boldsymbol{X})$ region that satisfies (1.7). In finite-dimensional setting, the credible region constructed will have the right asymptotic frequentist coverage in most cases, i.e., as $n \to \infty$,

$$P_{\theta_0}^{(n)}(\theta_0 \in C(\boldsymbol{X})) = 1-\gamma. \tag{1.8}$$

However, this is no longer true for credible region constructed for infinite-dimensional parameters. Using the Gaussian white noise model, Cox (1993) concluded that Bayesian credible sets at any level have frequentist coverage zero for almost all true θ_0. This disparaging conclusion was furthered enforced by Freedman (1999) and he provided instances of credible sets with asymptotically zero or sub-optimal coverage. It may be noted that their formulation consists of drawing true function (equivalently sequences) from the prior itself, and failure is shown in the almost sure sense with respect to the prior. The main cause of this is that in finite-dimensional problems, the Bernstein-von Mises theorem holds. That is, the posterior distribution converges to a normal distribution centered at the maximum likelihood estimate, with the inverse Fisher information as its variance. Hence, if the sample size is large enough, credible sets will behave like confidence regions and they will have asymptotically the right coverage. However, there is often no Bernstein-von Mises theorem for nonparametric models and credible sets constructed for these models are not guaranteed to have the right coverage. On the other hand, Leahu (2011) and Castillo and Nickl (2013) showed that if the parameter space is extended beyond ℓ_2 sequence space, and normal priors with large variances are used for each component, then the credible regions do possess adequate frequentist coverage.

It was only after the work of Knapik et al. (2011) that the entirety of the problem is known. Zero coverage occurs when the prior oversmoothes the truth. They further showed that by un-

dersmoothing the prior, we can construct conservative credible sets with coverage tending to 1. They then applied their findings to the problem of heat equation in Knapik et al. (2013). On the other hand, Castillo and Nickl (2013) extended Leahu (2011) work beyond conjugate priors and obtained parametric posterior convergence rates under norms weaker than the ℓ_2-norm, called negative Sobolev norms. Assuming a weak Bernstein-von Mises phenomenon, Castillo and Nickl (2013) constructed credible sets with right frequentist coverage for white noise models. They also showed that Bernstein-von Mises result holds for the negative Sobolev norm under appropriate condition on the prior. Recent research considers the issue of constructing adaptive credible sets, where these sets have guaranteed frequentist coverage with radius that adapts to the unknown regularity of θ_0, or function smoothness if Θ is a function space (see Hoffman et al., 2013 and Szabo et al., 2014 for more details).

Motivated by the previous discussion, we will consider a reformulation of the Bayesian credibility problem. For a given sequence $L_n \to \infty$, our aim is to find a subset $C(\boldsymbol{X}) \subseteq \Theta$ with the following properties, uniformly for $\theta_0 \in D$ where D is a ball in Θ:

1. $\Pi(\theta \in C(\boldsymbol{X})|\boldsymbol{X}) = 1 - \omega_n$ in $P_{\theta_0}^{(n)}$-probability,
2. $P_{\theta_0}^{(n)}(\theta_0 \in C(\boldsymbol{X})) \to 1$,
3. $\text{diam}(C(\boldsymbol{X})) = O_{P_{\theta_0}^{(n)}}(L_n \epsilon_n)$,

where ϵ_n is the minimax convergence rate for θ. The first condition creates a $1 - \omega_n$ credible region, where $\omega_n \in [0, 1]$ can be fixed or a sequence tending to 0. If ω_n is a constant, then the second condition holds if we inflate the radius of $C(\boldsymbol{X})$ by some factor, or we use a prior on θ that is less regular than the true regularity of θ_0 (undersmoothing if Θ is a function space with function smoothness as a measure of regularity). The second condition says that the frequentist coverage under the true distribution of the credible region goes to one asymptotically. The last condition limits the size of the constructed credible region by forcing it to have diameter at most $L_n \epsilon_n$.

1.6 Research questions and our contributions

There are roughly three problems that we hope to address in this dissertation. In what follows, our parameter space Θ is the function space $\mathcal{F}$. Consider again the nonparametric regression problem in (1.1). Let f_0 be the true f. Assume that f has a unique maximum at $\boldsymbol{\mu} \in U \subseteq \mathrm{R}^d$ with maximum value $M = f(\boldsymbol{\mu})$.

1.6.1 Sup-norm convergence rates for f and its mixed partial derivatives

Convergence rates under L_2-norm and Hellinger distance were the first to be derived and studied, and the corresponding theory is well established. However, theory for convergence rates in stronger norms such as the supremum norm is limited and has only caught attention in recent years. In addition, there is an inequality relating the sup-norm distance of partial derivatives of f and the Euclidean distance of $\boldsymbol{\mu}$. Also, similar inequality holds between f and M. This implies that the posteriors of f and its partial derivatives induce the same sup-norm rate on the posteriors of M and $\boldsymbol{\mu}$. Therefore, these two factors motivate the study of sup-norm posterior convergence rates for f and its mixed partial derivatives.

The first paper to tackle this problem was Giné and Nickl (2011), where they derived rates in L_r-metrics for all $1 \leq r \leq \infty$, and they constructed new tests with exponentially small error probabilities based on Talagrand's inequality to handle L_∞-convergence. However their rates are sub-optimal for $r > 2$, and they showed using a white noise model that optimal rate is achievable for conjugate priors with diagonal structure. Castillo (2014) introduced techniques based on semiparametric Bernstein-von Misses type results to obtain optimal sup-norm rates, for the white noise model and density estimation using priors based on wavelet series. He split the sup-norm distance into smaller components consisting of semiparametric functionals, and the corresponding influence function is estimated using various uniform approximation schemes. Hoffman et al. (2013) derived optimal adaptive sup-norm convergence rate for white noise model using spike and slab priors. They then generalized their result by introducing sieve priors that achieve the same adaptive sup-norm rate. For density estimation, Scricciolo (2014) employed Gaussian kernel mixtures by endowing the mixing distribution with a Pitman-Yor or normalized inverse Gaussian process. She obtained adaptive posterior convergence rates in L_r-metric for $1 \leq r \leq \infty$.

Most of the papers on sup-norm convergence rate focus on density estimation and white noise model with known error variance, and contain only brief mentions of nonparametric regression. Also, all papers considered univariate models. Most papers use random series prior with wavelets as basis. In this dissertation, we will derive sup-norm convergence rates for multivariate nonparametric regression in the setting of (1.1). Instead of using wavelets, we represent f using random series with tensor-product B-splines as basis. Although both B-splines and wavelets are widely used in this context, B-splines have the added advantage in that mixed partial derivatives of f are expressible also as tensor-product B-splines. Hence the same posterior analysis for f carries over to its mixed partial derivatives. This allows us to study posterior convergence rates for derivatives as well, a topic that was largely unaddressed so far, except

implicitly as inverse problems in white noise models. Moreover, we allow function to have different smoothness in different dimensions, which is a useful generalization. In addition, we let σ be unknown, thus making the results more relevant for practical applications. The sup-norm convergence rates under our proposed prior coincide with the minimax rates of estimating a function and its mixed partial derivatives.

1.6.2 L_∞-credible sets for f and its mixed partial derivatives

The issue of constructing nonparametric confidence set has been studied by many authors. In particular, Juditsky and Lambert-Lacroix (2003), Robins and van der Vaart (2006) and Cai and Low (2006) showed that it is impossible to construct adaptive confidence sets that are simultaneously honest, and have radius that adapt to the underlying function smoothness. Bayesian analog of a confidence set is given by a credible region in which has a guaranteed posterior probability. Construction of credible sets in the context of function estimation is a challenging problem, since optimal smoothing typically makes the order of the bias and variability the same. They are typically easy to obtain numerically from posterior sampling. Hence it is of considerable interest to know that a credible region possesses asymptotically correct frequentist coverage.

The issue of constructing credible sets for nonparametric models, and evaluating their performance has received very little attention in the Bayesian nonparametrics literature, with the main references mentioned in Section 1.5. Most of the credible sets for f in these papers are sets in sequence ℓ_2-or function L_2-Hilbert spaces, while sets in L_∞-Banach space are not well understood. Using our tensor-product B-spline prior, we propose methods to construct L_∞-credible sets for f and its mixed partial derivatives that satisfy the criteria listed in Section 1.5. In addition, we will also construct pointwise credible intervals for $f(\boldsymbol{x})$ at any $\boldsymbol{x} \in (0,1)^d$ and L_2-credible sets for f. Under appropriate conditions, all pointwise, L_2-and L_∞-credible sets are shown to have frequentist coverage tending to 1 asymptotically with optimal size up to a logarithmic factor. Moreover, we carried out extensive simulations to compare finite sample performance of our pointwise credible intervals and credible bands with pointwise confidence intervals and confidence bands proposed by Zhou et al. (1998).

1.6.3 Bayesian estimation of location of maximum $\boldsymbol{\mu}$ and maximum value M

The estimation of $\boldsymbol{\mu}$ and M in one-dimensional regression setting is well studied in the frequentist literature. For example Müller (1985, 1989) estimated $\boldsymbol{\mu}$ via kernel estimators, Shoung and Zhang (2001) used least squares and Klemelä (2005) considered adaptive estimation of $\boldsymbol{\mu}$.

Kiefer and Wolfowitz (1952) proposed a Robbins-Monro type algorithm using sequential sampling to estimate $\boldsymbol{\mu}$ in one-dimension. Blum (1954) extended their method to higher dimensions. Mokkadem and Pelletier (2007) provided a method to simultaneously estimate $\boldsymbol{\mu}$ and M in a sequential design setting. Belitser et al. (2012) proposed a two-stage procedure for estimation of these two quantities and studied its asymptotic properties.

However, the papers described above used frequentist methods in estimating $\boldsymbol{\mu}$ and M, and a Bayesian equivalent is lacking. Therefore, we will study this problem in a Bayesian setting. Inspired by Belitser et al. (2012), we will propose a Bayesian two-stage procedure to estimate $\boldsymbol{\mu}$ and M. We study its asymptotic properties and investigate its finite sample performance through a series of simulations. In particular, we show that our proposed Bayesian two-stage procedure performs better in terms of mean square error than other common single-stage methods, and is as good as the frequentist two-stage procedure proposed in Belitser et al. (2012). We hope that this dissertation is a first step towards providing a Bayesian solution to this problem.

1.7 Notations and preliminaries

We describe notations and technical definitions used throughout this dissertation. All asymptotic relations and symbols described below refer to the regime $n \to \infty$. Given two numerical sequences a_n and b_n, $a_n = O(b_n)$ or $a_n \lesssim b_n$ means a_n/b_n is bounded, while $a_n = o(b_n)$ or $a_n \ll b_n$ means $a_n/b_n \to 0$. Also, $a_n \asymp b_n$ means $a_n = O(b_n)$ and $b_n = O(a_n)$. Furthermore, $a_n \sim b_n$ is $a_n/b_n \to 1$. For stochastic sequences X_n, $X_n = O_P(a_n)$ means that $P(|X_n| \leq Ca_n) \to 1$ for some constant $C > 0$, while $X_n = o_P(a_n)$ means $P(|X_n| \leq \epsilon) \to 1, \forall \epsilon > 0$. Let $\mathbb{N} = \{1, 2, \ldots\}$ be the set of natural numbers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

Vectors are represented by bold symbols and can be upper or lower case English or Greek letters. All vectors are in column format and the corresponding non-bold letters with subscript denoting the components, i.e., for $\boldsymbol{x}, \boldsymbol{x}_i \in \mathbb{R}^d$, $\boldsymbol{x} = (x_1, \ldots, x_d)^T$ and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^T$ respectively. Let $\|\boldsymbol{x}\|_p$ be the vector p-norm, i.e., $\|\boldsymbol{x}\|_p = (\sum_{k=1}^d |x_k|^p)^{1/p}$. If $p = 2$, we suppress the subscript and simply write $\|\boldsymbol{x}\|$ to be the usual Euclidean norm. Given another vector $\boldsymbol{y}$ of the same dimension, we write $\boldsymbol{x} \leq \boldsymbol{y}$ if $x_k \leq y_k, k = 1, \ldots, d$. Matrices are written in bold and only upper case English letters are used to denote them. For a given symmetric matrix $\boldsymbol{A}$ of dimension $m \times m$, we denote its (i, j)th element by a_{ij}. Let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote the smallest and largest eigenvalues respectively. Denote $\Lambda(\boldsymbol{A})$ to be the set of eigenvalues of $\boldsymbol{A}$. Define the (r, s) matrix norm of $\boldsymbol{A}$ as $\|\boldsymbol{A}\|_{(r,s)} = \sup\{\|\boldsymbol{A}\boldsymbol{x}\|_s : \|\boldsymbol{x}\|_r \leq 1\}$. In particular, the

$(2,2)$ and (∞,∞) norms are defined as

$$\|\boldsymbol{A}\|_{(2,2)} = \sqrt{\lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})} = |\lambda_{\max}(\boldsymbol{A})|,$$
$$\|\boldsymbol{A}\|_{(\infty,\infty)} = \max_{1\le i\le m}\sum_{j=1}^{m}|a_{ij}| = \max_{1\le j\le m}\sum_{i=1}^{m}|a_{ij}|,$$

where they are the usual operator norms if we view $\boldsymbol{A}$ as a continuous linear operator on $(\mathbb{R}^m, \|\cdot\|)$ and $(\mathbb{R}^m, \|\cdot\|_\infty)$ respectively. In addition, define the matrix max-norm as $\|\boldsymbol{A}\|_\infty = \max_{1\le i,j\le m}|a_{ij}|$. Then the max-norm, $(2,2)$ and (∞,∞) matrix norms are related by $\|\boldsymbol{A}\|_\infty \le \|\boldsymbol{A}\|_{(2,2)} \le \|\boldsymbol{A}\|_{(\infty,\infty)}$. With another symmetric and square matrix $\boldsymbol{B}$ of the same size, $\boldsymbol{A} \le \boldsymbol{B}$ means $\boldsymbol{B}-\boldsymbol{A}$ is non-negative definite and $\boldsymbol{A} < \boldsymbol{B}$ means $\boldsymbol{B}-\boldsymbol{A}$ is positive definite. We say that $\boldsymbol{A}$ is a q-banded matrix if $a_{ij} = 0$ for $|i-j| > q$. Also, we denote $\boldsymbol{I}_m$ as the $m\times m$ identity matrix and $\mathbf{1}_d$ as the vector of ones with dimension d. We define $\mathrm{diag}(a_1,\ldots,a_m)$ to be the diagonal matrix with diagonal elements $(a_1,\ldots,a_m)^T$.

For $f: U\to\mathbb{R}$ on some bounded open set $U\subseteq\mathbb{R}^d$, let $\|f\|_p$ be the L_p-norm, i.e., $\|f\|_p = (\int|f|^p d\nu)^{1/p}$ for some sigma-finite measure ν. If the integrating measure needs to be emphasized, we write $\|f\|_{p,G} = (\int|f|^p dG)^{1/p}$. If $p=\infty$, $\|f\|_\infty = \sup_{x\in U}|f(x)|$. For a real-valued random variable X and a function $\psi:[0,\infty)\to[0,\infty)$, which is nondecreasing, convex and $\psi(0)=0$, we define the Orlicz norm of X as

$$\|X\|_\psi = \inf\left\{C>0 : \mathrm{E}\left[\psi\left(\frac{|X|}{C}\right)\right]\le 1\right\}.$$

We note that if $\psi(x)=x^p$ for $p\ge 1$, then $\|X\|_\psi = \|X\|_p$. We will mainly use the exponential Orlicz norm, $\psi_p(x) = e^{x^p}-1$ for $p=2$ throughout this dissertation. For $I\subset U$, let $f|_I$ denote the restriction of f onto I such that $f(x) = f|_I(x)$ for $x\in I$. Furthermore, let $\mathbb{1}_U(x)$ be the indicator function on U such that $\mathbb{1}_U(x)=1$ for $x\in U$ and 0 otherwise. Define an Euclidean ball of center $\boldsymbol{a}$ with radius R in U to be $B(\boldsymbol{a},R) = \{\boldsymbol{y} : \|\boldsymbol{y}-\boldsymbol{a}\|\le R\}$. For $\delta>0$, let $N(\delta,U,\|\cdot\|)$ be the covering number of U, and is defined as the minimum number of Euclidean balls of radius δ needed such that their union contains U. Let $\mathrm{diam}(U) = \sup_{\boldsymbol{x},\boldsymbol{y}\in U}\|\boldsymbol{x}-\boldsymbol{y}\|$ be the diameter of U. We define U^d to be the d-fold Cartesian product of U, and $|U|$ to be the cardinality of U.

For multi-index $\boldsymbol{i} = (i_1,\ldots,i_d)^T\in\mathbb{N}_0^d$ and a vector $\boldsymbol{x}\in\mathbb{R}^d$, define

$$|\boldsymbol{i}| = \sum_{k=1}^{d} i_k, \qquad \boldsymbol{i}! = \prod_{k=1}^{d} i_k, \qquad \boldsymbol{x}^{\boldsymbol{i}} = \prod_{k=1}^{d} x_k^{i_k}.$$

We denote $\mathcal{P}_n = \text{span}\{\boldsymbol{x}^{\boldsymbol{i}} : |\boldsymbol{i}| \leq n\}$ to be the space of multivariate polynomials of order n. Here, $\text{span}\{a_1, \ldots, a_d\}$ means the set of all possible linear combinations of $a_1, \ldots, a_d$. Furthermore, for $k, r \in \mathbb{N}_0$ and $d \in \mathbb{N}$, we define

$$\mathbb{I}_k = \{\boldsymbol{i} \in \mathbb{N}_0^d : i_1 + \ldots + i_d = k\}, \qquad \mathbb{I}(r) = \bigcup_{k=0}^{r} \mathbb{I}_k.$$

Here we enumerate $\mathbb{I}(r)$ as $\mathbb{I}_0 \cup \mathbb{I}_1 \cup \ldots \cup \mathbb{I}_r$. Also, the elements within each $\mathbb{I}_k$ are arranged lexicographically. For $i \neq j$, $\mathbb{I}_i \cap \mathbb{I}_j = \emptyset$. The cardinality of $\mathbb{I}_k$ and $\mathbb{I}(r)$ are given by

$$|\mathbb{I}_k| = \binom{d+k-1}{d-1}, \qquad |\mathbb{I}(r)| = \sum_{k=0}^{r} |\mathbb{I}_k| = \sum_{k=0}^{r} \binom{d+k-1}{d-1}.$$

Let $\boldsymbol{H}f(\boldsymbol{x}_0)$ be the Hessian matrix of f at $\boldsymbol{x}_0 \in \mathbb{R}^d$, whose (i,j)th entry evaluated at $\boldsymbol{x} = \boldsymbol{x}_0$ is $\partial^2 f(\boldsymbol{x})/\partial x_i \partial x_j|_{\boldsymbol{x}=\boldsymbol{x}_0}$ for $i, j = 1, \ldots, d$. Given a vector $\boldsymbol{r} = (r_1, \ldots, r_d)^T \in \mathbb{N}_0^d$, let $D^{\boldsymbol{r}}$ denote the partial derivative operator

$$\frac{\partial^{|\boldsymbol{r}|}}{\partial x_1^{r_1} \ldots \partial x_d^{r_d}}.$$

If $\boldsymbol{r} = \boldsymbol{0}$, we view $D^{\boldsymbol{0}} f \equiv f$. If $\boldsymbol{r} = \boldsymbol{e}_k$, where $\boldsymbol{e}_k = (0, \ldots, 0, 1, 0 \ldots, 0)^T$ with 1 in the kth position, then we write $D^{\boldsymbol{e}_k} = D_k$. Moreover, we denote $\nabla f(\boldsymbol{x}) = (D_1 f(\boldsymbol{x}), \ldots, D_d f(\boldsymbol{x}))^T$ to be the gradient of f at $\boldsymbol{x}$. For any $\alpha > 0$, let $\lceil \alpha \rceil$ be the smallest integer bigger than or equal to α. We then define the Hölder norm $\|\cdot\|_{\mathcal{H}^\alpha}$ as

$$\|f\|_{\mathcal{H}^\alpha} = \max_{\boldsymbol{r}:|\boldsymbol{r}| \leq m_\alpha} \sup_{x \in U} |D^{\boldsymbol{r}} f(\boldsymbol{x})| + \max_{\boldsymbol{r}:|\boldsymbol{r}|=m_\alpha} \sup_{\boldsymbol{x},\boldsymbol{y} \in U : \boldsymbol{x} \neq \boldsymbol{y}} \frac{|D^{\boldsymbol{r}} f(\boldsymbol{x}) - D^{\boldsymbol{r}} f(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|^{\alpha - m_\alpha}}, \tag{1.9}$$

where $m_\alpha = \lceil \alpha - 1 \rceil$ is the largest integer strictly smaller than α. We introduce the isotropic Hölder function space $\mathcal{H}^\alpha(U)$ of order $\alpha > 0$ with domain U, consisting of functions $f : U \to \mathbb{R}$ such that $\|f\|_{\mathcal{H}^\alpha} < \infty$ and for $\boldsymbol{x}, \boldsymbol{x}_0 \in U$ with constant $C > 0$,

$$|D^{\boldsymbol{r}} f(\boldsymbol{x}) - D^{\boldsymbol{r}} T_{\boldsymbol{x}_0} f(\boldsymbol{x})| \leq C \|\boldsymbol{x} - \boldsymbol{x}_0\|^{\alpha - |\boldsymbol{r}|}, \tag{1.10}$$

where $\boldsymbol{r} \in \mathbb{N}_0^d$, $|\boldsymbol{r}| \leq m_\alpha$ and

$$T_{\boldsymbol{x}_0} f(\boldsymbol{x}) = \sum_{\boldsymbol{i} \in \mathbb{I}(m_\alpha)} \frac{1}{\boldsymbol{i}!} D^{\boldsymbol{i}} f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0)^{\boldsymbol{i}} \tag{1.11}$$

is the Taylor polynomial of order m_α by expanding f around $\boldsymbol{x}_0$. This implies that for all $\boldsymbol{r} \in \mathbb{N}_0^d$ such that $|\boldsymbol{r}| \leq m_\alpha$, $D^{\boldsymbol{r}} f$ is uniformly bounded and Lipschitz on U as a result of the mean value theorem. Also, it follows that f has a unique continuous extension onto the closure of U.

We write $X \sim \mathrm{N}(\xi, \Omega)$ if X has a one-dimensional normal distribution with mean ξ and variance Ω. If $X \sim \mathrm{N}(0,1)$, then X is a standard normal. Define $\Phi(x)$ for $x \in \mathbb{R}$ to be the cumulative distribution function of a standard normal, i.e., $\Phi(x) = \int_{-\infty}^{x} (2\pi)^{-1/2} \exp(-x^2/2)$. We say $\boldsymbol{X} \sim \mathrm{N}_J(\boldsymbol{\xi}, \boldsymbol{\Omega})$ if $\boldsymbol{X}$ has a J-dimensional normal distribution with mean vector $\boldsymbol{\xi}$ and covariance matrix $\boldsymbol{\Omega}$. For a random function $\{X(t), t \in U\}$, we say that $X \sim \mathrm{GP}(\xi, \Omega)$ if X is a Gaussian process with $\mathrm{E}X(t) = \xi(t)$ and $\mathrm{Cov}(X(s), X(t)) = \Omega(s,t)$ for any $s, t \in U$. We say that X is a sub-Gaussian random variable if $P(|X| > x) \leq a \exp(-bx^2)$ for any $x > 0$ and constants $a, b > 0$. In addition, it has finite Orlicz ψ_2-norm, i.e., $\|X\|_{\psi_2} < \infty$. We say that X is a sub-Gaussian process with respect to a semi-metric d on its index set U if for any $t, s \in U$ and $x > 0$, we have

$$P(|X(t) - X(s)| > x) \leq 2 \exp\left(-\frac{1}{2} \frac{x^2}{d(t,s)^2}\right). \tag{1.12}$$

Further properties of sub-Gaussian random variables and processes can be found in Appendix C.

1.8 Chapter organization

1.8.1 Chapter 2: Sup-norm posterior convergence rate for univariate regression

To avoid the added technicalities in dealing with higher dimensions, we first consider one-dimensional version of (1.1). We represent f using univariate B-spline with normal coefficients. We study pointwise, L_2-and L_∞-posterior convergence rates for f and its derivatives. In addition, we construct pointwise, L_2-and L_∞-credible sets for f and its derivatives to quantify their estimation uncertainties. We compare the finite sample performance of our proposed pointwise credible intervals and credible bands with the pointwise confidence intervals and confidence bands proposed by Zhou et al. (1998).

1.8.2 Chapter 3: Sup-norm posterior convergence rate for multivariate regression

Multivariate extension of the previous chapter using random series based on tensor-product B-splines. As before, we derive pointwise, L_2-and L_∞-posterior convergence rates for f and all its mixed partial derivatives. We investigate the issue of constructing pointwise, L_2-and L_∞-credible sets for f and its mixed partial derivatives. Also, we consider anisotropic extension of our results for functions will different smoothness in different dimensions.

1.8.3 Chapter 4: Bayesian procedure for estimating location of maximum

We establish consistency for the posterior of $\boldsymbol{\mu}$. We derive an inequality relating sup-norm distance of partial derivatives of f with Euclidean distance of $\boldsymbol{\mu}$, and also prove a similar inequality between f and M. As an implication, posterior distributions of f and its partial derivatives will induce their corresponding posterior convergence rates on M and $\boldsymbol{\mu}$ respectively. In addition, we propose two different methods to construct credible regions for $\boldsymbol{\mu}$ and M.

1.8.4 Chapter 5: Two-stage Bayesian estimation of $\boldsymbol{\mu}$ and M

We propose a Bayesian two-stage procedure to estimate $\boldsymbol{\mu}$ and M. Asymptotic properties of our proposed method will be studied. In particular, we derive the corresponding two-stage posterior convergence rates of $\boldsymbol{\mu}$ and M. Performance of our proposed method will be evaluated against other single-stage methods, and also against the frequentist two-stage procedure proposed by Belitser et al. (2012).

1.8.5 Appendices

We present new results on univariate and tensor-product B-splines in Appendices A and B respectively. These results are used expensively to derive results on posterior convergence rates and credible sets in Chapters 1–4. In Appendix C, we list several important properties of sub-Gaussian random variables, and derive an upper bound for the L_p-moments of the sup-norm of a sub-Gaussian process with $p \geq 1$. Appendix D contains miscellaneous results on bounds for banded matrices and quadratic forms.

Chapter 2

Sup-norm posterior convergence rate for univariate regression

2.1 Introduction

Consider the univariate version of the nonparametric regression problem in (1.1), such that for some $f : U \subseteq \mathbb{R} \to \mathbb{R}$, we have

$$Y_i = f(X_i) + \varepsilon_i, \tag{2.1}$$

where ε_i are independent and identically distributed (i.i.d.) as $\mathrm{N}(0, \sigma^2)$ and $0 < \sigma < \infty$ is unknown. The covariates can be deterministic or randomly distributed. In both cases, $X_i \in U$ for $i = 1, \ldots, n$, where U is some open interval in $\mathbb{R}$. Since any such intervals is an affine transformation of $(0, 1)$, we can let $U = (0, 1)$ without loss of generality.

In this chapter, we will estimate f and its derivatives using Bayesian procedures, i.e., by expanding f using univariate B-splines and assigning normal priors on the basis coefficients. To deal with unknown σ^2, we consider two approaches: by estimating σ^2 using empirical Bayes, or further endowing σ^2 with a conjugate inverse-gamma prior. Conjugacy with the model (2.1) above enables explicit expression for the posterior distribution to be derived and analyzed. We will study pointwise, L_2-and L_∞-posterior convergence rates for f and its derivatives. Moreover, pointwise, L_2-and L_∞-credible sets will be constructed to quantify their estimation uncertainties. Under appropriate conditions, we show that they have guaranteed frequentist coverage with optimal size up to a logarithmic factor.

Since univariate B-splines are used extensively in this chapter, we give a brief introduction

to B-splines and discuss some of their important properties. Suppose we divide our domain $(0,1)$ into $N+1$ non-overlapping subintervals $[t_l, t_{l+1})$ for $l = 0, \ldots, N-1$ and $[t_N, t_{N+1}]$ with knot points $0 = t_0 < t_1 < \cdots < t_N < t_{N+1} = 1$. A B-spline or Basis-spline of order q is a piecewise polynomial defined on $(0,1)$, such that when restricted to a subinterval, B-spline is a polynomial with degree no more than $q-1$. It is differentiable up to $q-2$ times at the knot points, and is a continuous function globally if $q \geq 2$ (B-splines are step-functions if $q = 1$). For a given order q, there will be a set of $J = q + N$ B-splines, which we denote as $B_{j,q}(\cdot)$ for $j = 1, \ldots, J$. Exact expression of each $B_{j,q}(\cdot)$ is given by (see Section 2 of Zhou et al., 1998)

$$B_{j,q}(x) = (t_j - t_{j-q})[t_{j-q}, \ldots, t_j](t-x)_+^{q-1}, \tag{2.2}$$

where $[t_{j-q}, \ldots, t_j]h$ is the qth order divided difference of function h and $z_+ = z$ if $z > 0$ and 0 otherwise. In the above formula, we use $t_j = 0$ for $j < 0$ and $t_j = 1$ for $j > N+1$. We note that the expression above differs slightly with (4.16) of Schumaker (2007), where $B_{j,q}(\cdot)$ is constructed using knots from t_j to t_{j+q}. Alternatively, $B_{j,q}(\cdot)$ can be derived recursively from equation (14) of Chapter IX, de Boor (2001). This set of B-splines spans the entire J-dimensional space of polynomial splines. Figure 2.1 shows a set of 10 cubic ($q = 4$) B-splines.

Below are important properties of univariate B-splines used in this dissertation. Proofs of these results can be found in Chapter IX of de Boor (2001) or Section 4.3 of Schumaker (2007). New results on univariate B-splines that we developed can be found in Appendix A.

1. $0 \leq B_{j,q}(x) \leq 1$ for all $x \in (0,1)$ and $1 \leq j \leq J$.
2. The support of $B_{j,q}(x)$ is (t_{j-q}, t_j), i.e., $B_{j,q}(x) > 0$ on $x \in (t_{j-q}, t_j)$ and is zero for $x \notin [t_{j-q}, t_j]$.
3. For a given $x \in [t_{l-1}, t_l]$, only q adjacent B-splines $(B_{l,q}(x), \ldots, B_{l+q-1,q}(x))^T$ are nonzero (positive).
4. Partition of unity: $\sum_{j=1}^{J} B_{j,q}(x) = 1$ for any $x \in (0,1)$.

Let f_0 be the true regression function of f in (2.1). The assumption on f_0 in this chapter is as follows:

Assumption 1. Under the true distribution P_0, we assume $Y_i = f_0(X_i) + \varepsilon_i$ such that ε_i are i.i.d. with mean 0, variance σ_0^2 and $\|\varepsilon_i\|_{\psi_2} < \infty$ for $i = 1, \ldots, n$. Also, f_0 belongs to the isotropic Hölder space $\mathcal{H}^\alpha(0,1)$ for $\alpha > 2$. To avoid boundary effects, f_0 can be extended to a slightly bigger set $(-\epsilon, 1+\epsilon)$ for some $\epsilon > 0$.

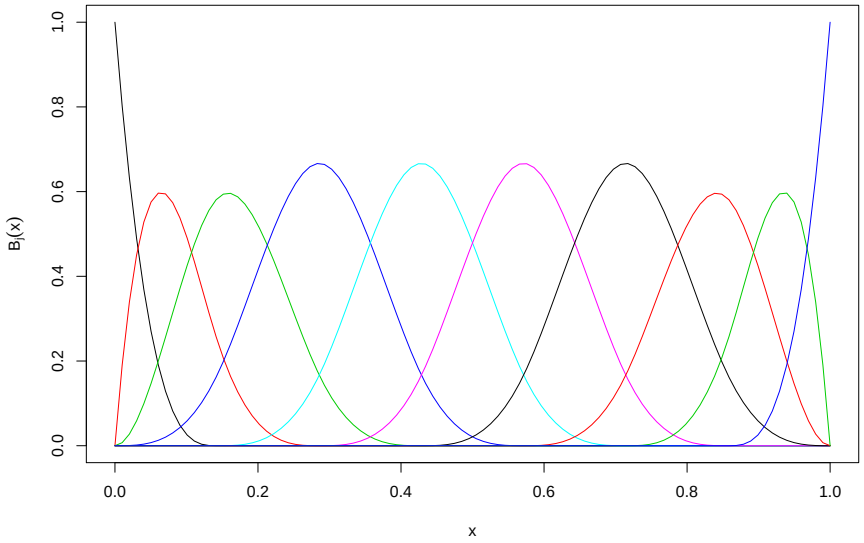


Figure 2.1: A set of 10 cubic ($q = 4$) B-splines.

Let p_0 be the density of P_0 with respect to the Lebesgue measure. Let $\mathrm{E}_0(\cdot)$ and $\mathrm{Var}_0(\cdot)$ be the expectation and variance operators taken with respect to P_0. For notational conciseness, write $\boldsymbol{F}_0 = (f_0(X_1), \ldots, f_0(X_n))^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$. We first describe in detail our prior specifications and assumptions, and then derive the corresponding posterior distributions of f and its derivatives.

2.2 Prior and posterior conjugacy

For any $x \in (0, 1)$, we represent $f(x)$ as

$$f(x) = \sum_{j=1}^{J} \theta_j B_{j,q}(x) = \boldsymbol{b}_{J,q}(x)^T \boldsymbol{\theta}, \tag{2.3}$$

where $\boldsymbol{b}_{J,q}(x) = (B_{1,q}(x), \ldots, B_{J,q}(x))^T$ is the set of J B-spline basis functions of fixed order $q \geq \alpha$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^T$ are the corresponding basis coefficients. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ and $\boldsymbol{B} = (\boldsymbol{b}_{J,q}(X_1), \boldsymbol{b}_{J,q}(X_2), \ldots, \boldsymbol{b}_{J,q}(X_n))^T$. Then the model for (2.1) can be written compactly as

$$\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{\theta}, \sigma^2 \sim \mathrm{N}_n(\boldsymbol{B}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}_n). \tag{2.4}$$

In this chapter, we treat J as deterministic and allow it to depend on n and α. Each B-spline

function has knot sequence $0 = t_{-q} = t_{1-q} = \cdots = t_0 < t_1 < t_2 < \cdots < t_N < t_{N+1} = t_{N+2} = \cdots = t_{N+q} = 1$. The q duplicate knots at each boundary point $\{0, 1\}$ are needed, because the recursive formulas for B-spline and its derivatives (see equations (4.22) and (4.23) of Schumaker, 2007) entail using knot points with index exceeding N or have negative index. Although the notation does not reflect, all knot points $\{t_{-q}, \ldots, t_0, t_1, \ldots, t_{N+1}, \ldots, t_{N+q-1}\}$ and the number of interior knots N may depend on n. We then have $J = q + N$, where N and thus J are increasing with n. Moreover, we assume that $J \leq n$.

Define $\delta_l = t_l - t_{l-1}$ for $l = 1, \ldots, N$ to be the one-step knot increment, and let $\Delta = \max_{1 \leq l \leq N} \delta_l$ be the mesh size. We assume that the knot sequence is quasi-uniform (see Definition 6.4 of Schumaker, 2007), that is,

$$\frac{\Delta}{\min_{1 \leq l \leq N} \delta_l} \leq C \tag{2.5}$$

for some $C > 0$. This assumption is reasonable in view of Lemma 6.17 from Schumaker (2007), which says that we can always choose a subset of knots from any given knot sequence to form a quasi-uniform sequence with $C = 3$. In addition, this class of partition is general enough for most applications and includes the uniform and nested uniform partitions as special cases (see Examples 6.6 and 6.7 of Schumaker, 2007 respectively).

If the design points $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ are deterministic, assume that there exists a cumulative distribution function $G(x)$ with positive and continuous density $g(x)$ on $[0, 1]$ such that

$$\sup_{x \in [0,1]} |G_n(x) - G(x)| = o(N^{-1}), \tag{2.6}$$

where $G_n(x) = n^{-1} \sum_{i=1}^n \mathbb{1}_{[X_i, \infty)}(x)$ is the empirical distribution of $(X_1 \ldots, X_n)^T$. As an example, the discrete uniform design $X_i = (i-1)/(n-1)$ for $i = 1, \ldots, n$ satisfies (2.6) with G being the uniform distribution on $[0, 1]$ and $N \lesssim n^{1/(2\alpha+1)}$.

For random design points, we assume that $X_i \overset{i.i.d.}{\sim} G(x)$, and they are mutually independent of ε_i for $i = 1, \ldots, n$. Here, $G(x) = P(X_i \leq x)$ is a cumulative distribution function with continuous density $g(x)$ on $[0, 1]$. By Donsker's theorem, $\|G_n - G\|_\infty = O_P(1/\sqrt{n})$. If $N \lesssim n^{1/(2\alpha+1)}$ and $\alpha > 1/2$, then

$$\|G_n - G\|_\infty = O_P\left(\frac{1}{\sqrt{n}}\right) = o_P(N^{-1}). \tag{2.7}$$

In this dissertation, we will prove results on posterior convergence rates and credible sets based on deterministic design points. The established theorems will then translate directly to the

random case by conditioning $\boldsymbol{X}$ in the posterior distribution, i.e., by considering $\Pi(\cdot|\boldsymbol{Y}, \boldsymbol{X})$.

On the basis coefficients, we assign

$$\boldsymbol{\theta}|\sigma^2 \sim \mathrm{N}_J(\boldsymbol{\eta}, \sigma^2\boldsymbol{\Omega}). \tag{2.8}$$

We assume that $\|\boldsymbol{\eta}\|_\infty < \infty$ and the entries of $\boldsymbol{\Omega}$ do not depend on n. Moreover, $\boldsymbol{\Omega}^{-1}$ is a v-banded matrix for some fixed v not depending on n. The bandedness of $\boldsymbol{\Omega}^{-1}$ ensures that the posterior precision matrix is also banded, which allows upper bound on the bias of the posterior mean to be established. We note that $\boldsymbol{\Omega}$ depends on n only through its dimension, which is $J \times J$. Furthermore, as $n \to \infty$, we assume that there exists constants $0 < c_1 \leq c_2 < \infty$ not depending on n such that

$$c_1\boldsymbol{I}_J \leq \boldsymbol{\Omega} \leq c_2\boldsymbol{I}_J. \tag{2.9}$$

Therefore, the prior support of f in (2.3) is given by the closure of its reproducing kernel Hilbert space (van der Vaart and van Zanten 2008a), which is the J-dimensional space of polynomial splines spanned by elements of $\boldsymbol{b}_{J,q}(\cdot)$. Since $q \geq \alpha$, this implies that the prior support of f is a subspace of $\mathcal{H}^\alpha(0,1)$. Based on (2.4) and (2.8), the conditional posterior of $\boldsymbol{\theta}$ is

$$\Pi(\boldsymbol{\theta}|\boldsymbol{Y}, \sigma^2) \sim \mathrm{N}_J\left[\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\left(\boldsymbol{B}^T\boldsymbol{Y} + \boldsymbol{\Omega}^{-1}\boldsymbol{\eta}\right), \sigma^2\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\right]. \tag{2.10}$$

It then follows that the induced conditional posterior for f is $\Pi(f|\boldsymbol{Y}, \sigma^2) \sim \mathrm{GP}(\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}\boldsymbol{\eta}, \sigma^2\Sigma)$, where $\boldsymbol{A}$ and $\boldsymbol{c}$ are bounded linear operators mapping $\mathbb{R}^n$ and $\mathbb{R}^J$ respectively to $\mathcal{H}^\alpha(0,1)$, and Σ is the covariance function defined on $(0,1) \times (0,1)$ such that for any $x, y \in (0,1)$,

$$\boldsymbol{A}(x) = \boldsymbol{b}_{J,q}(x)^T\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\boldsymbol{B}^T, \tag{2.11}$$

$$\boldsymbol{c}(x) = \boldsymbol{b}_{J,q}(x)^T\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\boldsymbol{\Omega}^{-1}, \tag{2.12}$$

$$\Sigma(x,y) = \boldsymbol{b}_{J,q}(x)^T\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\boldsymbol{b}_{J,q}(y). \tag{2.13}$$

Note that since the posterior mean is an affine transformation of $\boldsymbol{Y}$, Assumption 1 with Properties 1 and 2 of Appendix C imply that $\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}\boldsymbol{\eta}$ is a sub-Gaussian process under P_0.

To deal with unknown σ^2, we will consider two approaches. The first is empirical Bayes estimation. Based on (2.4) and (2.8), it follows that $\boldsymbol{Y}|\sigma^2 \sim \mathrm{N}_n[\boldsymbol{B}\boldsymbol{\eta}, \sigma^2(\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^T + \boldsymbol{I}_n)]$. The

corresponding marginal log-likelihood function is

$$l(\sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\log\{\det(\boldsymbol{B\Omega B}^T + \boldsymbol{I}_n)\} \\ - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{B\eta})^T(\boldsymbol{B\Omega B}^T + \boldsymbol{I}_n)^{-1}(\boldsymbol{Y} - \boldsymbol{B\eta}),$$

with score function

$$\frac{dl(\sigma^2)}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{Y} - \boldsymbol{B\eta})^T(\boldsymbol{B\Omega B}^T + \boldsymbol{I}_n)^{-1}(\boldsymbol{Y} - \boldsymbol{B\eta}).$$

Setting the above to zero and solving for σ^2, we obtain

$$\widehat{\sigma}_n^2 = \frac{(\boldsymbol{Y} - \boldsymbol{B\eta})^T(\boldsymbol{B\Omega B}^T + \boldsymbol{I}_n)^{-1}(\boldsymbol{Y} - \boldsymbol{B\eta})}{n}. \tag{2.14}$$

Note that $P_0(\widehat{\sigma}_n^2 > 0) = 1$ since $P_0(\boldsymbol{Y} = \boldsymbol{B\eta}) = 0$. By the second order condition,

$$\left.\frac{d^2 l(\sigma^2)}{d(\sigma^2)^2}\right|_{\sigma^2=\widehat{\sigma}_n^2} = -\frac{n}{2\widehat{\sigma}_n^4} < 0.$$

Hence with P_0 almost surely, $\widehat{\sigma}_n^2$ is a local maximum. Since $\widehat{\sigma}_n^2$ is the unique critical point, we conclude that $\widehat{\sigma}_n^2$ is the maximum marginal likelihood estimator for σ^2. Empirical Bayes then entails substituting σ^2 for $\widehat{\sigma}_n^2$ in the conditional posterior of f, i.e.,

$$\Pi(f|\boldsymbol{Y}, \sigma^2)|_{\sigma^2=\widehat{\sigma}_n^2} = \Pi_{\widehat{\sigma}_n}(f|\boldsymbol{Y}) \sim \mathrm{GP}(\boldsymbol{AY} + \boldsymbol{c\eta}, \widehat{\sigma}_n^2\Sigma). \tag{2.15}$$

We then write $\mathrm{E}_{\widehat{\sigma}_n}(\cdot|\boldsymbol{Y})$ and $\mathrm{Var}_{\widehat{\sigma}_n}(\cdot|\boldsymbol{Y})$ for the expectation and variance operators taken with respect to this empirical posterior. For the second approach, we will consider full hierarchical Bayes by further endowing σ^2 with a conjugate inverse-gamma (IG) prior

$$\sigma^2 \sim \mathrm{IG}\left(\frac{\beta}{2}, \frac{\gamma}{2}\right) \tag{2.16}$$

for some hyperparameters $\beta > 4$ and $\gamma > 0$. Here, $\beta > 4$ ensures that the prior mean and variance of σ^2 exist. By direct calculations, the posterior of σ^2 is

$$\sigma^2|\boldsymbol{Y} \sim \mathrm{IG}\left(\frac{\beta + n}{2}, \frac{\gamma + n\widehat{\sigma}_n^2}{2}\right). \tag{2.17}$$

The (i, j)th element of $\boldsymbol{B}^T\boldsymbol{B}$ is $\sum_{m=1}^n B_i(X_m)B_j(X_m)$. If $X_m \in [t_{l-1}, t_l]$ for some $1 \le l \le N + 1$, only q adjacent basis functions $(B_{l,q}(X_m), \ldots, B_{l+q-1,q}(X_m))^T$ will be nonzero for each $m = 1, \ldots, n$. Hence when $|i - j| > q$, $B_i(X_m)B_j(X_m) = 0$ for all $m = 1, \ldots, n$, and this implies

that $\boldsymbol{B}^T\boldsymbol{B}$ is q-banded. Furthermore under assumptions (2.5) and (2.6), Lemma 6.1 of Zhou et al. (1998) implies that there exist constants $0 < C_1 \leq C_2 < \infty$ not depending on n such that as $n \to \infty$,

$$C_1 \frac{n}{J} \boldsymbol{I}_J \leq \boldsymbol{B}^T\boldsymbol{B} \leq C_2 \frac{n}{J} \boldsymbol{I}_J. \tag{2.18}$$

In particular, $\|\boldsymbol{B}^T\boldsymbol{B}\|_{(2,2)} \asymp n/J$. Combining the above with (2.9), it follows that

$$\left(C_1 \frac{n}{J} + \frac{1}{c_2}\right) \leq \lambda_{\min}\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right) \leq \lambda_{\max}\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right) \leq \left(C_2 \frac{n}{J} + \frac{1}{c_1}\right). \tag{2.19}$$

By Theorem 22 of de Boor (2001) Chapter XII, for $f_0 \in \mathcal{H}^\alpha(0,1)$, there exists a $\boldsymbol{\theta}_\infty \in \mathbb{R}^J$ such that for constant $C_{\alpha,q} > 0$ depending only on α and q with $R > 0$, we have

$$\sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \|\boldsymbol{b}_{J,q}(\cdot)^T\boldsymbol{\theta}_\infty - f_0\|_\infty \leq \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} C_{\alpha,q} J^{-\alpha} \|f_0\|_{\mathcal{H}^\alpha} \lesssim J^{-\alpha}. \tag{2.20}$$

Since $\|\boldsymbol{b}_{J,q}(\cdot)^T\boldsymbol{\theta}_\infty\|_\infty \leq \|f_0\|_{\mathcal{H}^\alpha} + C_{\alpha,q}J^{-\alpha}\|f_0\|_{\mathcal{H}^\alpha} \leq R + 1$ uniformly in $\|f_0\|_{\mathcal{H}^\alpha} \leq R$, we have

$$\sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \|\boldsymbol{\theta}_\infty\|_\infty \lesssim \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \|\boldsymbol{b}_{J,q}(\cdot)^T\boldsymbol{\theta}_\infty\|_\infty = O(1) \tag{2.21}$$

by Corollary 8 of de Boor (2001) Chapter XI. The next theorem shows that $\widehat{\sigma}_n^2$ derived in (2.14) is a $\sqrt{n}$-consistent estimator of σ_0^2, uniformly over $\|f_0\|_{\mathcal{H}^\alpha} \leq R$.

Theorem 2.1. *If $J \asymp n^{1/(2\alpha+1)}$ and $\alpha > 1/2$, we have for $R > 0$,*

$$\sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0(\widehat{\sigma}_n^2 - \sigma_0^2)^2 = O\left(\frac{1}{n}\right). \tag{2.22}$$

Proof. Let $\boldsymbol{U} = (\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^T + \boldsymbol{I}_n)^{-1}$. By equation (33) of page 355 in Searle (1982), the absolute bias $|\mathrm{E}_0(\widehat{\sigma}_n^2) - \sigma_0^2|$ is

$$\begin{aligned}
&\left|\frac{\sigma_0^2}{n}\mathrm{tr}(\boldsymbol{U}) - \sigma_0^2\right| + \frac{1}{n}(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\eta})^T\boldsymbol{U}(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\eta}) \\
&\lesssim \frac{1}{n}[\mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{U}) + (\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty)^T\boldsymbol{U}(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty) + (\boldsymbol{B}\boldsymbol{\theta}_\infty - \boldsymbol{B}\boldsymbol{\eta})^T\boldsymbol{U}(\boldsymbol{B}\boldsymbol{\theta}_\infty - \boldsymbol{B}\boldsymbol{\eta})],
\end{aligned} \tag{2.23}$$

where we used $(\boldsymbol{x} + \boldsymbol{y})^T\boldsymbol{D}(\boldsymbol{x} + \boldsymbol{y}) \leq 2\boldsymbol{x}^T\boldsymbol{D}\boldsymbol{x} + 2\boldsymbol{y}^T\boldsymbol{D}\boldsymbol{y}$ for any $\boldsymbol{D} \geq \boldsymbol{0}$ in the last line. Let $\boldsymbol{P}_{\boldsymbol{B}} = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T$. Suppose that $\boldsymbol{A}$ is an $m \times m$ matrix, $\boldsymbol{G}$ an $m \times r$ matrix, $\boldsymbol{T}$ an $r \times r$ matrix, and $\boldsymbol{W}$ an $r \times m$ matrix. Assume that $\boldsymbol{A}$ and $\boldsymbol{T}$ are invertible. Then the binomial inverse theorem (see Theorem 18.2.8 of Harville, 1997 for a proof) says

$$(\boldsymbol{A} + \boldsymbol{G}\boldsymbol{T}\boldsymbol{W})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{G}(\boldsymbol{T}^{-1} + \boldsymbol{W}\boldsymbol{A}^{-1}\boldsymbol{G})^{-1}\boldsymbol{W}\boldsymbol{A}^{-1}. \tag{2.24}$$

Therefore, two applications of (2.24) to $\boldsymbol{U}$ yield

$$\boldsymbol{U} = (\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^T + \boldsymbol{I}_n)^{-1} = \boldsymbol{I}_n - \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1})^{-1}\boldsymbol{B}^T = \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{B}} + \boldsymbol{V}, \tag{2.25}$$

where $\boldsymbol{V} = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}[\boldsymbol{\Omega} + (\boldsymbol{B}^T\boldsymbol{B})^{-1}]^{-1}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T \geq \boldsymbol{0}$. Hence the first term in (2.23) is

$$\frac{1}{n}\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{B}} - \boldsymbol{V}) \leq \frac{1}{n}\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{B}}) = \frac{J}{n}. \tag{2.26}$$

Note that $\boldsymbol{U} \leq \boldsymbol{I}_n$ since $\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^T \geq \boldsymbol{0}$, and the second term in (2.23) is bounded by

$$\frac{1}{n}\|\boldsymbol{U}\|_{(2,2)}\|\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty\|^2 \leq \|\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty\|_\infty^2 \lesssim J^{-2\alpha} \tag{2.27}$$

in view of (2.20). By (2.25) and the fact that $(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{B}})\boldsymbol{B} = \boldsymbol{0}$, the last term in (2.23) is

$$\frac{1}{n}(\boldsymbol{\theta}_\infty - \boldsymbol{\eta})^T[\boldsymbol{\Omega} + (\boldsymbol{B}^T\boldsymbol{B})^{-1}]^{-1}(\boldsymbol{\theta}_\infty - \boldsymbol{\eta}) \leq \frac{1}{n}\left(c_1 + \frac{1}{C_2}\frac{J}{n}\right)^{-1} J\|\boldsymbol{\theta}_\infty - \boldsymbol{\eta}\|_\infty^2 \lesssim \frac{J}{n}, \tag{2.28}$$

where we used (2.9) and (2.18) to bound the maximum eigenvalue of $[\boldsymbol{\Omega} + (\boldsymbol{B}^T\boldsymbol{B})^{-1}]^{-1}$. By (2.21) and assumption on the prior, $\|\boldsymbol{\theta}_\infty - \boldsymbol{\eta}\|_\infty^2 = O(1)$. Combining the bounds in (2.26), (2.27) and (2.28) into (2.23), we obtain $|\mathrm{E}_0(\widehat{\sigma}_n^2) - \sigma_0^2| \lesssim J/n + J^{-2\alpha}$.

Let $\boldsymbol{Y} = \boldsymbol{F}_0 + \boldsymbol{\varepsilon}$ and write $n\widehat{\sigma}_n^2 = (\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\eta})^T\boldsymbol{U}(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\eta}) + 2(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\eta})^T\boldsymbol{U}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T\boldsymbol{U}\boldsymbol{\varepsilon}$. Using the fact $\mathrm{Var}(T_1 + T_2) \leq 2\mathrm{Var}(T_1) + 2\mathrm{Var}(T_2)$, it follows that $\mathrm{Var}_0(\widehat{\sigma}_n^2)$ is bounded up to a constant multiple by

$$\frac{1}{n^2}[(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty)^T\boldsymbol{U}^2(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty) + (\boldsymbol{B}\boldsymbol{\theta}_\infty - \boldsymbol{B}\boldsymbol{\eta})^T\boldsymbol{U}^2(\boldsymbol{B}\boldsymbol{\theta}_\infty - \boldsymbol{B}\boldsymbol{\eta}) + \mathrm{Var}_0(\boldsymbol{\varepsilon}^T\boldsymbol{U}\boldsymbol{\varepsilon})]. \tag{2.29}$$

In view of (2.20) and $\boldsymbol{U} \leq \boldsymbol{I}_n$, the first term above is bounded by

$$\frac{1}{n^2}\|\boldsymbol{U}\|_{(2,2)}^2\|\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty\|^2 \leq \frac{1}{n}\|\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty\|_\infty^2 \lesssim \frac{1}{n}J^{-2\alpha}. \tag{2.30}$$

By repeated applications of (2.9) and (2.18),

$$\begin{aligned}\|\boldsymbol{V}\|_{(2,2)} &= \|\boldsymbol{B}^T\boldsymbol{B}\|_{(2,2)}\|(\boldsymbol{B}^T\boldsymbol{B})^{-1}\|_{(2,2)}\|[\boldsymbol{\Omega} + (\boldsymbol{B}^T\boldsymbol{B})^{-1}]^{-1}\|_{(2,2)} \\ &\lesssim \left(\frac{n}{J}\right)\left(\frac{J}{n}\right)^2\left(c_1 + \frac{1}{C_2}\frac{J}{n}\right)^{-1} \lesssim \frac{J}{n}.\end{aligned} \tag{2.31}$$

Using (2.25), idempotency of $\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{B}}$ and $(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{B}})\boldsymbol{B} = \boldsymbol{0}$, the second term in (2.29) can be

written as

$$\frac{1}{n^2}(\boldsymbol{\theta}_\infty - \boldsymbol{\eta})^T \boldsymbol{B}^T (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{B}} + \boldsymbol{V})^2 \boldsymbol{B}(\boldsymbol{\theta}_\infty - \boldsymbol{\eta}) = \frac{1}{n^2}(\boldsymbol{\theta}_\infty - \boldsymbol{\eta})^T \boldsymbol{B}^T \boldsymbol{V}^2 \boldsymbol{B}(\boldsymbol{\theta}_\infty - \boldsymbol{\eta}).$$

In view of (2.31) and (2.18), the right hand side above is bounded by

$$\frac{1}{n^2}\|\boldsymbol{V}\|_{(2,2)}^2 \|\boldsymbol{B}(\boldsymbol{\theta}_\infty - \boldsymbol{\eta})\|^2 \lesssim \frac{1}{n^2}\left(\frac{J}{n}\right)^2 \|\boldsymbol{B}^T\boldsymbol{B}\|_{(2,2)} J \|\boldsymbol{\theta}_\infty - \boldsymbol{\eta}\|_\infty^2 \lesssim \frac{1}{n}\left(\frac{J}{n}\right)^2, \tag{2.32}$$

where $\|\boldsymbol{\theta}_\infty - \boldsymbol{\eta}\|_\infty^2$ is bounded using (2.21) and assumption on the prior. By Lemma D.3, the last term in (2.29) is $O(1/n)$. Combining this with the bounds established in (2.30) and (2.32) into (2.29), we obtain $\mathrm{Var}_0(\widehat{\sigma}_n^2) \lesssim 1/n$. By taking $J \asymp n^{1/(2\alpha+1)} \ll n$ and $\alpha > 1/2$, the mean square error is

$$\mathrm{E}_0(\widehat{\sigma}_n^2 - \sigma_0^2)^2 \lesssim \frac{1}{n} + \left(\frac{J}{n}\right)^2 + J^{-4\alpha} \lesssim \frac{1}{n}. \quad \square$$

A Bayesian analog of Theorem 2.1 above is given below, which shows that the posterior of σ^2 in (2.17) converges to σ_0^2 at the parametric rate $1/\sqrt{n}$.

Theorem 2.2. *If $J \asymp n^{1/(2\alpha+1)}$ and $\alpha > 1/2$, we have for $R > 0$ and any $M_n \to \infty$,*

$$\sup_{\|f_0\|_{\mathcal{H}^\alpha} \le R} \mathrm{E}_0 \Pi\left(|\sigma^2 - \sigma_0^2| > \frac{M_n}{\sqrt{n}} \middle| \boldsymbol{Y}\right) \to 0. \tag{2.33}$$

Proof. By Markov's inequality, it suffices to show that $\mathrm{E}_0\mathrm{E}[(\sigma^2 - \sigma_0^2)^2|\boldsymbol{Y}] = O(n^{-1})$. The posterior mean and variance of σ^2 are

$$\begin{aligned} \mathrm{E}(\sigma^2|\boldsymbol{Y}) &= \frac{\gamma}{\beta+n-2} + \frac{n}{\beta+n-2}\widehat{\sigma}_n^2, \\ \mathrm{Var}(\sigma^2|\boldsymbol{Y}) &= \frac{2}{\beta+n-4}\left(\frac{\gamma}{\beta+n-2} + \frac{n}{\beta+n-2}\widehat{\sigma}_n^2\right)^2. \end{aligned} \tag{2.34}$$

We have $\mathrm{E}[(\sigma^2 - \sigma_0^2)^2|\boldsymbol{Y}] = [\mathrm{E}(\sigma^2|\boldsymbol{Y}) - \sigma_0^2]^2 + \mathrm{Var}(\sigma^2|\boldsymbol{Y})$. Theorem 2.1 implies that $\mathrm{E}_0(\widehat{\sigma}_n^2) = \sigma_0^2 + O(1/\sqrt{n})$, $\mathrm{Var}_0(\widehat{\sigma}_n^2) = O(1/n)$ and $\mathrm{E}_0(\widehat{\sigma}_n^2)^2 = \sigma_0^4 + O(1/n)$. Hence, $\mathrm{E}_0[\mathrm{E}(\sigma^2|\boldsymbol{Y}) - \sigma_0^2)]^2$ is

$$\begin{aligned} &[\mathrm{E}_0\mathrm{E}(\sigma^2|\boldsymbol{Y}) - \sigma_0^2]^2 + \mathrm{Var}_0[\mathrm{E}(\sigma^2|\boldsymbol{Y})] \\ &= \left\{O\left(\frac{1}{n}\right) + \left[1 + O\left(\frac{1}{n}\right)\right]\left[\sigma_0^2 + O\left(\frac{1}{\sqrt{n}}\right)\right] - \sigma_0^2\right\}^2 + \left[1 + O\left(\frac{1}{n}\right)\right]^2 O\left(\frac{1}{n}\right) \\ &= O\left(\frac{1}{n}\right), \end{aligned}$$

while $\mathrm{E}_0\mathrm{Var}(\sigma^2|\boldsymbol{Y})$ is

$$O\left(\frac{1}{n}\right)\left\{O\left(\frac{1}{n^2}\right)+\left[1+O\left(\frac{1}{n}\right)\right]^2\left[\sigma_0^4+O\left(\frac{1}{n}\right)\right]\right\}=O\left(\frac{1}{n}\right). \quad \square$$

2.3 Posterior convergence rates for f

Since the explicit expression for the posterior of f given $(\boldsymbol{Y},\sigma^2)$ is available due to normal-normal conjugacy, we will derive posterior convergence rates by directly analyzing the posterior distribution. This is done by decomposing the posterior mean square error into three parts corresponding to: posterior variance, variance and bias of the posterior mean. We then bound these three terms separately and use Markov's inequality to arrive at our desired results. Theorem 2.3 below is for pointwise posterior convergence and is a crucial en-route step in deriving Theorems 2.4 and 2.5. For notational simplicity, we write $\sigma^2 \in \mathcal{U}_n$ if $|\sigma^2-\sigma_0^2| \le \zeta_n/\sqrt{n}$ for $\zeta_n = o(\sqrt{n})$.

Theorem 2.3. *If $J \asymp n^{1/(2\alpha+1)}$, we have for $R>0$, every fixed $x \in (0,1)$ and any $M_n \to \infty$,*

$$\textit{Empirical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R} \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(f: |f(x)-f_0(x)| > M_n n^{-\alpha/(2\alpha+1)}|\boldsymbol{Y}) \to 0. \tag{2.35}$$

$$\textit{Hierarchical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R} \mathrm{E}_0 \Pi(f: |f(x)-f_0(x)| > M_n n^{-\alpha/(2\alpha+1)}|\boldsymbol{Y}) \to 0. \tag{2.36}$$

Proof. At any $x \in (0,1)$, $\Pi(f(x)|\boldsymbol{Y},\sigma^2) \sim \mathrm{N}(\boldsymbol{A}(x)\boldsymbol{Y}+\boldsymbol{c}(x)\boldsymbol{\eta}, \sigma^2\Sigma(x,x))$ where $\boldsymbol{A}(x), \boldsymbol{c}(x)$ and $\Sigma(x,x)$ were given by (2.11), (2.12) and (2.13) respectively. Under P_0, $\boldsymbol{A}(x)\boldsymbol{Y}+\boldsymbol{c}(x)\boldsymbol{\eta}$ is a sub-Gaussian random variable with mean $\boldsymbol{A}(x)\boldsymbol{F}_0+\boldsymbol{c}(x)\boldsymbol{\eta}$ and variance $\sigma_0^2\Psi(x,x)$, where $\sigma_0^2\Psi$ is the covariance function of the sub-Gaussian process $\boldsymbol{AY}+\boldsymbol{c\eta}$ under P_0, such that for $x,y \in (0,1)$,

$$\Psi(x,y) = \boldsymbol{b}_{J,q}(x)^T\left(\boldsymbol{B}^T\boldsymbol{B}+\boldsymbol{\Omega}^{-1}\right)^{-1}\boldsymbol{B}^T\boldsymbol{B}\left(\boldsymbol{B}^T\boldsymbol{B}+\boldsymbol{\Omega}^{-1}\right)^{-1}\boldsymbol{b}_{J,q}(y). \tag{2.37}$$

Note that the posterior variance $\sigma^2\Sigma(x,x)$ does not depend on $\boldsymbol{Y}$ and f_0, while the posterior mean $\boldsymbol{A}(x)\boldsymbol{Y}+\boldsymbol{c}(x)\boldsymbol{\eta}$ does not depend on σ^2. Therefore,

$$\begin{aligned}
&\sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R} \mathrm{E}_0 \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}\{[f(x)-f_0(x)]^2|\boldsymbol{Y},\sigma^2\} \\
&\le \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}\{[f(x)-\boldsymbol{A}(x)\boldsymbol{Y}-\boldsymbol{c}(x)\boldsymbol{\eta}]^2|\sigma^2\} + \sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R} \mathrm{E}_0\mathrm{E}\{[\boldsymbol{A}(x)\boldsymbol{Y}+\boldsymbol{c}(x)\boldsymbol{\eta}-f_0(x)]^2|\boldsymbol{Y}\} \\
&= \sup_{\sigma^2\in\mathcal{U}_n} \sigma^2\Sigma(x,x)+\sigma_0^2\Psi(x,x) + \sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R} [\boldsymbol{A}(x)\boldsymbol{F}_0+\boldsymbol{c}(x)\boldsymbol{\eta}-f_0(x)]^2.
\end{aligned} \tag{2.38}$$

To bound the posterior variance $\sigma^2\Sigma(x,x)$ uniformly over $\sigma^2 \in \mathcal{U}_n$, first observe that

$$\|\boldsymbol{b}_{J,q}(x)\|^2 = \sum_{j=1}^{J} B_{j,q}^2(x) \leq \max_{1\leq j\leq J} B_{j,q}(x) \sum_{j=1}^{J} B_{j,q}(x) \leq 1. \tag{2.39}$$

Cauchy-Schwarz inequality and the definition of spectral norm then imply

$$\begin{aligned}\sup_{\sigma^2\in\mathcal{U}_n} \sigma^2\Sigma(x,x) &\leq \left(\sigma_0^2 + \frac{\zeta_n}{\sqrt{n}}\right) \left\|\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\right\|_{(2,2)} \\ &\lesssim \left(C_1\frac{n}{J} + \frac{1}{c_2}\right)^{-1} = O\left(\frac{J}{n}\right), \end{aligned} \tag{2.40}$$

where the last line follows from (2.19). In view of (2.37), the variance of the posterior mean, i.e., $\sigma_0^2\Psi(x,x)$ can be bounded above by

$$\begin{aligned}\sigma_0^2\Psi(x,x) &\leq \sigma_0^2 \left\|\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\right\|_{(2,2)}^2 \|\boldsymbol{B}^T\boldsymbol{B}\|_{(2,2)} \|\boldsymbol{b}_{J,q}(x)\|^2 \\ &= O\left(\frac{J}{n}\right)^2 O\left(\frac{n}{J}\right) = O\left(\frac{J}{n}\right), \end{aligned} \tag{2.41}$$

where we have used (2.19), (2.18) and (2.39) to bound the three norms respectively. To bound the last squared bias term in (2.38), we will appeal to (2.20) and the fact that $\|\boldsymbol{b}_{J,q}(x)\|_1 = \sum_{j=1}^{J} B_{j,q}(x) = 1$. Hölder's inequality then implies uniformly in $\|f_0\|_{\mathcal{H}^\alpha} \leq R$ that

$$\begin{aligned}&|\boldsymbol{A}(x)\boldsymbol{F}_0 + \boldsymbol{c}(x)\boldsymbol{\eta} - f_0(x)| \\ &= \left|\boldsymbol{b}_{J,q}(x)^T \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \left(\boldsymbol{B}^T\boldsymbol{F}_0 + \boldsymbol{\Omega}^{-1}\boldsymbol{\eta}\right) - f_0(x)\right| \\ &\leq \left|\boldsymbol{b}_{J,q}(x)^T \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \left(\boldsymbol{B}^T\boldsymbol{F}_0 + \boldsymbol{\Omega}^{-1}\boldsymbol{\eta}\right) - \boldsymbol{b}_{J,q}(x)^T\boldsymbol{\theta}_\infty\right| + |f_0(x) - \boldsymbol{b}_{J,q}(x)^T\boldsymbol{\theta}_\infty| \\ &\leq \left|\boldsymbol{b}_{J,q}(x)^T \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \left[\boldsymbol{B}^T(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty) + \boldsymbol{\Omega}^{-1}(\boldsymbol{\eta} - \boldsymbol{\theta}_\infty)\right]\right| + O(J^{-\alpha}) \\ &\leq \left\|\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\right\|_{(\infty,\infty)} \left\{\|\boldsymbol{B}^T(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty)\|_\infty + \left\|\boldsymbol{\Omega}^{-1}\right\|_{(\infty,\infty)} (\|\boldsymbol{\theta}_\infty\|_\infty + \|\boldsymbol{\eta}\|_\infty)\right\} \\ &\quad + O(J^{-\alpha}). \end{aligned} \tag{2.42}$$

By assumption, $\boldsymbol{\Omega}^{-1}$ is v-banded with fixed v and uniformly bounded entries. Therefore, $\|\boldsymbol{\Omega}^{-1}\|_{(\infty,\infty)} = O(1)$. Since $\boldsymbol{B}^T\boldsymbol{B}$ is q-banded, Lemma D.1 and equation (2.19) imply that $\|(\boldsymbol{B}^T\boldsymbol{B}+\boldsymbol{\Omega}^{-1})^{-1}\|_{(\infty,\infty)} = O(J/n)$. Also, $\sup_{\|f_0\|_{\mathcal{H}^\alpha}\leq R} \|\boldsymbol{\theta}\|_\infty$ and $\|\boldsymbol{\eta}\|_\infty$ are both $O(1)$ by (2.21) and the assumption on the prior. To bound $\sup_{\|f_0\|_{\mathcal{H}^\alpha}\leq R} \|\boldsymbol{B}^T(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty)\|_\infty$, we use the non-

negativity of B-splines, (2.20) and Lemma A.2 to write

$$\begin{aligned}\sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R}\|\boldsymbol{B}^T(\boldsymbol{F}_0-\boldsymbol{B}\boldsymbol{\theta}_\infty)\|_\infty &\le \sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R}\max_{1\le j\le J}\sum_{i=1}^n B_{j,q}(X_i)|f_0(X_i)-\boldsymbol{b}_{J,q}(X_i)^T\boldsymbol{\theta}_\infty| \\ &\le \sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R} C_{\alpha,q}J^{-\alpha}\|f_0\|_{\mathcal{H}^\alpha}\max_{1\le j\le J}\sum_{i=1}^n B_{j,q}(X_i) \\ &\lesssim nJ^{-(1+\alpha)}. \end{aligned} \tag{2.43}$$

Therefore, combining the bounds obtained and squaring (2.42), we have for any $x\in(0,1)$,

$$\sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R}[\boldsymbol{A}(x)\boldsymbol{F}_0+\boldsymbol{c}(x)\boldsymbol{\eta}-f_0(x)]^2 = O\left(\frac{J^2}{n^2}\right)+O(J^{-2\alpha}). \tag{2.44}$$

Combining all bounds from (2.40), (2.41) and (2.44) into (2.38), we obtain

$$\begin{aligned}\sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R}\mathrm{E}_0\sup_{\sigma^2\in\mathcal{U}_n}\mathrm{E}\left[(f(x)-f_0(x))^2|\boldsymbol{Y},\sigma^2\right] &= O\left(\frac{J}{n}\right)+O\left(\frac{J^2}{n^2}\right)+O(J^{-2\alpha}) \\ &= O\left(\frac{J}{n}\right)+O(J^{-2\alpha}), \end{aligned} \tag{2.45}$$

where the last line follows since $J\le n$ by assumption. The two terms on the right will have the same order when $J\asymp n^{1/(2\alpha+1)}$. This implies that the corresponding posterior mean square error in (2.45) is of order $n^{-2\alpha/(2\alpha+1)}$.

Observe that by Markov's inequality, Theorem 2.1 implies for $R>0$ and any $M_n\to\infty$,

$$\inf_{\|f_0\|_{\mathcal{H}^\alpha}\le R}P_0\left(\sigma_0^2-\frac{M_n}{\sqrt{n}}\le\widehat{\sigma}_n^2\le\sigma_0^2+\frac{M_n}{\sqrt{n}}\right)\to 1. \tag{2.46}$$

For the empirical Bayes version, by Markov's inequality, (2.45) and (2.46) above, we have for $\epsilon_n=n^{-\alpha/(2\alpha+1)}$ and any $M_n\to\infty$,

$$\mathrm{E}_0\Pi_{\widehat{\sigma}_n}(|f(x)-f_0(x)|>M_n\epsilon_n|\boldsymbol{Y}) \le \frac{\mathrm{E}_0\sup_{\sigma^2\in\mathcal{U}_n}\mathrm{E}([f(x)-f_0(x)]^2|\boldsymbol{Y},\sigma^2)}{M_n^2\epsilon_n^2}+o(1)\to 0, \tag{2.47}$$

uniformly over $\|f_0\|_{\mathcal{H}^\alpha}\le R$. For the hierarchical Bayes procedure, we have for $\epsilon_n=n^{-\alpha/(2\alpha+1)}$

and any $M_n \to \infty$,

$$\begin{aligned}
&\mathrm{E}_0 \Pi(|f(x) - f_0(x)| > M_n \epsilon_n | \boldsymbol{Y}) \\
&\leq \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \Pi(|f(x) - f_0(x)| > M_n \epsilon_n | \boldsymbol{Y}, \sigma^2) + \mathrm{E}_0 \Pi\left(|\sigma^2 - \sigma_0^2| > \frac{\zeta_n}{\sqrt{n}} \middle| \boldsymbol{Y} \right), \qquad (2.48)
\end{aligned}$$

uniformly over $\|f_0\|_{\mathcal{H}^\alpha} \leq R$. The first term on the right hand side is $o(1)$ by (2.45), while the second term goes to zero by Theorem 2.2. □

Theorem 2.4. *If $J \asymp n^{1/(2\alpha+1)}$, we have for $R > 0$ and every $M_n \to \infty$,*

$$\textit{Empirical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(f : \|f - f_0\|_2 > M_n n^{-\alpha/(2\alpha+1)} | \boldsymbol{Y}) \to 0. \qquad (2.49)$$

$$\textit{Hierarchical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \Pi(f : \|f - f_0\|_2 > M_n n^{-\alpha/(2\alpha+1)} | \boldsymbol{Y}) \to 0. \qquad (2.50)$$

Proof. By Fubini's theorem,

$$\begin{aligned}
\sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}(\|f - f_0\|_2^2 | \boldsymbol{Y}, \sigma^2) &\leq \int_0^1 \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}(|f(x) - f_0(x)|^2 | \boldsymbol{Y}, \sigma^2) dx \\
&\lesssim \frac{J}{n} + J^{-2\alpha}, \qquad (2.51)
\end{aligned}$$

where the last line follows from (2.45). Setting $J \asymp n^{1/(2\alpha+1)}$ will balance the orders of these two terms, yielding the rate of $n^{-2\alpha/(2\alpha+1)}$. The empirical and hierarchical Bayes posterior convergence rates then follow from (2.47) and (2.48) respectively with absolute values replaced by L_2-norms. □

Theorem 2.5. *If $J \asymp (n/\log n)^{1/(2\alpha+1)}$, we have for $R > 0$ and every $M_n \to \infty$,*

$$\textit{Empirical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(f : \|f - f_0\|_\infty > M_n (\log n/n)^{\alpha/(2\alpha+1)} | \boldsymbol{Y}) \to 0.$$

$$\textit{Hierarchical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \Pi(f : \|f - f_0\|_\infty > M_n (\log n/n)^{\alpha/(2\alpha+1)} | \boldsymbol{Y}) \to 0.$$

Proof. Recall that $\Pi(f | \boldsymbol{Y}, \sigma^2) \sim \mathrm{GP}(\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}\boldsymbol{\eta}, \sigma^2 \Sigma)$, with $\boldsymbol{A}, \boldsymbol{c}$ and Σ defined in (2.11), (2.12) and (2.13) respectively. Let $Z_n \sim \mathrm{GP}(0, \Sigma)$. Under the true distribution P_0, $\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}\boldsymbol{\eta}$ is a sub-Gaussian process with mean function $\boldsymbol{A}\boldsymbol{F}_0 + \boldsymbol{c}\boldsymbol{\eta}$ and covariance function $\sigma_0^2 \Psi$, where Ψ was defined in (2.37). Let Q_n be a sub-Gaussian process with mean function 0 and covariance function $\sigma_0^2 \Psi$. Note that Z_n does not depend on $\boldsymbol{Y}$ and f_0, while the posterior mean $\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}\boldsymbol{\eta}$

does not depend on σ^2. Therefore,

$$\begin{aligned}
&\sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R} \mathrm{E}_0 \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}(\|f-f_0\|_\infty^2|\boldsymbol{Y},\sigma^2)\\
&\le \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}(\|f-\boldsymbol{AY}-\boldsymbol{c\eta}\|_\infty^2|\sigma^2) + \sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R} \mathrm{E}_0\{\mathrm{E}(\|\boldsymbol{AY}+\boldsymbol{c\eta}-f_0\|_\infty^2|\boldsymbol{Y})\}\\
&= \sup_{\sigma^2\in\mathcal{U}_n} \sigma^2\mathrm{E}\|Z_n\|_\infty^2 + \mathrm{E}\|Q_n\|_\infty^2 + \sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R} \|\boldsymbol{AF}_0+\boldsymbol{c\eta}-f_0\|_\infty^2. \qquad (2.52)
\end{aligned}$$

Since $Q_n = \boldsymbol{A\varepsilon}$, then by Property 2 of Appendix C and Assumption 1, Q_n is sub-Gaussian with respect to the semi-metric $d(t,s) = \sqrt{\mathrm{Var}(Q_n(t)-Q_n(s))}$. By Lemma A.4 with $m=0$, Z_n and Q_n satisfy the condition for Lemma C.1. Applying Lemma C.1 with $p=2$, we see that for any $0<\delta_n<1$ and $T_n = N(\delta_n,(0,1),\|\cdot\|)$,

$$\begin{aligned}
\sup_{\sigma^2\in\mathcal{U}_n} \sigma^2\mathrm{E}\|Z_n\|_\infty^2 &\lesssim \sigma_0^2\log(1/\delta_n)\left\{n\delta_n^2 + \max_{1\le i\le T_n}\mathrm{E}[Z_n(u_i)^2]\right\}\\
&\lesssim \log(1/\delta_n)\left(n\delta_n^2+\frac{J}{n}\right), \qquad (2.53)
\end{aligned}$$

where the last line follows from (2.40). Similarly,

$$\mathrm{E}\|Q_n\|_\infty^2 \lesssim \log(1/\delta_n)\left\{n\delta_n^2 + \max_{1\le i\le T_n}\mathrm{E}[Q_n(u_i)^2]\right\} \lesssim \log(1/\delta_n)\left(n\delta_n^2+\frac{J}{n}\right), \qquad (2.54)$$

where the last inequality follows from (2.41). Setting $\delta_n \asymp \sqrt{J}/n$ for both (2.53) and (2.54),

$$\sup_{\sigma^2\in\mathcal{U}_n} \sigma^2\mathrm{E}\|Z_n\|_\infty^2 \lesssim \frac{J}{n}\log n, \qquad \mathrm{E}\|Q_n\|_\infty^2 \lesssim \frac{J}{n}\log n. \qquad (2.55)$$

Since the bound for (2.44) is uniform for all $x\in(0,1)$ and all $\|f_0\|_{\mathcal{H}^\alpha}\le R$, the last deterministic squared bias term in (2.52) is

$$\sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R} \|A\boldsymbol{F}_0+\boldsymbol{c\eta}-f_0\|_\infty^2 = O\left(\frac{J^2}{n^2}\right) + O(J^{-2\alpha}). \qquad (2.56)$$

Therefore, combining the bounds for (2.55) and (2.56) into (2.52), we will obtain

$$\begin{aligned}
\sup_{\|f_0\|_{\mathcal{H}^\alpha}\le R} \mathrm{E}_0 \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}(\|f-f_0\|_\infty^2|\boldsymbol{Y},\sigma^2) &\lesssim \frac{J}{n}\log n + \frac{J^2}{n^2} + J^{-2\alpha}\\
&\lesssim \frac{J}{n}\log n + J^{-2\alpha}, \qquad (2.57)
\end{aligned}$$

where the last line follows since $J\le n$ by assumption. Setting $J \asymp (n/\log n)^{1/(2\alpha+1)}$ will balance

the orders of these two terms, yielding the rate $(\log n/n)^{2\alpha/(2\alpha+1)}$ for the posterior mean square error in (2.57). The empirical and hierarchical Bayes posterior convergence rates then follow from (2.47) and (2.48) respectively with absolute values replaced by sup-norms. □

Note that we will incur an extra logarithmic factor in the rate once we move from pointwise to sup-norm convergence. The pointwise, L_2-and L_∞-rates established above are the minimax rates (see Stone, 1980 for pointwise and Theorem 1 of Stone, 1982 for general L_r-norms, with $1 \leq r \leq \infty$). By examining the proofs of Theorems 2.3, 2.4 and 2.5 above, it also follows that the point estimator of $f_0(x)$, i.e., the posterior mean $\boldsymbol{A}(x)\boldsymbol{Y} + \boldsymbol{c}(x)\boldsymbol{\eta}$ converges to $f_0(x)$ at the same $n^{-\alpha/(2\alpha+1)}$ rate. Similarly, the posterior mean of f, i.e., $\boldsymbol{AY} + \boldsymbol{c\eta}$ converges to f_0 under L_2-norm at the rate $n^{-\alpha/(2\alpha+1)}$, while in L_∞-norm, it is $(\log n/n)^{\alpha/(2\alpha+1)}$.

2.4 Posterior convergence rates for derivatives

Given our prior specification $f(x) = \boldsymbol{b}_{J,q}(x)^T\boldsymbol{\theta}$ and $\boldsymbol{\theta}|\sigma^2 \sim \mathrm{N}(\boldsymbol{\eta}, \sigma^2\boldsymbol{\Omega})$, the mth derivative of f for $m \in \mathbb{N}_0$ and $m < \alpha$ can be written as $d^m f(x)/dx^m = \sum_{j=1}^{J} \theta_j d^m B_{j,q}(x)/dx^m$. Using equations (15) and (16) of Chapter X in de Boor (2001), this expression reduces to

$$f^{(m)}(x) = \frac{d^m f(x)}{dx^m} = \sum_{j=1}^{J-m} \theta_j^{(m)} B_{j,q-m}(x), \tag{2.58}$$

for

$$\theta_j^{(m)} = \begin{cases} \theta_j & \text{if } m = 0, \\ (q-m)\dfrac{\theta_{j+1}^{(m-1)} - \theta_j^{(m-1)}}{t_j - t_{j-q+m}} & \text{if } m \geq 1, \end{cases} \tag{2.59}$$

where $B_{j,q-m}(\cdot)$ is the corresponding $q-m$ order B-spline.

Define $\boldsymbol{b}_{J,q,m}(x) = (B_{1,q-m}(x), \ldots, B_{J-m,q-m}(x))^T$. Then $f^{(m)}(x) = \boldsymbol{b}_{J,q,m}(x)^T\boldsymbol{W}_m\boldsymbol{\theta}$, where $\boldsymbol{W}_m$ is a $(J-m) \times J$ matrix. Each row of $\boldsymbol{W}_m$ has only $m+1$ consecutive nonzero entries and their arrangement is analogues to a banded matrix with bandwidth $m+1$, namely the position of the $m+1$ consecutive entries in the current row is a shift of one entry to the right of the $m+1$ consecutive entries' position in the previous row. The first m columns will have 1 to m nonzero entries respectively and the last m columns will have m to 1 nonzero entries. The rest $J-2m$ columns will have exactly $m+1$ nonzero entries. All other entries are zero.

In view of (2.59), the consecutive $m+1$ entries in each row of $\boldsymbol{W}_m$ are actually the coefficients

associated with taking m iterated finite difference of elements in $\boldsymbol{\theta}$, and divided by products of difference quotients of the knot sequence. Let $w_{ij}^{(m)}$ denote the (i,j)th entry of $\boldsymbol{W}_m$. The expressions for the nonzero entries can be described as follows: for each row $i = 1, \ldots, J-m$, the endpoints of the $m+1$ consecutive entries are

$$w_{i,i}^{(m)} = (-1)^m \prod_{l=1}^{m} \frac{q-l}{t_i - t_{i-q+l}}, \qquad w_{i,i+m}^{(m)} = \prod_{l=1}^{m} \frac{q-l}{t_{i+l-1} - t_{i-q+m}}. \tag{2.60}$$

If m is odd, the intermediate entries for $1 \leq l \leq (m-1)/2$ are

$$w_{i,i+l}^{(m)} = (-1)^{m-l} w_{i,i}^{(m)} \left[1 + \sum_{r=1}^{\binom{m}{l}-1} \prod_{s=1}^{r} \frac{t_i - t_{i-q+s}}{t_{i+1} - t_{i+1-q+s}} \right], \tag{2.61}$$

while for $(m+1)/2 \leq l \leq m$, it is

$$w_{i,i+l}^{(m)} = (-1)^{m-l} w_{i,i+m}^{(m)} \left[1 + \sum_{r=1}^{\binom{m}{l}-1} \prod_{s=1}^{r} \frac{t_{i+m-s} - t_{i+m-q}}{t_{i+m-1-s} - t_{i+m-1-q}} \right]. \tag{2.62}$$

If m is even, then the expression for $w_{i,i+l}^{(m)}$ is (2.61) for $1 \leq l \leq m/2-1$ and is (2.62) for $m/2+1 \leq l \leq m$. When $l = m/2$, we have that $w_{i,i+m/2}^{(m)}$ is

$$\begin{aligned} & w_{i,i}^{(m)} \left(\frac{t_i - t_{i-q+1}}{t_{i+1} - t_{i+2-q}} \right) \left[1 + \sum_{r=1}^{\binom{m-1}{m/2-1}-1} \prod_{s=0}^{r-1} \frac{t_{i+s} - t_{i+2-q}}{t_{i+1+s} - t_{i+3-q}} \right] \\ & \quad + w_{i,i+m}^{(m)} \left(\frac{t_{i+m-1} - t_{i+m-q}}{t_{i+m-2} - t_{i+m-1-q}} \right) \left[1 + \sum_{r=1}^{\binom{m-1}{m/2}-1} \prod_{s=0}^{r-1} \frac{t_{i+m-2} - t_{i+m-q-s}}{t_{i+m-3} - t_{i+m-1-q-s}} \right]. \end{aligned} \tag{2.63}$$

The conditional posterior distribution for $f^{(m)}$ is $\Pi(f^{(m)} | \boldsymbol{Y}, \sigma^2) \sim \mathrm{GP}(\boldsymbol{A}_m \boldsymbol{Y} + \boldsymbol{c}_m \boldsymbol{\eta}, \sigma^2 \Sigma_m)$, where for $x, y \in (0,1)$,

$$\boldsymbol{A}_m(x) = \boldsymbol{b}_{J,q,m}(x)^T \boldsymbol{W}_m \left(\boldsymbol{B}^T \boldsymbol{B} + \boldsymbol{\Omega}^{-1} \right)^{-1} \boldsymbol{B}^T, \tag{2.64}$$

$$\boldsymbol{c}_m(x) = \boldsymbol{b}_{J,q,m}(x)^T \boldsymbol{W}_m \left(\boldsymbol{B}^T \boldsymbol{B} + \boldsymbol{\Omega}^{-1} \right)^{-1} \boldsymbol{\Omega}^{-1}, \tag{2.65}$$

$$\Sigma_m(x,y) = \boldsymbol{b}_{J,q,m}(x)^T \boldsymbol{W}_m \left(\boldsymbol{B}^T \boldsymbol{B} + \boldsymbol{\Omega}^{-1} \right)^{-1} \boldsymbol{W}_m^T \boldsymbol{b}_{J,q,m}(y). \tag{2.66}$$

Note that if $m = 0$ and defining $\boldsymbol{W}_0 = \boldsymbol{I}_J$, the three operators above reduce to that defined in (2.11), (2.12) and (2.13) respectively. Using the same reasoning as before, we see that under

Assumption 1 with Properties 1 and 2 of Appendix C, $\boldsymbol{A}_m\boldsymbol{Y} + \boldsymbol{c}_m\boldsymbol{\eta}$ is a sub-Gaussian process under P_0. Analytical expression for the conditional posterior of $f^{(m)}$ given $(\boldsymbol{Y}, \sigma^2)$ with $m \in \mathbb{N}_0$ and $m < \alpha$ is available due to normal-normal conjugacy, and we derive its pointwise, L_2-and L_∞-posterior convergence rates by directly bounding the posterior mean square error. Note that Theorem 2.6 below is an important en-route step in deriving Theorems 2.7 and 2.8.

Theorem 2.6. *If $J \asymp n^{1/(2\alpha+1)}$, we have for $R > 0$, every fixed $x \in (0,1)$ and any $M_n \to \infty$,*

$$\textit{Empirical Bayes:} \sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R} \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(f : |f^{(m)}(x) - f_0^{(m)}(x)| > M_n n^{-(\alpha-m)/(2\alpha+1)} | \boldsymbol{Y}) \to 0.$$

$$\textit{Hierarchical Bayes:} \sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R} \mathrm{E}_0 \Pi(f : |f^{(m)}(x) - f_0^{(m)}(x)| > M_n n^{-(\alpha-m)/(2\alpha+1)} | \boldsymbol{Y}) \to 0.$$

Proof. The posterior rate calculation proceeds similarly as illustrated in the proof of Theorem 2.3. That is, by bounding the corresponding posterior variance, variance of the posterior mean, and the squared bias as in (2.38), and making the substitutions $\boldsymbol{A} = \boldsymbol{A}_m$, $\boldsymbol{c} = \boldsymbol{c}_m$ and $\Sigma = \Sigma_m$, where $\boldsymbol{A}_m, \boldsymbol{c}_m$ and Σ_m were defined in (2.64), (2.65) and (2.66) respectively. The only difference is the extra matrix factor $\boldsymbol{W}_m$ in these calculations.

Recall that each row of $\boldsymbol{W}_m$ has $m+1$ nonzero entries and each column has at most $m+1$ nonzero entries. From the explicit expressions of these nonzero entries given in (2.60)–(2.63), assumption (2.5) and Lemma A.1 imply that

$$|w_{i,i+h}^{(m)}| \asymp \frac{1}{\Delta^m} \asymp J^m \tag{2.67}$$

for $1 \le i \le J-m$, $0 \le h \le m$ and zero otherwise. This further implies that

$$\|\boldsymbol{W}_m\|_{(\infty,\infty)} = \max_{1 \le i \le J-m} \sum_{j=i}^{i+m} |w_{ij}^{(m)}| \asymp \frac{1}{\Delta^m} \asymp J^m,$$

$$\|\boldsymbol{W}_m^T\|_{(\infty,\infty)} = \max_{1 \le j \le J} \sum_{i=1}^{j} |w_{ij}^{(m)}| \asymp \frac{1}{\Delta^m} \asymp J^m$$

and

$$\|\boldsymbol{W}_m^T \boldsymbol{W}_m\|_{(2,2)} \le \|\boldsymbol{W}_m^T \boldsymbol{W}_m\|_{(\infty,\infty)} \le \|\boldsymbol{W}_m\|_{(\infty,\infty)} \|\boldsymbol{W}_m^T\|_{(\infty,\infty)} \asymp J^{2m}. \tag{2.68}$$

It follows that the posterior mean square error is inflated by the introduction of a J^{2m} factor in the bounds for the posterior variance and variance of the posterior mean, and a J^m in the bias. Also, the bound on B-spline approximation error in (2.20) but for $f_0^{(m)}$ is of the order $J^{-(\alpha-m)}$

uniformly in $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R$, as a result of applying Corollary 6.21 of Schumaker, 2007 (or Theorem 26, Chapter XII of de Boor, 2001). Therefore, the derivative analogs of (2.44) and (2.45) for any $x \in (0,1)$ are given by

$$\sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R} |\boldsymbol{A}_m(x)\boldsymbol{F}_0 + \boldsymbol{c}_m(x)\boldsymbol{\eta} - f_0^{(m)}(x)|^2 \lesssim \frac{J^{2m+2}}{n^2} + J^{-2(\alpha-m)}, \tag{2.69}$$

and

$$\begin{aligned}\sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R} \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}\left[\left(f^{(m)}(x) - f_0^{(m)}(x)\right)^2 \middle| \boldsymbol{Y}\right] &\lesssim \frac{J^{2m+1}}{n} + \frac{J^{2m+2}}{n^2} + J^{-2(\alpha-m)} \\ &\lesssim \frac{J^{2m+1}}{n} + J^{-2(\alpha-m)} \end{aligned} \tag{2.70}$$

respectively, where the last line above follows since $J \leq n$ by assumption. Setting the same order of J as in Theorem 2.3, i.e., $J \asymp n^{1/(2\alpha+1)}$ will balance the two orders on the right of (2.70) for all $m < \alpha$ including $m = 0$. Hence, the posterior mean square error will have order $n^{-2(\alpha-m)/(2\alpha+1)}$. The empirical and hierarchical Bayes posterior convergence rates then follow from (2.47) and (2.48) respectively with f and f_0 replaced by $f^{(m)}$ and $f_0^{(m)}$ respectively. □

Theorem 2.7. *If $J \asymp n^{1/(2\alpha+1)}$, we have for $R > 0$ and every $M_n \to \infty$,*

$$\begin{aligned} &\textit{Empirical Bayes:} && \sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R} \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(f : \|f^{(m)} - f_0^{(m)}\|_2 > M_n n^{-(\alpha-m)/(2\alpha+1)} | \boldsymbol{Y}) \to 0. \\ &\textit{Hierarchical Bayes:} && \sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R} \mathrm{E}_0 \Pi(f : \|f^{(m)} - f_0^{(m)}\|_2 > M_n n^{-(\alpha-m)/(2\alpha+1)} | \boldsymbol{Y}) \to 0. \end{aligned}$$

Proof. Following the steps in (2.51) with f and f_0 replaced by $f^{(m)}$ and $f_0^{(m)}$ respectively, the bound is $J^{2m+1}/n + J^{-2(\alpha-m)}$ in view of (2.70). Set $J \asymp n^{1/(2\alpha+1)}$ to balance the orders of the two terms. The result then follows from Markov's inequality. □

Theorem 2.8. *If $J \asymp (n/\log n)^{1/(2\alpha+1)}$, then for $R > 0$ and every $M_n \to \infty$, we have uniformly over $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R$,*

$$\begin{aligned} &\textit{Empirical Bayes:} && \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(f : \|f^{(m)} - f_0^{(m)}\|_\infty > M_n (\log n/n)^{(\alpha-m)/(2\alpha+1)} | \boldsymbol{Y}) \to 0. \\ &\textit{Hierarchical Bayes:} && \mathrm{E}_0 \Pi(f : \|f^{(m)} - f_0^{(m)}\|_\infty > M_n (\log n/n)^{(\alpha-m)/(2\alpha+1)} | \boldsymbol{Y}) \to 0. \end{aligned}$$

Proof. The calculations for finding the sup-norm rate as described in Theorem 2.5 apply with slight modifications. The only difference is the extra $\boldsymbol{W}_m$ matrix appearing in $\boldsymbol{A}_m, \boldsymbol{c}_m$ and Σ_m from (2.64), (2.65) and (2.66) respectively. Recall that $\Pi(f^{(m)}|\boldsymbol{Y}, \sigma^2) \sim \mathrm{GP}(\boldsymbol{A}_m\boldsymbol{Y} +$

$\boldsymbol{c}_m\boldsymbol{\eta}, \sigma^2\Sigma_m)$. Define $Z_{n,m} \sim \mathrm{GP}(0, \Sigma_m)$. Under P_0, $\boldsymbol{A}_m\boldsymbol{Y} + \boldsymbol{c}_m\boldsymbol{\eta}$ is a sub-Gaussian process with mean function $\boldsymbol{A}_m\boldsymbol{F}_0 + \boldsymbol{c}_m\boldsymbol{\eta}$ and covariance function $\sigma_0^2\Psi_m$, such that for any $x, y \in (0,1)$,

$$\Psi_m(x,y) = \boldsymbol{b}_{J,q,m}(x)^T\boldsymbol{W}_m\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\boldsymbol{B}^T\boldsymbol{B}\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\boldsymbol{W}_m^T\boldsymbol{b}_{J,q,m}(y). \quad (2.71)$$

Define $Q_{n,m}$ to be a sub-Gaussian process with mean function 0 and covariance function $\sigma_0^2\Psi_m$. From (2.68), we see that the $\|\cdot\|_{(2,2)}$ matrix norm of $\boldsymbol{W}_m$ is $O(J^m)$. For any fixed $x \in (0,1)$,

$$\begin{aligned}\Sigma_m(x,x) &= \boldsymbol{b}_{J,q,m}(x)^T\boldsymbol{W}_m\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\boldsymbol{W}_m^T\boldsymbol{b}_{J,q,m}(x)\\ &\le \|\boldsymbol{b}_{J,q,m}(x)\|^2\|\boldsymbol{W}_m^T\boldsymbol{W}_m\|_{(2,2)}\|\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\|_{(2,2)}\\ &\lesssim \frac{J^{2m+1}}{n}, \end{aligned} \quad (2.72)$$

where we have used (2.39), (2.68) and (2.19) to bound the norms in line 2 respectively. By similar reasoning, we have for any $x \in (0,1)$,

$$\Psi_m(x,x) \le \|\boldsymbol{b}_{J,q,m}(x)\|^2\|\boldsymbol{W}_m^T\boldsymbol{W}_m\|_{(2,2)}\|\boldsymbol{B}^T\boldsymbol{B}\|_{(2,2)}\left\|\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\right\|_{(2,2)} \lesssim \frac{J^{2m+1}}{n}, \quad (2.73)$$

where we have used (2.39), (2.68), (2.18) and (2.19) to bound the norms respectively. Since $Q_{n,m} = \boldsymbol{A}_m\boldsymbol{\varepsilon}$, then by Property 2 of Appendix C and Assumption 1, $Q_{n,m}$ is sub-Gaussian with respect to the semi-metric $d(t,s) = \sqrt{\mathrm{Var}(Q_{n,m}(t) - Q_{n,m}(s))}$. By Lemma A.4, both $Z_{n,m}$ and $Q_{n,m}$ satisfy the condition for Lemma C.1. Therefore, Lemma C.1 with $p = 2$ and the two upper bounds from (2.72), (2.73) imply that for any $0 < \delta_n < 1$ and $T_n = N(\delta_n, (0,1), \|\cdot\|)$,

$$\begin{aligned}\sup_{\sigma^2\in\mathcal{U}_n}\sigma^2\mathrm{E}(\|Z_{n,m}\|_\infty^2) &\lesssim \sigma_0^2\log(1/\delta_n)\left[n\delta_n^2 + \max_{1\le i\le T_n}\Sigma_m(u_i,u_i)\right]\\ &\lesssim \log(1/\delta_n)\left(n\delta_n^2 + \frac{J^{2m+1}}{n}\right)\end{aligned} \quad (2.74)$$

and

$$\mathrm{E}(\|Q_{n,m}\|_\infty^2) \lesssim \log(1/\delta_n)\left[n\delta_n^2 + \max_{1\le i\le T_n}\sigma_0^2\Psi_m(u_i,u_i)\right] \lesssim \log(1/\delta_n)\left(n\delta_n^2 + \frac{J^{2m+1}}{n}\right). \quad (2.75)$$

By setting $\delta_n^2 \asymp J^{2m+1}/n^2$ for both (2.74) and (2.75), we have

$$\sup_{\sigma^2\in\mathcal{U}_n}\sigma^2\mathrm{E}(\|Z_{n,m}\|_\infty^2) \lesssim (J^{2m+1}/n)\log n, \qquad \mathrm{E}(\|Q_{n,m}\|_\infty^2) \lesssim (J^{2m+1}/n)\log n. \quad (2.76)$$

Since the bound for (2.69) is uniform for all $x \in (0,1)$ and all $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R$, we have

$$\sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R} \|\boldsymbol{A}_m \boldsymbol{F}_0 + \boldsymbol{c}_m \boldsymbol{\eta} - f_0^{(m)}\|_\infty^2 = O\left(\frac{J^{2m+2}}{n^2}\right) + O(J^{-2(\alpha-m)}). \tag{2.77}$$

Therefore, the derivative analog of (2.57) with f and f_0 replaced by $f^{(m)}$ and $f_0^{(m)}$ is

$$\sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R} \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}(\|f^{(m)} - f_0^{(m)}\|_\infty^2 | \boldsymbol{Y}) \lesssim \frac{J^{2m+1}}{n} \log n + J^{-2(\alpha-m)}. \tag{2.78}$$

Setting the same $J \asymp (n/\log n)^{1/(2\alpha+1)}$ as in Theorem 2.5 will balance the two terms above for all $m < \alpha$ including $m = 0$, yielding the rate $(\log n/n)^{2(\alpha-m)/(2\alpha+1)}$ for the posterior mean square error in (2.78). The empirical and hierarchical Bayes posterior convergence rates then follow from (2.47) and (2.48) respectively with absolute values replaced by sup-norms and f, f_0 replaced by $f^{(m)}, f_0^{(m)}$ respectively. □

Given any mode of convergence, i.e., under pointwise, L_2-or L_∞-norms, the same optimal J applies to both f and all its derivatives such that $m < \alpha$. If we view $f^{(0)} \equiv f$, then posterior convergence rates established in Section 2.3 are special cases of Section 2.4 by setting $m = 0$. These rates coincide with the minimax rates for estimating $f^{(m)}(x)$ at any $x \in (0,1)$ (Stone, 1980), $f^{(m)}$ in L_2-and L_∞-norms (Theorem 1 of Stone, 1982). It also follows from the proof that the posterior mean of $f^{(m)}(x)$ converges to $f_0^{(m)}(x)$ at any $x \in (0,1)$ at the rate given in Theorem 2.6. By the same reasoning, the posterior mean of $f^{(m)}$ converges to $f_0^{(m)}$ under L_2-and L_∞-norms at the rates given by Theorems 2.7 and 2.8 respectively.

2.5 Credible sets for f and its derivatives

In this section, we will treat f and its derivatives in a unified framework by viewing f as $f^{(0)}$. Then if we define $\boldsymbol{W}_0 = \boldsymbol{I}_J$, we will have $\boldsymbol{A}_0 = \boldsymbol{A}, \boldsymbol{c}_0 = \boldsymbol{c}$ and $\Sigma_0 = \Sigma$ where $\boldsymbol{A}_m, \boldsymbol{c}_m$ and Σ_m were given in (2.64), (2.65) and (2.66) respectively. We begin by constructing pointwise credible set for $f^{(m)}(x)$ at $x \in (0,1)$. Let $\omega_n \in [0,1]$ be a sequence such that $\omega_n \to 0$ as $n \to \infty$. Define z_δ to be the upper δ-quantile of a standard normal. Using the fact that $P(Z > z) \leq z^{-1} \exp(-z^2/2)$ for $Z \sim \mathrm{N}(0,1)$ and $z > 0$, we then have

$$z_{\omega_n/2} = O(\sqrt{-\log \omega_n}). \tag{2.79}$$

In particular, if $\omega_n \gtrsim 1/n$, then $z_{\omega_n/2} = O(\sqrt{\log n})$; and if $\omega_n \gtrsim 1/\log n$, then $z_{\omega_n/2} = O(\sqrt{\log \log n})$. Since $\Pi(f^{(m)}(x)|\boldsymbol{Y}, \sigma^2) \sim \mathrm{N}(\boldsymbol{A}_m(x)\boldsymbol{Y} + \boldsymbol{c}_m(x)\boldsymbol{\eta}, \sigma^2 \Sigma_m(x,x))$, we can construct

a $1-\omega_n$ pointwise credible interval for $f^{(m)}(x)$ at any $x \in (0,1)$ as

$$\Pi(f : |f^{(m)}(x) - \boldsymbol{A}_m(x)\boldsymbol{Y} - \boldsymbol{c}_m(x)\boldsymbol{\eta}| \le z_{\omega_n/2}\sigma\sqrt{\Sigma_m(x,x)}|\boldsymbol{Y}, \sigma^2) = 1-\omega_n. \tag{2.80}$$

Since σ^2 is unknown, we will consider the empirical Bayes approach by substituting σ^2 for $\widehat{\sigma}_n^2$ derived in (2.14), leading to the following empirical credible set:

$$\widehat{\mathcal{C}}(x) = \{f : |f^{(m)}(x) - \boldsymbol{A}_m(x)\boldsymbol{Y} - \boldsymbol{c}_m(x)\boldsymbol{\eta}| \le z_{\omega_n/2}\widehat{\sigma}_n\sqrt{\Sigma_m(x,x)}\}, \tag{2.81}$$

where by construction, $\Pi_{\widehat{\sigma}_n}(\widehat{\mathcal{C}}(x)|\boldsymbol{Y}) = 1-\omega_n$. We shall study the frequentist coverage of $\widehat{\mathcal{C}}(x)$ at any $x \in (0,1)$ as $n \to \infty$, which is given by

$$\inf_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R} P_0(f_0 \in \widehat{\mathcal{C}}(x)). \tag{2.82}$$

Theorem 2.9. *If $J \asymp n^{1/(2\alpha+1)}$, then the asymptotic coverage of (2.82) is 1, and the radius of $\widehat{\mathcal{C}}(x)$ is $O_{P_0}(n^{-(\alpha-m)/(2\alpha+1)}\sqrt{-\log\omega_n})$ uniformly over $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R$ at any $x \in (0,1)$, $R>0$ and all $m < \alpha$.*

Proof. For any sequence $\zeta_n = o(\sqrt{n})$, define $t_m(x) = \inf_{|\sigma^2-\sigma_0^2| \le \zeta_n/\sqrt{n}} z_{\omega_n/2}\sigma\sqrt{\Sigma_m(x,x)}$. By (2.46), to show (2.82) has asymptotic coverage of 1, it suffices to show that

$$\sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R} P_0\left(|f_0^{(m)}(x) - \boldsymbol{A}_m(x)\boldsymbol{Y} - \boldsymbol{c}_m(x)\boldsymbol{\eta}| > t_m(x)\right) \to 0. \tag{2.83}$$

Since $z_{\omega_n/2} \to \infty$ as $\omega_n \to 0$ and $\inf_{|\sigma^2-\sigma_0^2| \le \zeta_n/\sqrt{n}} \sigma^2 = \sigma_0^2 + o(1)$, we have $t_m(x)^2 \gg \Sigma_m(x,x)$. By (2.19), $\Sigma_m(x,x) \ge \lambda_{\min}\{(\boldsymbol{B}^T\boldsymbol{B}+\boldsymbol{\Omega}^{-1})^{-1}\}\|\boldsymbol{W}_m^T\boldsymbol{b}_{J,q,m}(x)\|^2 \gtrsim (J/n)\|\boldsymbol{W}_m^T\boldsymbol{b}_{J,q,m}(x)\|^2$. Hence, we need a lower bound for $\|\boldsymbol{W}_m^T\boldsymbol{b}_{J,q,m}(x)\|^2 = \sum_{j=1}^J B_{j,q}^{(m)}(x)^2$. For any $x \in (0,1)$, let i_x be a positive integer such that $x \in [t_{i_x-1}, t_{i_x}]$. Hence, only $(B_{i_x,q-m}(x)\ldots, B_{i_x+q-m-1,q-m}(x))^T$ are nonzero. By (2.60)–(2.63), $B_{j,q}^{(m)}(x)$ is a linear combination of $(B_{j,q-m}(x),\ldots,B_{j+m,q-m}(x))^T$ for any $1 \le j \le J$. Choose $j = i_x + q - m - 1$, and we have

$$\begin{aligned}\|\boldsymbol{W}_m^T\boldsymbol{b}_{J,q,m}(x)\|^2 = \sum_{j=1}^J (B_{j,q}^{(m)}(x))^2 &\ge \prod_{u=1}^m \frac{(q-u)^2}{(t_{i_x+q-m-1} - t_{i_x-m-1+u})^2} B_{i_x+q-m-1,q-m}(x)^2 \\ &\ge \frac{1}{\Delta^{2m}}\left(\frac{q-m}{q}\right)^2 \gtrsim J^{2m},\end{aligned} \tag{2.84}$$

where the last line follows since (2.5) implies that $t_{i_x+q-m-1} - t_{i_x-m-1+u} \le (q-u+1)\Delta$, and we used Lemma A.1 to obtain the final lower bound. Consequently, $t_m(x)^2 \gg J^{2m+1}/n$. Under the true distribution P_0, $\boldsymbol{A}_m(x)\boldsymbol{Y} + \boldsymbol{c}_m(x)\boldsymbol{\eta}$ is a sub-Gaussian random variable with mean

$\boldsymbol{A}_m(x)\boldsymbol{F}_0 + \boldsymbol{c}_m(x)\boldsymbol{\eta}$ and variance $\sigma_0^2\Psi_m(x,x)$, where the expression for $\Psi_m(x,x)$ at $x \in (0,1)$ is given by (2.71). Define $Q_{n,m}(x)$ to be a sub-Gaussian random variable with mean 0 and variance $\sigma_0^2\Psi_m(x,x)$. Then uniformly in $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R$,

$$\begin{aligned}\mathrm{E}_0|f_0^{(m)}(x) - \boldsymbol{A}_m(x)\boldsymbol{Y} - \boldsymbol{c}_m(x)\boldsymbol{\eta}|^2 &= \mathrm{E}_0[\boldsymbol{A}_m(x)\boldsymbol{F}_0 + \boldsymbol{c}_m(x)\boldsymbol{\eta} - f_0^{(m)}(x) + Q_{n,m}(x)]^2 \\ &\le 2\mathrm{E}[Q_{n,m}(x)^2] + 2|\boldsymbol{A}_m(x)\boldsymbol{F}_0 + \boldsymbol{c}_m(x)\boldsymbol{\eta} - f_0^{(m)}(x)|^2 \\ &\lesssim \frac{J^{2m+1}}{n} + J^{-2(\alpha-m)},\end{aligned}$$

where the last line follows from (2.73) and (2.69) respectively. By Markov's inequality,

$$\sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R} P_0(|f_0^{(m)}(x) - \boldsymbol{A}_m(x)\boldsymbol{Y} - \boldsymbol{c}_m(x)\boldsymbol{\eta}| > t_m(x)) \lesssim \frac{J^{2m+1}/n + J^{-2(\alpha-m)}}{t_m(x)^2}.$$

Since $t_m(x)^2 \gg J^{2m+1}/n$, we have $t_m(x)^2 \gg J^{-2(\alpha-m)}$ if $J \asymp n^{1/(2\alpha+1)}$. Hence the right hand side above will go to zero as $n \to \infty$. We conclude that $\inf_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R} P_0(|f_0^{(m)}(x) - \boldsymbol{A}_m(x)\boldsymbol{Y} - \boldsymbol{c}_m(x)\boldsymbol{\eta}| \le t_m(x)) \to 1$ if $J \asymp n^{1/(2\alpha+1)}$. Let $h_m(x) = z_{\omega_n/2}\sqrt{\Sigma_m(x,x)}$ at $x \in (0,1)$. By (2.79), we have $z_{\omega_n/2} = O(\sqrt{-\log\omega_n})$. Since $\Sigma_m(x,x) \lesssim J^{2m+1}/n$ by (2.72), we conclude using (2.46) that $\widehat{\sigma}_n h_m(x) = O_{P_0}(n^{-(\alpha-m)/(2\alpha+1)}\sqrt{-\log\omega_n})$ uniformly over $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R$ at any $x \in (0,1)$. □

The conditional posterior distribution for $f^{(m)}$ is $\Pi(f^{(m)}|\boldsymbol{Y},\sigma^2) \sim \mathrm{GP}(\boldsymbol{A}_m\boldsymbol{Y} + \boldsymbol{c}_m\boldsymbol{\eta}, \sigma^2\Sigma_m)$. The conditional distribution of $(f^{(m)} - \boldsymbol{A}_m\boldsymbol{Y} - \boldsymbol{c}_m\boldsymbol{\eta})/\sigma$ given $(\boldsymbol{Y},\sigma^2)$ is $\mathrm{GP}(0,\Sigma_m)$, and hence does not depend on $\boldsymbol{Y}$ and σ^2. This enables us to construct L_2-credible ball for $f^{(m)}$, centered at its posterior mean with radius σh_m, where h_m is chosen such that its posterior mass is exactly $1-\omega_n$. That is,

$$\Pi(f : \|f^{(m)} - \boldsymbol{A}_m\boldsymbol{Y} - \boldsymbol{c}_m\boldsymbol{\eta}\|_2 \le \sigma h_m|\boldsymbol{Y},\sigma^2) = 1-\omega_n, \tag{2.85}$$

where $h_m \to \infty$ as $\omega_n \to 0$ and h_m does not depend on $\boldsymbol{Y}$. The corresponding empirical L_2-credible set is

$$\widehat{\mathcal{C}}_2 = \{f : \|f^{(m)} - \boldsymbol{A}_m\boldsymbol{Y} - \boldsymbol{c}_m\boldsymbol{\eta}\|_2 \le \widehat{\sigma}_n h_m\}, \tag{2.86}$$

where by construction, $\Pi_{\widehat{\sigma}_n}(\widehat{\mathcal{C}}_2|\boldsymbol{Y}) = 1-\omega_n$. We shall study the frequentist coverage of $\widehat{\mathcal{C}}_2$ as $n \to \infty$, which is given by

$$\inf_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R} P_0(f_0 \in \widehat{\mathcal{C}}_2). \tag{2.87}$$

Theorem 2.10. *If $J \asymp n^{1/(2\alpha+1)}$, the asymptotic coverage of* (2.87) *is* 1*, and the radius of $\widehat{\mathcal{C}}_2$ is $O_{P_0}(n^{-(\alpha-m)/(2\alpha+1)}\sqrt{-\log\omega_n})$ uniformly over $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R$ for $R > 0$ and all $m < \alpha$.*

Proof. For any sequence $\zeta_n = o(\sqrt{n})$, define $t_m = \inf_{|\sigma^2-\sigma_0^2|\le\zeta_n/\sqrt{n}} \sigma h_m$. By (2.46), to show (2.87) has asymptotic coverage of 1, it suffices to show

$$\sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}}\le R} P_0\left(\|f_0^{(m)} - \boldsymbol{A}_m\boldsymbol{Y} - \boldsymbol{c}_m\boldsymbol{\eta}\|_2 > t_m\right) \to 0. \tag{2.88}$$

Let $Z_{n,m} \sim \mathrm{GP}(0, \Sigma_m)$. Now $\Pi_{\widehat{\sigma}_n}(\widehat{\mathcal{C}}_2|\boldsymbol{Y}) = 1-\omega_n$ implies that $P(\|Z_{n,m}\|_2 > h_m) \to 0$. Observe that $\int_0^1 Z_{n,m}(x)dx$ is normally distributed, and by Fubini's theorem, its mean is 0 and variance is $\int_0^1\int_0^1 \Sigma_m(x,y)dxdy$. Now since $\int_0^1 Z_{n,m}(x)dx \le \|Z_{n,m}\|_1 \le \|Z_{n,m}\|_2$ a.s., we have

$$1 - \Phi\left(\frac{h_m}{\sqrt{\int_0^1\int_0^1 \Sigma_m(x,y)dxdy}}\right) \le P(\|Z_{n,m}\|_2 > h_m) \to 0. \tag{2.89}$$

Hence, $h_m^2 \gg \int_0^1\int_0^1 \Sigma_m(x,y)dxdy$. To derive a lower bound for $\int_0^1\int_0^1 \Sigma_m(x,y)dxdy$, we first need to establish lower bounds for integrals of B-splines, it turns out that by using the last two equations in the proof of Lemma 6.7 in Zhou et al. (1998), we have

$$\int_0^1 B_{j,q}(x)dx \ge \frac{1}{C^q q!} \tag{2.90}$$

for any $1 \le j \le J$, where the constant C is from the quasi-uniform knot assumption in (2.5). Using the expression for $\Sigma_m(x,y)$ at $x, y \in (0,1)$ given in (2.66),

$$\begin{aligned}
\int_0^1\int_0^1 \Sigma_m(x,y)dxdy &= \left(\int_0^1 \boldsymbol{b}_{J,q,m}(x)dx\right)^T \boldsymbol{W}_m(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1})^{-1}\boldsymbol{W}_m^T\left(\int_0^1 \boldsymbol{b}_{J,q,m}(x)dx\right) \\
&\ge \lambda_{\min}\left\{(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1})^{-1}\right\}\left\|\boldsymbol{W}_m^T\left(\int_0^1 \boldsymbol{b}_{J,q,m}(x)dx\right)\right\|^2 \\
&\gtrsim \frac{J}{n}\left\|\boldsymbol{W}_m^T\left(\int_0^1 \boldsymbol{b}_{J,q,m}(x)dx\right)\right\|^2,
\end{aligned} \tag{2.91}$$

where the last line follows from (2.19). Now using expressions for the entries of $\boldsymbol{W}_m$ described

in (2.60)–(2.63) and the lower bound in (2.90),

$$
\begin{aligned}
&\left\| \boldsymbol{W}_m^T \int_0^1 \boldsymbol{b}_{J,q,m}(x)dx \right\|^2 \\
&= \sum_{j=1}^{m} \left(\sum_{i=1}^{j} w_{ij}^{(m)} \int_0^1 B_{i,q-m}(x)dx \right)^2 + \sum_{j=1}^{J-2m} \left(\sum_{i=0}^{m} w_{j+i,j+m}^{(m)} \int_0^1 B_{j+i,q-m}(x)dx \right)^2 \\
&\quad + \sum_{j=0}^{m-1} \left(\sum_{i=0}^{j} w_{J-m-i,J-j}^{(m)} \int_0^1 B_{J-m-i,q-m}(x)dx \right)^2 \\
&\geq \left(\prod_{l=1}^{m} \frac{q-l}{t_1 - t_{1-q+l}} \right)^2 \left(\int_0^1 B_{1,q-m}(x)dx \right)^2 \\
&\geq \frac{1}{[C^{q-1}(q-1)!]^2} \left(\frac{q-m}{q} \right)^2 \frac{1}{\Delta^{2m}} \asymp N^{2m} \asymp J^{2m}.
\end{aligned}
$$

The last line follows since (2.5) implies that $t_1 - t_{1-q+l} \leq (q-l+1)\Delta$ for $l = 1, \ldots, m$, and we applied Lemma A.1 to obtain the final lower bound. In view of (2.91), we have $\int_0^1 \int_0^1 \Sigma_m(x,y)dxdy \gtrsim J^{2m+1}/n$. Therefore, $t_m^2 \gtrsim \sigma_0^2 h_m \gg J^{2m+1}/n$.

Under the true distribution P_0, $\boldsymbol{A}_m\boldsymbol{Y} + \boldsymbol{c}_m\boldsymbol{\eta}$ is a sub-Gaussian process with mean function $\boldsymbol{A}_m\boldsymbol{F}_0 + \boldsymbol{c}_m\boldsymbol{\eta}$ and covariance function $\sigma_0^2\Psi_m$, where the expression for $\Psi_m(x,y)$ at $x, y \in (0,1)$ is given by (2.71). Define $Q_{n,m}$ to be a sub-Gaussian process with mean function 0 and covariance function $\sigma_0^2\Psi_m$. Then uniformly in $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R$,

$$
\begin{aligned}
\mathrm{E}_0\|f_0^{(m)} - \boldsymbol{A}_m\boldsymbol{Y} - \boldsymbol{c}_m\boldsymbol{\eta}\|_2^2 &= \mathrm{E}\|\boldsymbol{A}_m\boldsymbol{F}_0 + \boldsymbol{c}_m\boldsymbol{\eta} - f_0^{(m)} + Q_{n,m}\|_2^2 \\
&\leq 2\mathrm{E}\|Q_{n,m}\|_2^2 + 2\|\boldsymbol{A}_m\boldsymbol{F}_0 + \boldsymbol{c}_m\boldsymbol{\eta} - f_0^{(m)}\|_2^2 \\
&= 2\int_0^1 \Psi_m(x,x)dx + 2\int_0^1 |\boldsymbol{A}_m(x)\boldsymbol{F}_0 + \boldsymbol{c}_m(x)\boldsymbol{\eta} - f_0^{(m)}(x)|^2 dx \\
&\lesssim \frac{J^{2m+1}}{n} + J^{-2(\alpha-m)},
\end{aligned}
$$

where the last line follows from (2.73) and (2.69) respectively. By Markov's inequality,

$$
\sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R} P_0(\|f_0^{(m)} - \boldsymbol{A}_m\boldsymbol{Y} - \boldsymbol{c}_m\boldsymbol{\eta}\|_2 > t_m) \lesssim \frac{J^{2m+1}/n + J^{-2(\alpha-m)}}{t_m^2}.
$$

Since $t_m^2 \gg J^{2m+1}/n$, we have $t_m^2 \gg J^{-2(\alpha-m)}$ if $J \asymp n^{1/(2\alpha+1)}$. Hence the right hand side will go to zero as $n \to \infty$. We conclude that $\inf_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R} P_0(\|f_0^{(m)} - \boldsymbol{A}_m\boldsymbol{Y} - \boldsymbol{c}_m\boldsymbol{\eta}\|_2 \leq t_m) \to 1$ if $J \asymp n^{1/(2\alpha+1)}$. Applying (3.5) from Ledoux and Talagrand (1991) to $Z_{n,m}$ with $\|\cdot\|$ as the

L_2-norm, we have

$$\omega_n = P(\|Z_{n,m}\|_2 > h_m) \le 4\exp\left(-\frac{h_m^2}{8\mathrm{E}\|Z_{n,m}\|_2^2}\right).$$

It then follows that $h_m \lesssim (\mathrm{E}\|Z_{n,m}\|_2^2)^{1/2}\sqrt{-\log\omega_n}$. By Fubini's theorem and (2.40), we have $\mathrm{E}\|Z_{n,m}\|_2^2 = \int_0^1 \Sigma(x,x)dx \lesssim J^{2m+1}/n$. In view of (2.46), we have uniformly over $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R$ that the radius of $\widehat{\mathcal{C}}_2$ is $\widehat{\sigma}_n h_m = O_{P_0}(n^{-(\alpha-m)/(2\alpha+1)}\sqrt{-\log\omega_n})$. □

We will use a slightly different approach to construct L_∞-credible region for $f^{(m)}$ of size $1-\omega$. Note that here $\omega \in [0,1]$ does not depend on n and we assume that $\omega < 1/2$. We choose a sequence $h_{\infty,m}$ such that

$$\Pi_{\widehat{\sigma}_n}(f : \|f^{(m)} - \boldsymbol{A}_m\boldsymbol{Y} - \boldsymbol{c}_m\boldsymbol{\eta}\|_\infty \le \widehat{\sigma}_n h_{\infty,m}|\boldsymbol{Y}) = 1-\omega, \tag{2.92}$$

where $h_{\infty,m}$ does not depend on $\boldsymbol{Y}$. As we shall see in Theorem 2.11 below, in order to ensure frequentist coverage of 1 asymptotically, we need to introduce an inflation factor ρ_n in the radius, such that $\rho_n \to \infty$ as $n \to \infty$. That is, we consider the following inflated credible region

$$\widehat{\mathcal{C}}_\infty = \{f : \|f^{(m)} - \boldsymbol{A}_m\boldsymbol{Y} - \boldsymbol{c}_m\boldsymbol{\eta}\|_\infty \le \rho_n\widehat{\sigma}_n h_{\infty,m}\}, \tag{2.93}$$

where by construction, $\Pi_{\widehat{\sigma}_n}(\widehat{\mathcal{C}}_\infty|\boldsymbol{Y}) \ge 1-\omega$ if $\rho_n \ge 1$. We shall study the frequentist coverage of $\widehat{\mathcal{C}}_\infty$ as $n \to \infty$, which is given by

$$\inf_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R} P_0(f_0 \in \widehat{\mathcal{C}}_\infty). \tag{2.94}$$

Theorem 2.11. *If $J \asymp (n/\log n)^{1/(2\alpha+1)}$, the asymptotic coverage of* (2.94) *is* 1 *and the radius of $\widehat{\mathcal{C}}_\infty$ is $O_{P_0}((\log n/n)^{(\alpha-m)/(2\alpha+1)}\rho_n)$, uniformly in $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R$ for $R > 0$ and any $m < \alpha$.*

Remark To simultaneously control the size of $\widehat{\mathcal{C}}_\infty$ and ensure guaranteed frequentist coverage, we can take ρ_n to be a slowly varying factor tending to infinity. For example, $\rho_n = (\log\log n)^{1/3}$ gives excellent results (see the simulation experiment in Section 2.6).

Proof. For any $\zeta_n = o(\sqrt{n})$, let $t_{\infty,m} = \inf_{|\sigma^2-\sigma_0^2|\le\zeta_n/\sqrt{n}}\sigma h_{\infty,m}$. By (2.46), to show (2.94) has asymptotic coverage of 1 , it suffices to show

$$\sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R} P_0(\|f_0^{(m)} - \boldsymbol{A}_m\boldsymbol{Y} - \boldsymbol{c}_m\boldsymbol{\eta}\|_\infty > \rho_n t_{\infty,m}) \to 0. \tag{2.95}$$

Let $Z_{n,m} \sim \mathrm{GP}(0, \Sigma_m)$. Define $M(Z_{n,m})$ to be the median of $\|Z_{n,m}\|_\infty$, i.e., $M(Z_{n,m})$ is a number satisfying both $P(\|Z_{n,m}\|_\infty \leq M(Z_{n,m})) \geq 1/2$ and $P(\|Z_{n,m}\|_\infty \geq M(Z_{n,m})) \geq 1/2$. In addition, we define $\sigma_Z^2 = \sup_{x\in(0,1)} \mathrm{Var}(Z_{n,m}(x))$ and note that by (2.72), $\sigma_Z^2 \lesssim J^{2m+1}/n$. Using the fact that $\sigma_Z \leq 2M(Z_{n,m})$ and $|\mathrm{E}\|Z_{n,m}\|_\infty - M(Z_{n,m})| \leq \sigma_Z(\pi/2)^{1/2}$ (see Pages 52 and 54 of Ledoux and Talagrand, 1991), we have $\mathrm{E}\|Z_{n,m}\|_\infty \leq (1 + 2(\pi/2)^{1/2})M(Z_{n,m})$. Since $\sigma_Z^2 = o(1)$ for $J \asymp (n/\log n)^{1/(2\alpha+1)}$, we have $\mathrm{E}\|Z_{n,m}\|_\infty \geq M(Z_{n,m}) + o(1)$. Consequently, $\mathrm{E}\|Z_{n,m}\|_\infty \asymp M(Z_{n,m})$.

Now, (2.92) implies that $P(\|Z_{n,m}\|_\infty > h_{\infty,m}) = \omega$. Since $\omega < 1/2$ by assumption, this implies that $h_{\infty,m} \geq M(Z_{n,m}) \asymp \mathrm{E}\|Z_{n,m}\|_\infty$. Hence, we need a lower bound for $\mathrm{E}\|Z_{n,m}\|_\infty$. Define $V_i = Z_{n,m}(t_i)$, where t_i with $1 \leq i \leq N$ are the interior knots of the B-spline. By the mean value theorem and (2.84), we have for some point $t^* \in \lambda t_i + (1-\lambda)t_r$ with $\lambda \in [0,1]$ that

$$\sum_{j=1}^{J} \left(B_{j,q}^{(m)}(t_i) - B_{j,q}^{(m)}(t_r)\right)^2 = \sum_{j=1}^{J} \left(B_{j,q}^{(m+1)}(t^*)\right)^2 (t_i - t_r)^2 \gtrsim J^{2m+2} \min_{1\leq l\leq N} \delta_l^2 \gtrsim J^{2m}, \quad (2.96)$$

where $\delta_l = t_l - t_{l-1}$ for $1 \leq l \leq N$, and the last inequality follows from assumption (2.5) and Lemma A.1. Then for any $1 \leq i, r \leq N$,

$$\begin{aligned}
\mathrm{E}(V_i - V_r)^2 &= (\boldsymbol{b}_{J,q,m}(t_i) - \boldsymbol{b}_{J,q,m}(t_r))^T \boldsymbol{W}_m (\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1})^{-1} \boldsymbol{W}_m^T (\boldsymbol{b}_{J,q,m}(t_i) - \boldsymbol{b}_{J,q,m}(t_r)) \\
&\geq \lambda_{\min}\{(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1})^{-1}\} \|\boldsymbol{W}_m^T (\boldsymbol{b}_{J,q,m}(t_i) - \boldsymbol{b}_{J,q,m}(t_r))\|^2 \\
&\gtrsim \frac{J}{n} \sum_{j=1}^{J} \left(B_{j,q}^{(m)}(t_i) - B_{j,q}^{(m)}(t_r)\right)^2 \geq c\frac{J^{2m+1}}{n}, \qquad (2.97)
\end{aligned}$$

where $c > 0$ is a universal constant not depending on n, and we used (2.19) to lower bound the minimum eigenvalue of $(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1})^{-1}$. Define $U_i = \sqrt{2n/(cJ^{2m+1})} V_i$ and H_i i.i.d $\mathrm{N}(0,1)$ for $1 \leq i \leq N$. By equation (3.14) of Ledoux and Talagrand (1991), we have $\mathrm{E}(\max_{1\leq i\leq N} H_i) \gtrsim \sqrt{\log N}$. Therefore for any $1 \leq i, r \leq N$,

$$\mathrm{E}(U_i - U_r)^2 \geq 2 = \mathrm{E}(H_i - H_r)^2, \qquad (2.98)$$

and by Slepian's Lemma (Corollary 3.14 of Ledoux and Talagrand, 1991), we have

$$\mathrm{E}\left(\max_{1\leq i\leq N} U_i\right) \geq \mathrm{E}\left(\max_{1\leq i\leq N} H_i\right) \gtrsim \sqrt{\log N}. \qquad (2.99)$$

Since $N \sim J$ by Lemma A.1, we have $t_{\infty,m}^2 \gtrsim \sigma_0^2 h_{\infty,m}^2 \gtrsim (\mathrm{E}\|Z_{n,m}\|_\infty)^2 \gtrsim (J^{2m+1}/n)\log n$. Then

uniformly in $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R$,

$$\begin{aligned}
\mathrm{E}_0(\|f_0^{(m)} - \boldsymbol{A}_m\boldsymbol{Y} - \boldsymbol{c}_m\boldsymbol{\eta}\|_\infty^2) &= \mathrm{E}(\|\boldsymbol{A}_m\boldsymbol{F}_0 + \boldsymbol{c}_m\boldsymbol{\eta} - f_0^{(m)} + Q_{n,m}\|_\infty^2) \\
&\leq 2\mathrm{E}(\|Q_{n,m}\|_\infty^2) + 2\mathrm{E}(\|\boldsymbol{A}_m\boldsymbol{F}_0 + \boldsymbol{c}_m\boldsymbol{\eta} - f_0^{(m)}\|_\infty^2) \\
&\lesssim \frac{J^{2m+1}}{n}\log n + \frac{J^{2m+2}}{n^2} + J^{-2(\alpha-m)} \\
&\lesssim \frac{J^{2m+1}}{n}\log n + J^{-2(\alpha-m)}, \qquad (2.100)
\end{aligned}$$

where the last two lines follow from (2.76) and (2.69) respectively. By Markov's inequality,

$$\sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R} P_0(\|f_0^{(m)} - \boldsymbol{A}_m\boldsymbol{Y} - \boldsymbol{c}_m\boldsymbol{\eta}\|_\infty > \rho_n t_{\infty,m}) \lesssim \frac{(J^{2m+1}/n)\log n + J^{-2(\alpha-m)}}{\rho_n^2 t_{\infty,m}^2}. \quad (2.101)$$

Now, since $t_{\infty,m}^2 \gtrsim (J^{2m+1}/n)\log n$, then $\rho_n^2 t_{\infty,m}^2 \gg (J^{2m+1}/n)\log n$. If $J \asymp (n/\log n)^{1/(2\alpha+1)}$, we have $\rho_n^2 t_{\infty,m}^2 \gg J^{-2(\alpha-m)}$. Hence the fraction on the right hand side goes to 0 as $n \to \infty$. We conclude that $\inf_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R} P_0(\|f_0^{(m)} - \boldsymbol{A}_m\boldsymbol{Y} - \boldsymbol{c}_m\boldsymbol{\eta}\|_\infty \leq \rho_n t_{\infty,m}) \to 1$ for $J \asymp (n/\log n)^{1/(2\alpha+1)}$. Applying (3.5) from Ledoux and Talagrand (1991) to $Z_{n,m}$ with $\|\cdot\|$ as the L_∞-norm, we have

$$\omega = P(\|Z_{n,m}\|_\infty > h_{\infty,m}) \leq 4\exp\left(-\frac{h_{\infty,m}^2}{8\mathrm{E}\|Z_{n,m}\|_\infty^2}\right).$$

Hence, $h_{\infty,m} \lesssim (\mathrm{E}\|Z_{n,m}\|_\infty^2)^{1/2}\sqrt{-\log\omega}$. By (2.76), we have $\mathrm{E}\|Z_{n,m}\|_\infty^2 \lesssim (J^{2m+1}/n)\log n$. In view of (2.46), we have uniformly over $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \leq R$ that the radius of $\widehat{\mathcal{C}}_\infty$ is $\rho_n\widehat{\sigma}_n h_{\infty,m} = O_{P_0}((\log n/n)^{(\alpha-m)/(2\alpha+1)}\rho_n)$. □

Remark If we allow $\omega = \omega_n \to 0$ without using ρ_n, as in the construction for pointwise and L_2-credible regions, we need $h_{\infty,m}^2 \gg (J^{2m+1}/n)\log n$ to make the right hand side of (2.101) approach 0. We currently are unable to show this but conjecture that the lower bound for $h_{\infty,m}^2$ established above, i.e., $(J^{2m+1}/n)\log n$ is sharp regardless whether $\omega < 1/2$ is fixed or approaching 0. The introduction of an inflation factor is not unique to this problem, and it was used by Szabo et al. (2014) in the context of constructing adaptive L_2-credible regions.

Another way to ensure guaranteed frequentist coverage is to undersmooth the prior, i.e., using more B-splines in the expansion than the optimal number. However, it turns out that we need two different priors to accomplish this task for L_∞-credible sets. Namely, we use the optimal J to estimate the center $\boldsymbol{A}_m\boldsymbol{Y} + \boldsymbol{c}_m\boldsymbol{\eta}$ and a larger J to estimate the size of the radius. We write f as f_J to show that J B-splines have been used in the expansion of (2.3). In a similar manner,

we write $\boldsymbol{A}_m^J$ for $\boldsymbol{A}_m$, $\boldsymbol{c}_m^J$ for $\boldsymbol{c}_m$, Σ_m^J for Σ_m and $\widehat{\sigma}_{n,J}^2$ for $\widehat{\sigma}_n^2$. Let $J_n \asymp (n/\log n)^{1/(2\alpha+1)}$ be the optimal number of basis. Using the same construction as in (2.93), we consider the following empirical credible region for f_J with radius $\widehat{\sigma}_{n,J} h_{J,m}$,

$$\widehat{\mathcal{C}}_\infty^J = \{f_J : \|f_J^{(m)} - \boldsymbol{A}_m^J \boldsymbol{Y} - \boldsymbol{c}_m^J \boldsymbol{\eta}\|_\infty \le \widehat{\sigma}_{n,J} h_{J,m}\}, \tag{2.102}$$

where $h_{J,m}$ does not depend on $\boldsymbol{Y}$. We first choose $h_{J,m}$ such that $\Pi_{\widehat{\sigma}_{n,J}}(\widehat{\mathcal{C}}_\infty^J | \boldsymbol{Y}) = 1 - \omega$ with $\omega < 1/2$, where the appropriate order of J will be determined in Theorem 2.12 below. Using this $h_{J,m}$, we then study the frequentist coverage of the following set:

$$\inf_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R} P_0(\|f_0^{(m)} - \boldsymbol{A}_m^{J_n} \boldsymbol{Y} - \boldsymbol{c}_m^{J_n} \boldsymbol{\eta}\|_\infty \le \widehat{\sigma}_{n,J} h_{J,m}). \tag{2.103}$$

Theorem 2.12. *Let $m < \beta < \alpha$. If $J \asymp (n/\log n)^{1/(2\beta+1)}$, the asymptotic coverage of* (2.103) *is* 1*, and the radius of $\widehat{\mathcal{C}}_\infty^J$ is $O_{P_0}((\log n/n)^{(\beta-m)/(2\beta+1)})$, uniformly over $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R$ for $R > 0$ and any $m < \beta$.*

Proof. Define $t_{J,m} = \inf_{|\sigma^2 - \sigma_0^2| \le \zeta_n/\sqrt{n}} \sigma h_{J,m}$. Following the same argument as in Theorem 2.11, we see that $t_{J,m}^2 \gtrsim (J^{2m+1}/n) \log n$. Using the optimal $J_n \asymp (n/\log n)^{1/(2\alpha+1)}$ to estimate the posterior mean, we see that $\mathrm{E}_0(\|f_0^{(m)} - \boldsymbol{A}_m^{J_n} \boldsymbol{Y} - \boldsymbol{c}_m^{J_n} \boldsymbol{\eta}\|_\infty^2) \lesssim (J_n^{2m+1}/n) \log n + J_n^{-2(\alpha-m)}$ uniformly in $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R$, by substituting J for J_n in (2.100) of Theorem 2.11. By Markov's inequality,

$$\sup_{\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R} P_0(\|f_0^{(m)} - \boldsymbol{A}_m^{J_n} \boldsymbol{Y} - \boldsymbol{c}_m^{J_n} \boldsymbol{\eta}\|_\infty > t_{J,m}) \lesssim \frac{(J_n^{2m+1}/n) \log n + J_n^{-2(\alpha-m)}}{(J^{2m+1}/n) \log n}. \tag{2.104}$$

Since $J \gg J_n$, we have $(J^{2m+1}/n) \log n \gg (J_n^{2m+1}/n) \log n$ and $(J^{2m+1}/n) \log n \gg J_n^{-2(\alpha-m)}$. These then imply the right hand side above will go to 0 as $n \to \infty$. We conclude that $P_0(\|f_0^{(m)} - \boldsymbol{A}_m^{J_n} \boldsymbol{Y} - \boldsymbol{c}_m^{J_n} \boldsymbol{\eta}\|_\infty \le t_{J,m}) \to 1$ for $J \asymp (n/\log n)^{1/(2\beta+1)}$ uniformly over $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R$. As in the proof of Theorem 2.11, it follows from (3.5) of Ledoux and Talagrand (1991) that $h_{J,m} \lesssim (\mathrm{E}\|Z_{n,m}^J\|_\infty^2)^{1/2} \sqrt{-\log \omega}$ for $Z_{n,m}^J \sim \mathrm{GP}(0, \Sigma_m^J)$. By (2.76), we have $\mathrm{E}\|Z_{n,m}^J\|_\infty^2 \lesssim (J^{2m+1}/n) \log n$. In view of (2.46), we have uniformly over $\|f_0^{(m)}\|_{\mathcal{H}^{\alpha-m}} \le R$ that the radius of $\widehat{\mathcal{C}}_\infty^J$ is $\widehat{\sigma}_{n,J} h_{J,m} = O_{P_0}((\log n/n)^{(\beta-m)/(2\beta+1)})$. □

Remark If we had used the optimal J_n to estimate both the radius and the posterior mean, then the right hand side of (2.104) does not approach 0, and hence the need for two different priors. This is in contrast to the undersmoothing method described in Knapik et al. (2011) where they considered L_2-credible region. In Theorem 4.2 of that paper, one prior is sufficient to do undersmoothing. The main difference is that unlike in the L_2 case, the expectation and

standard deviation of $\|Z^J_{n,m}\|_\infty$ has the same order of magnitude, and their argument using asymptotically completely separated distributions (see page 22 of the same paper) does not apply here.

Theorem 2.11 uses the optimal number of B-spline basis to estimate both the center and radius but inflates the width of the credible region by some factor to ensure frequentist coverage of 1 asymptotically. Theorem 2.12 achieves the same aim by using the optimal number to estimate the center but uses a larger J to estimate the radius. In terms of implementation, Theorem 2.11 enjoys two advantages over Theorem 2.12. The first advantage is that we do not have to use two priors, one with optimal J_n and the other with larger J for point estimation and credible region construction respectively. Since the inflation factor ρ_n does not depend on α, which is the underlying smoothness of f_0, it can be used in the adaptive context, i.e, in the case of unknown α and we put a prior on J. On the other hand, we cannot undersmooth if we use an adaptive prior on J, since adaptive prior by definition smooths the right amount.

In view of Theorems 2.9–2.12, we see that there are three different types of credible region constructions that will guarantee frequentist coverage of 1 asymptotically. The first is to let the credible level $1-\omega_n$ approach one as $n \to \infty$, and it was used to construct pointwise and L_2-credible regions (Theorems 2.9 and 2.10). If ω_n does not depend on n, then one could either inflate the radius by some slowly varying factor ρ_n (Theorem 2.11), or do undersmoothing by using a larger J in determining the radius (Theorem 2.12).

2.6 Simulation

In this section, we construct pointwise credible intervals and credible bands based on results established in the previous section. In addition, we will compare finite sample performance of our credible intervals and bands with the confidence intervals and bands proposed by Zhou et al. (1998). Consider the following true function,

$$f_0(x) = \sqrt{2}\sum_{i=1}^{\infty} i^{-3/2}\sin(i)\cos\{(i-1/2)\pi x\}, \tag{2.105}$$

where $x \in [0,1]$. This function is taken from Knapik et al. (2011), and they showed that it has smoothness of $\alpha = 1$ in the Sobolev sense. Figure 2.2 shows a plot of this function. We assume that the data are generated using

$$Y_i = f_0(X_i) + \sigma_0 Z_i, \quad i = 1, \ldots, n, \tag{2.106}$$

where $Z_i \overset{i.i.d.}{\sim} \mathrm{N}(0,1)$ and we set $\sigma_0^2 = 0.1$. The covariates follow a discrete uniform design, i.e., $X_i = (i-1)/(n-1)$ for $i = 1, \ldots, n$. We use B-splines of order 4 (cubic) and uniform knot sequence where we added 4 duplicate knots at each of the boundary points $\{0, 1\}$. For the prior parameters, we set $\boldsymbol{\eta} = \mathbf{0}$ and $\boldsymbol{\Omega} = \boldsymbol{I}_J$, where J is determined by cross-validation (to be discussed below). Using the methods discussed in Section 2.5, we construct $1 - \omega_n$ pointwise credible intervals for f with $m = 0$ where $\omega_n = 5/n$. The constant 5 is chosen such that ω_n at $n = 100$ is 0.05. Here, $\widehat{\sigma}_n^2$ is computed using (2.14). However, direct computation is expensive since the matrix in the middle $\boldsymbol{B\Omega B}^T + \boldsymbol{I}_n$ is $n \times n$ and $n \gg J$. Instead, we use (2.25) and rewrite $\widehat{\sigma}_n^2$ as

$$\widehat{\sigma}_n^2 = \frac{(\boldsymbol{Y} - \boldsymbol{B\eta})^T(\boldsymbol{Y} - \boldsymbol{B\eta}) - (\boldsymbol{Y} - \boldsymbol{B\eta})\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1})^{-1}\boldsymbol{B}^T(\boldsymbol{Y} - \boldsymbol{B\eta})}{n}. \tag{2.107}$$

This then results in huge reduction in the number of flops, since now $\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}$ is $J \times J$ and $J \ll n$. Moreover, this matrix is 4-banded by our choice of prior, $\boldsymbol{\Omega} = \boldsymbol{I}_J$.

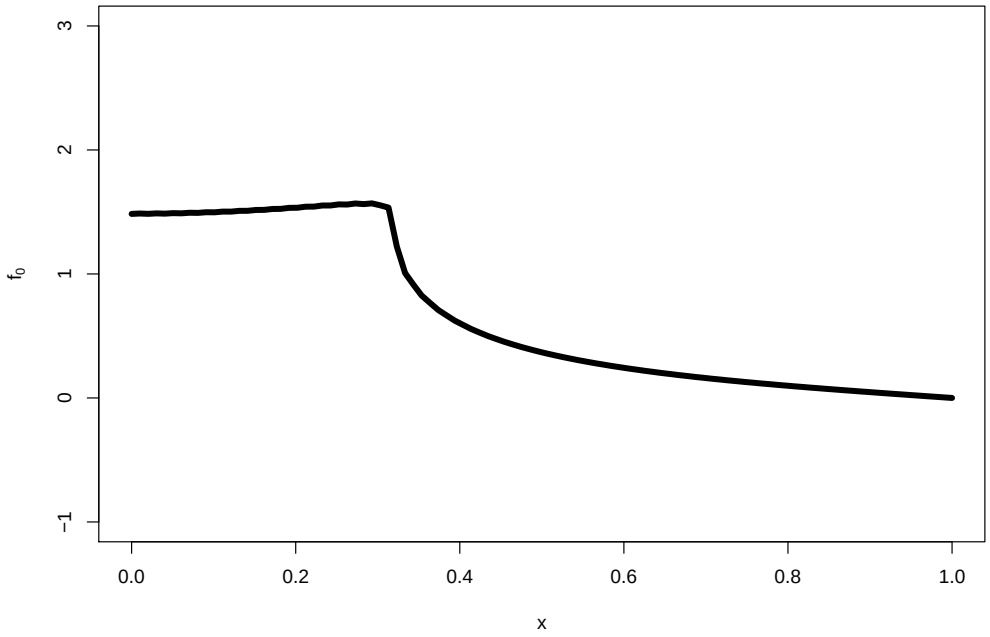


Figure 2.2: A plot of f_0, note the distinguishing bump near $x = 0.3$

For comparison, we will use methods proposed by Zhou et al. (1998) to construct $1 - \omega_n$ pointwise confidence intervals for f. Let $\widehat{f}(x) = \boldsymbol{b}_{J,q}(x)^T\widehat{\boldsymbol{\theta}}$ be the least squares estimator for f where $\widehat{\boldsymbol{\theta}} = (\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{Y}$. Their asymptotic $1 - \omega_n$ pointwise confidence intervals are

constructed as follows (see Theorem 4.1 of the same paper),

$$\widehat{f}(x) \pm z_{\omega_n/2}\sqrt{\mathrm{Var}(\widehat{f}(x))}, \tag{2.108}$$

where $\mathrm{Var}(\widehat{f}(x)) = \sigma^2 \boldsymbol{b}_{J,q}(x)^T(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{b}_{J,q}(x)$. Note that since $z_{\omega_n/2} \to \infty$ as $\omega_n \to 0$, undersmoothing is not needed and $J \asymp n^{1/(2\alpha+1)}$ will be enough to ensure asymptotic coverage of 1. However, since σ^2 is unknown, we use $\widetilde{\sigma}_n^2 = (\boldsymbol{Y}-\boldsymbol{B}\widehat{\boldsymbol{\theta}})^T(\boldsymbol{Y}-\boldsymbol{B}\widehat{\boldsymbol{\theta}})/(n-J)$ as plug-in estimator.

As we will see below in our simulation (see Figures 2.5–2.6), there is a point that has pointwise coverage of 0 for both Bayesian and frequentist methods, in all sample sizes considered. Consequently, we need a better way to construct credible and confidence intervals, which will ensure adequate coverage for all points under consideration. One way to accomplish this is to construct simultaneous credible and confidence bands. For our proposed method, we construct $1-\omega$ credible band as

$$(\boldsymbol{A}(x)\boldsymbol{Y} + \boldsymbol{c}(x)\boldsymbol{\eta}) \pm \rho_n \widehat{\sigma}_n h_{\infty,0} \tag{2.109}$$

in view of (2.93). Note that here ω does not depend on n. To actually determine $h_{\infty,0}$ in (2.109), we draw 1000 samples $\widetilde{f} \sim \Pi_{\widehat{\sigma}_n}(f|\boldsymbol{Y})$ and compute $\widetilde{Z}_\infty = \|\widetilde{f} - \boldsymbol{A}\boldsymbol{Y} - \boldsymbol{c}\boldsymbol{\eta}\|_\infty/\widehat{\sigma}_n$ for each drawn sample. Then, $h_{\infty,0}$ is estimated using the corresponding $1-\omega$ empirical quantile of $\widetilde{Z}_\infty$. To construct $1-\omega$ asymptotic confidence band, we use Theorem 4.2 of Zhou et al. (1998). That is,

$$\widehat{f}(x) \pm c(\omega)\sqrt{\mathrm{Var}(\widehat{f}(x))}, \tag{2.110}$$

where $c(\omega)$ solves the following equation

$$\frac{\omega}{2} = \frac{|\gamma|}{2\pi}\exp(-c(\omega)^2/2) + 1 - \Phi(c(\omega)).$$

The expression for $|\gamma|$ is given by (see equation (41) of Zhou et al., 1998)

$$|\gamma| = \int_0^1 \frac{(\boldsymbol{b}_{J,q}(x)^T(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{H}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{H}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{b}_{J,q}(x))^{1/2}}{(\boldsymbol{b}_{J,q}(x)^T(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{b}_{J,q}(x))^{3/2}}dx, \tag{2.111}$$

where $\boldsymbol{H} = \boldsymbol{W}_1^T\boldsymbol{b}_{J,q,1}(x)\boldsymbol{b}_{J,q}(x)^T - \boldsymbol{b}_{J,q}(x)\boldsymbol{b}_{J,q,1}(x)^T\boldsymbol{W}_1$. Hence, the confidence bands of Zhou et al. (1998) have varying length at each data point while our credible bands have fixed length. To determine the optimal J for both methods, we use leave-one-out cross validation (CV) based on the posterior mean $\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}\boldsymbol{\eta}$ and the least squares estimator $\widehat{f}$ respectively, with prediction mean square error as our criterion function.

We conduct our experiment across sample sizes $n = 100, 300, 500, 700, 1000, 2000$. For pointwise credible and confidence intervals, we report the empirical coverage based on 1000 Monte Carlo runs for each sample size. All computations were carried out in `R` and B-splines were constructed using the `bs` function from the `splines` package.

Figures 2.3 and 2.4 show the optimal N (vertical dash lines) chosen based on cross-validation for each level of sample size, where N is the number of interior knot points. The plots on the left panel (Bayes) use the posterior mean $\boldsymbol{AY} + \boldsymbol{c\eta}$ as estimator of f, while those on the right (Frequentist) use the least squares estimator $\widehat{f}$. We see that apart from the case where $n = 100$, both Bayesian and frequentist methods choose the same optimal N. Also, we notice that the plots for both methods become more and more similar as n increases. Hence, for the rest of the simulation study, the same N or equivalently J will suffice for both methods. Table 2.1 shows the growth rate of J selected by CV as n increases. Observe that there is a close agreement between the optimal J by CV with those computed using the formula $Cn^{1/(2\alpha+1)}$ for $\alpha = 1$, and $C > 0$ is a universal constant not depending on n. By trial and error, we found that $C = 1.65$ for the true function f_0 we considered. This formula is useful for predicting the optimal J for larger sample sizes. For example, one should use approximately 29 B-splines for $n = 5000$.

Table 2.1: Optimal J by CV (Bayes) vs. Computed J based on asymptotic formula with $\alpha = 1$.

n	**100**	**300**	**500**	**700**	**1000**	**2000**
J	8	10	13	14	17	20
$1.65n^{1/(2\alpha+1)}$	7.66	11.05	13.1	14.65	16.5	20.79

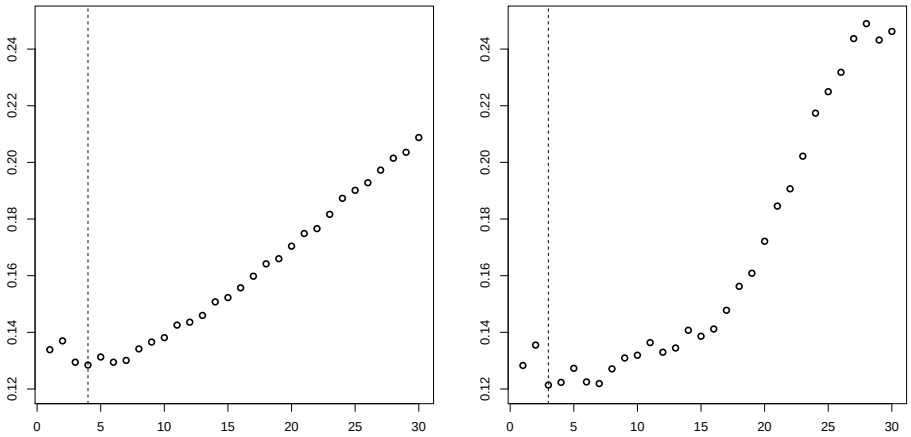
The pointwise coverage probabilities of the credible and confidence intervals for different n are shown in Figures 2.5 and 2.6. One distinguishing feature of these plots is the downward spike at around $x = 0.3$, which corresponds to the bump in the plot of the true function f_0 in Figure 2.2. We see that the plots gradually narrow down to this point as n increases. Moreover, the pointwise coverage is 0 at this point for both Bayesian and frequentist methods in all sample sizes considered. Other than this point, both methods yield almost the same pointwise coverage, especially in large samples sizes, i.e., $n = 700$ and above. We conclude that for high credibility and confidence levels, credible and confidence regions are equivalent in quantifying uncertainty of estimating f_0.

Table 2.2 shows the coverage of 95% ($\omega = 0.05$) simultaneous credible and confidence bands. By trial and error, we found that $\rho_n = (\log\log n)^{1/3}$ will cover the bump across all n. As a result,

our method has coverage of 1 for all sample sizes considered. At $n = 100$, the apparent higher coverage for the confidence bands is due to the positive bias of $\widetilde{\sigma}_n^2$ at small n, which increases the coverage probability. From $n = 300$ onward, the coverage of confidence bands increases with n. The corresponding graphical representations of these bands are shown in Figures 2.7 and 2.8. The left panel corresponds to our proposed Bayesian method, with the red line as the posterior mean and blue dash lines as our 95% credible bands. The right panel corresponds to the frequentist method of Zhou et al. (1998), with the red line as the least squares estimator $\widehat{f}$ and the blue dash lines as the 95% confidence bands. In both panels, the black line is the true function f_0. Except at the endpoints $\{0, 1\}$, we see that our credible band has larger width than the confidence band across all n, but the width of our credible band is decreasing with increasing n, a phenomenon similarly observed for confidence bands. At the endpoints, the frequentist method has much larger width due to the fact that there are fewer observations at these points. Note that our proposed credible bands cover the bump in f_0 for all sample sizes considered, while the frequentist method still misses the bump slightly, especially for $n = 100$ and sample sizes $n = 700$ and above.

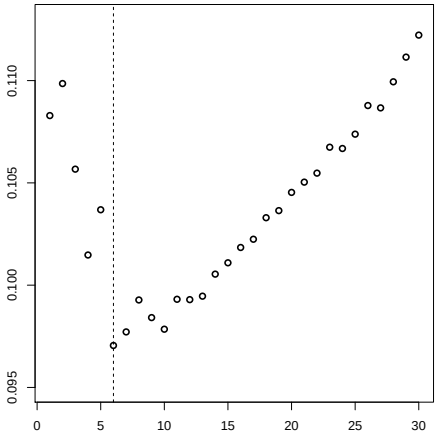
Table 2.2: 95% simultaneous credible and confidence bands.

n	**100**	**300**	**500**	**700**	**1000**	**2000**
Credible band coverage	1	1	1	1	1	1
Confidence band coverage	0.972	0.948	0.963	0.978	0.985	0.986

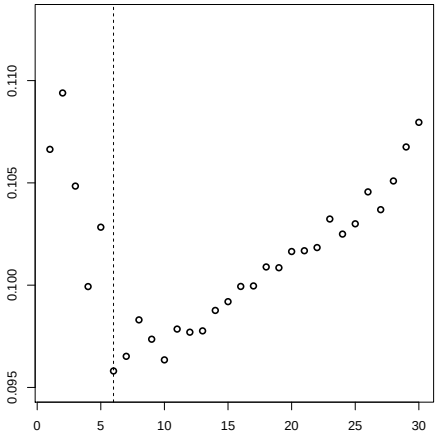


(a) Bayes: $n = 100$, $N = 4$

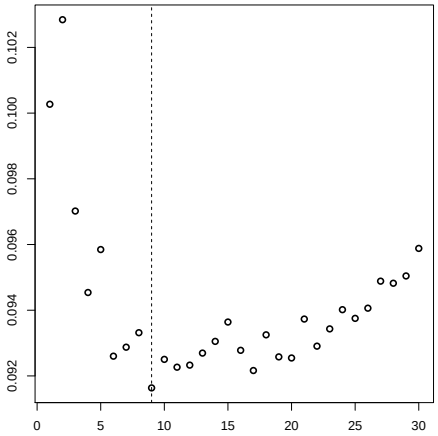
(b) Frequentist: $n = 100$, $N = 3$



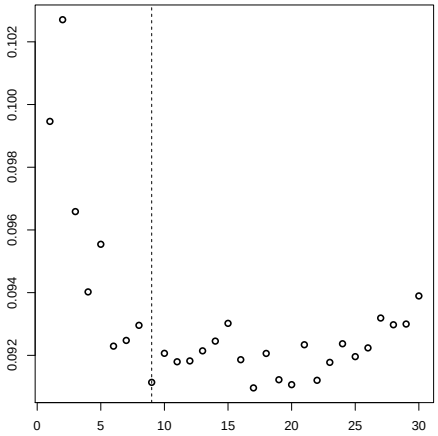
(c) Bayes: $n = 300$, $N = 6$



(d) Frequentist: $n = 300$, $N = 6$

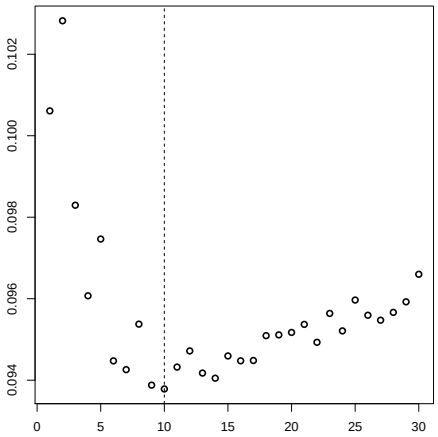


(e) Bayes: $n = 500$, $N = 9$

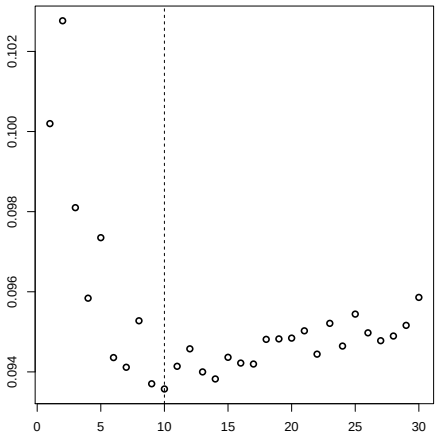


(f) Frequentist: $n = 500$, $N = 9$

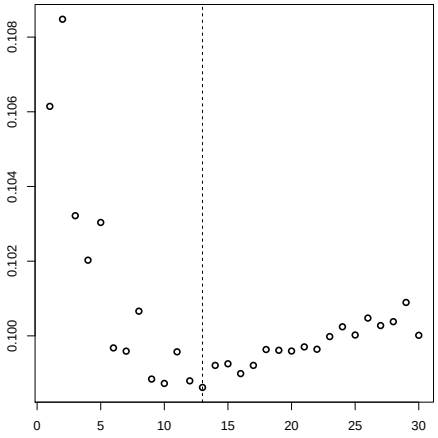
Figure 2.3: CV mean square error against N number of interior knots. Vertical dash lines indicate optimal N.



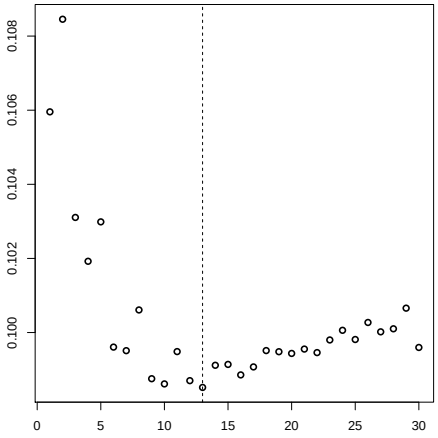
(a) Bayes: $n = 700$, $N = 10$



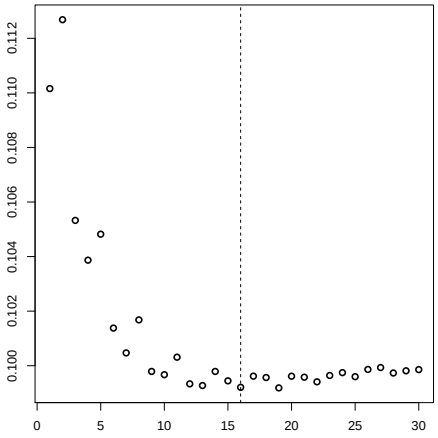
(b) Frequentist: $n = 700$, $N = 10$



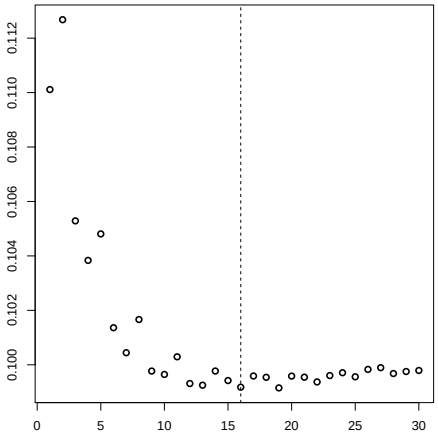
(c) Bayes: $n = 1000$, $N = 13$



(d) Frequentist: $n = 1000$, $N = 13$

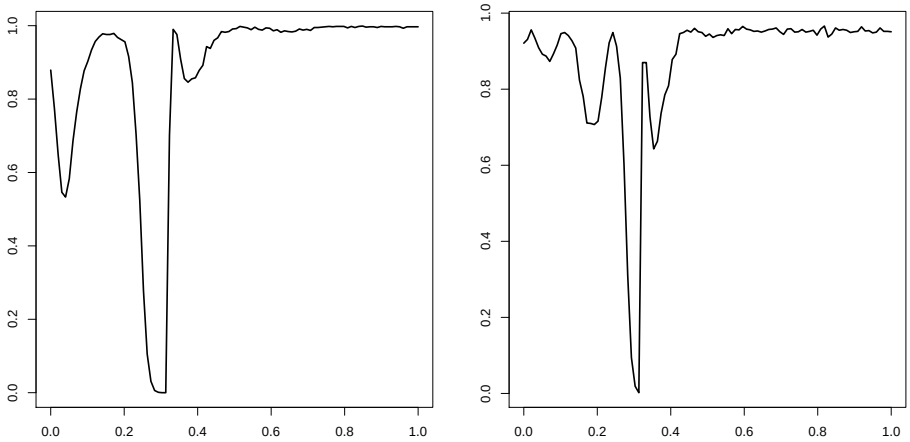


(e) Bayes: $n = 2000$, $N = 16$



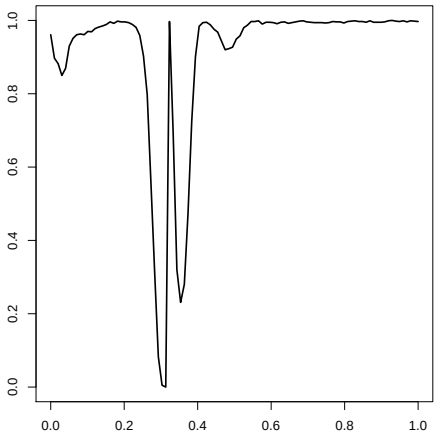
(f) Frequentist: $n = 2000$, $N = 16$

Figure 2.4: CV mean square error against N number of interior knots. Vertical dash lines indicate optimal N.

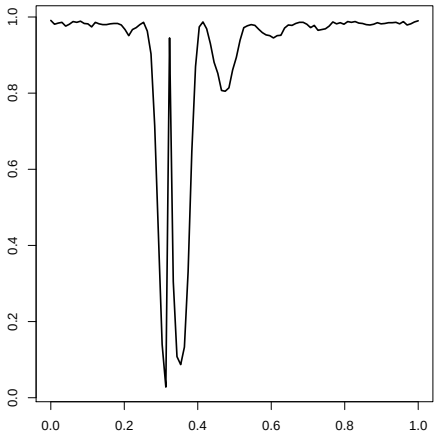


(a) $n = 100$, credibility $= 0.95$

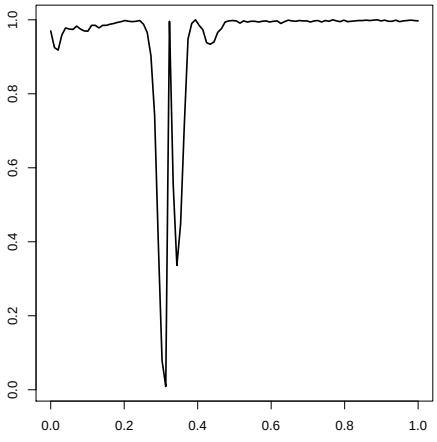
(b) $n = 100$, nominal $= 0.95$



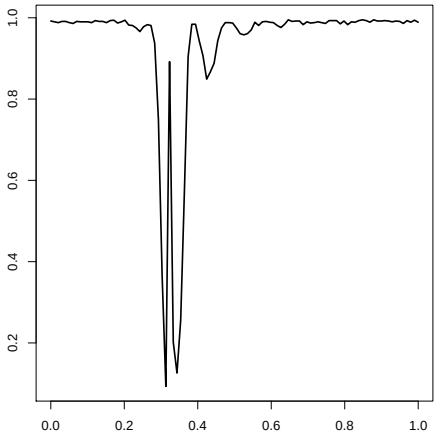
(c) $n = 300$, credibility $= 0.9833$



(d) $n = 300$, nominal $= 0.9833$

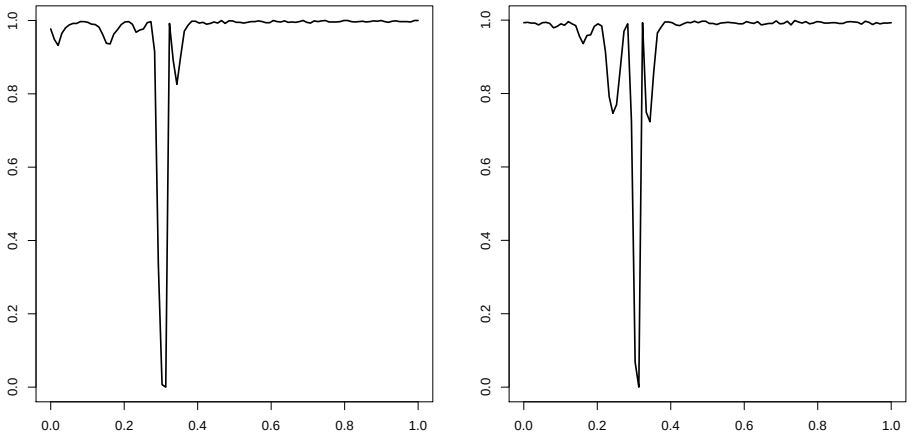


(e) $n = 500$, credibility $= 0.99$

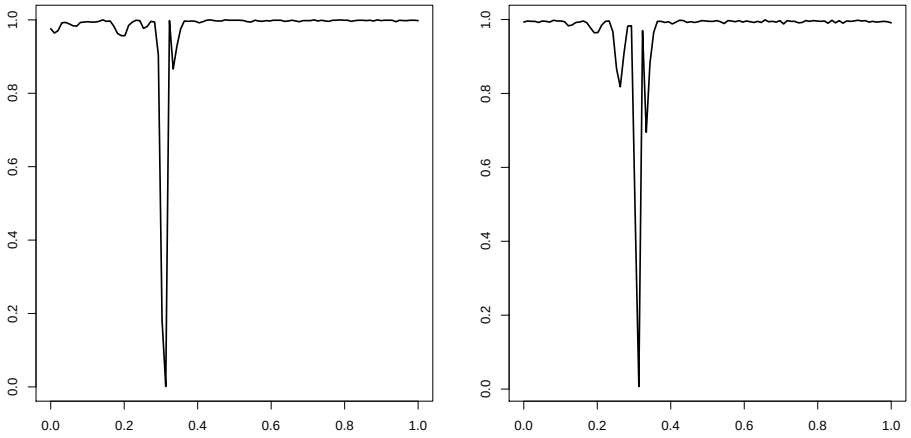


(f) $n = 500$, nominal $= 0.99$

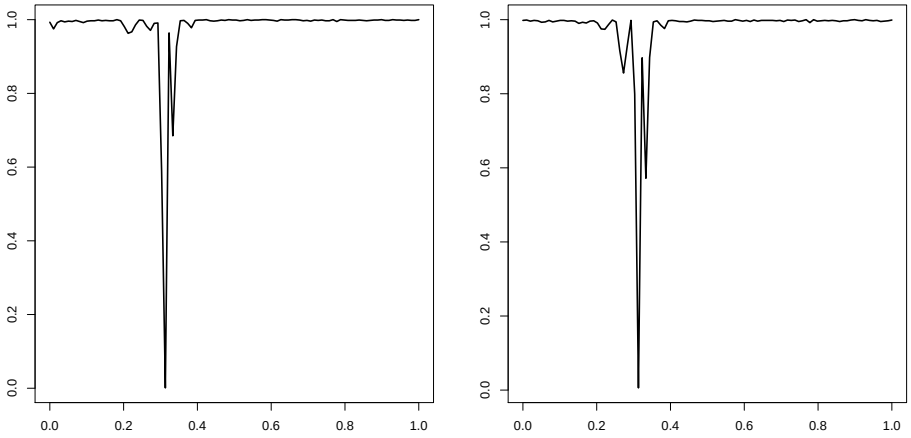
Figure 2.5: Pointwise coverage probabilities for credible and confidence intervals.



(a) $n = 700$, credibility $= 0.9929$ (b) $n = 700$, nominal $= 0.9929$

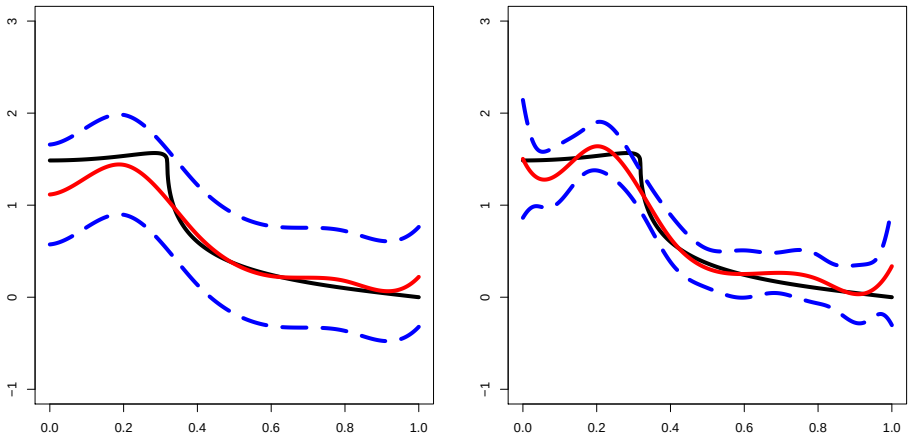


(c) $n = 1000$, credibility $= 0.995$ (d) $n = 1000$, nominal $= 0.995$



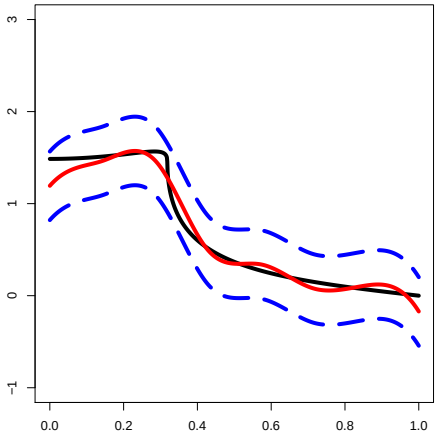
(e) $n = 2000$, credibility $= 0.9975$ (f) $n = 2000$, nominal $= 0.9975$

Figure 2.6: Pointwise coverage probabilities for credible and confidence intervals.

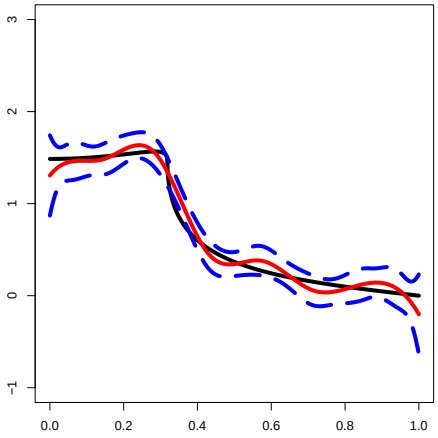


(a) Bayes: $n = 100$

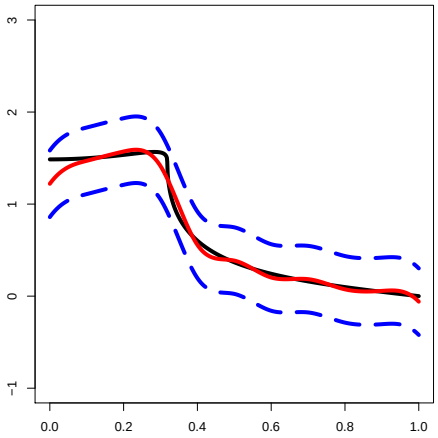
(b) Frequentist: $n = 100$



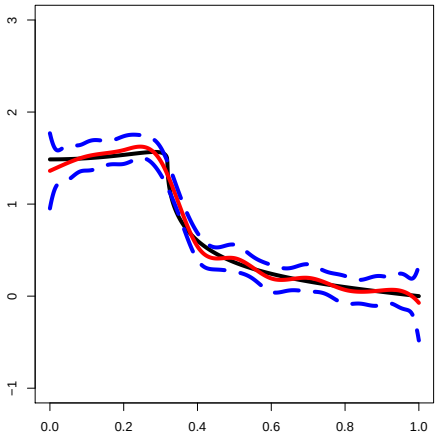
(c) Bayes: $n = 300$



(d) Frequentist: $n = 300$

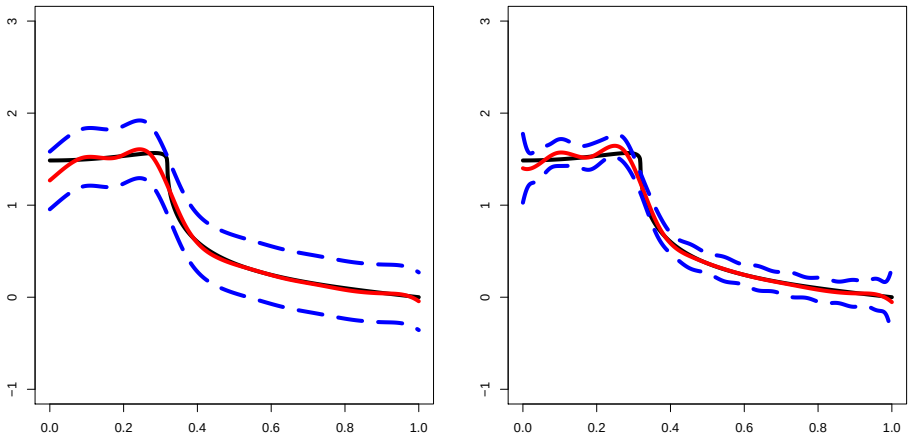


(e) Bayes: $n = 500$



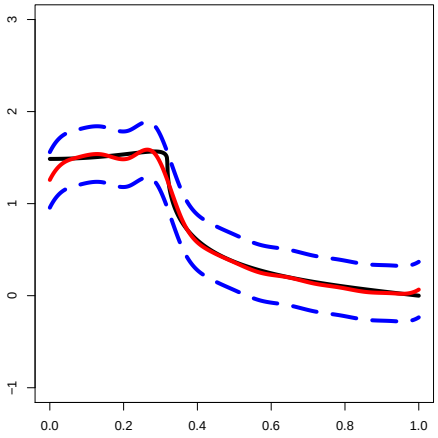
(f) Frequentist: $n = 500$

Figure 2.7: Red: posterior mean (left) and $\widehat{f}$ (right), Black: true function, Blue dash: 95% L_∞-credible (left) and confidence (right) bands.

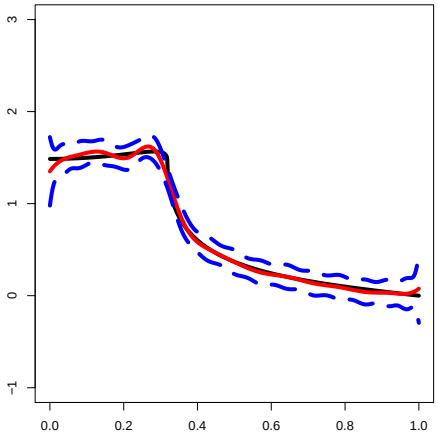


(a) Bayes: $n = 700$

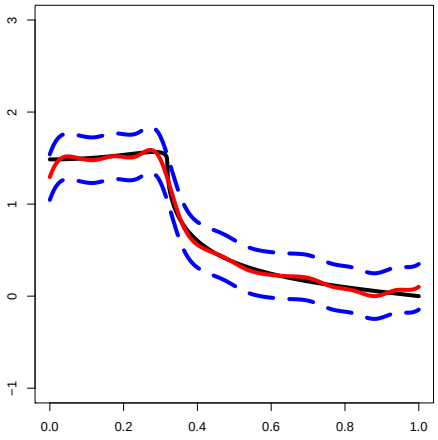
(b) Frequentist: $n = 700$



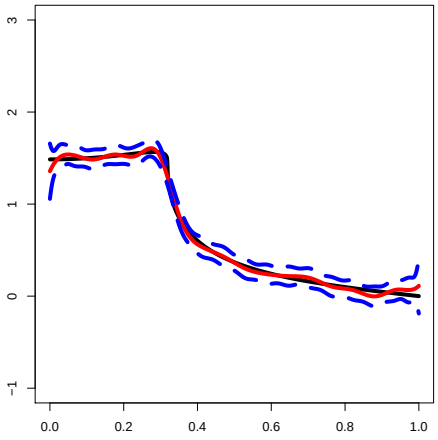
(c) Bayes: $n = 1000$



(d) Frequentist: $n = 1000$



(e) Bayes: $n = 2000$



(f) Frequentist: $n = 2000$

Figure 2.8: Red: posterior mean (left) and $\widehat{f}$ (right), Black: true function, Blue dash: 95% L_∞-credible (left) and confidence (right) bands.

Chapter 3

Sup-norm posterior convergence rate for multivariate regression

3.1 Introduction

Consider the nonparametric regression problem in (1.1), such that for some $f : U \subseteq \mathbb{R}^d \to \mathbb{R}$, we have

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i, \tag{3.1}$$

where ε_i are independent and identically distributed (i.i.d.) as $\mathrm{N}(0, \sigma^2)$ and $0 < \sigma < \infty$ is unknown. The covariates $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^T$ for $i = 1, \ldots, n$ can be deterministic or randomly distributed. In both cases, $\boldsymbol{X}_i \in U$ for $i = 1, \ldots, n$, where U is some open rectangular region in $\mathbb{R}^d$. Since any such regions is an affine transformation of the unit cube, we can let $U = (0, 1)^d$ without loss of generality.

This chapter is an extension of the previous chapter to higher dimensions. In the present chapter, we will estimate f and its mixed partial derivatives using Bayesian procedures, by expanding f using tensor-product B-splines and assigning normal priors on the basis coefficients. To deal with unknown σ^2, we consider two approaches: by estimating σ^2 using empirical Bayes, or further endowing σ^2 with a conjugate inverse-gamma prior. Conjugacy with the model (3.1) above then enables explicit expression for the posterior distribution to be derived and analyzed. We will study pointwise, L_2-and L_∞-posterior convergence rates for f and its mixed partial derivatives. Moreover, pointwise, L_2-and L_∞-credible sets will be constructed to quantify their estimation uncertainties. Under appropriate conditions, we show that they have guaranteed frequentist coverage with optimal size up to a logarithmic factor. Lastly, we will investigate the

issue of anisotropy, i.e., functions with different smoothness in different dimensions, and study its effect on posterior convergence rates and credible sets construction.

Since tensor-product B-splines are used extensively in this chapter, we will describe them briefly and discuss some of their important properties. Suppose we have d system of order q_k univariate B-splines with knot sequence $0 = t_{k,0} < t_{k,1} < \cdots < t_{k,N_k} < t_{k,N_k+1} = 1$, at each dimension $k = 1, \ldots, d$. That is, $\{B_{j_k,q_k}(\cdot) : j_k = 1, \ldots, J_k, k = 1, \ldots, d\}$ for $J_k = q_k + N_k$. We construct tensor-product B-splines at $\boldsymbol{x} = (x_1, \ldots, x_d)^T$ as

$$B_{j_1,\ldots,j_d}(x_1, \ldots, x_d) = \prod_{k=1}^{d} B_{j_k,q_k}(x_k).$$

Figure 3.1 shows 100 bivariate cubic B-splines, where we plotted only the last 50 to show its inner structure. Since tensor-product B-splines are products of univariate B-splines, its properties can be derived from those listed in Section 2.1 of the previous chapter. Below is a list of the most important properties of tensor-product B-splines used in this dissertation (see Chapter 12 Section 2 of Schumaker, 2007 for proofs of these properties). For new results on tensor-product B-splines that we developed for this dissertation, see Appendix B.

1. $0 \leq B_{j_1,\ldots,j_d}(\boldsymbol{x}) \leq 1$ for all $\boldsymbol{x} \in (0,1)^d$ and $1 \leq j_k \leq J_k$, $k = 1, \ldots, d$.
2. The support of $B_{j_1,\ldots,j_d}(\boldsymbol{x})$ is $\prod_{k=1}^{d}(t_{k,j_k-q_k}, t_{k,j_k})$. That is, we have $B_{j_1,\ldots,j_d}(\boldsymbol{x}) > 0$ on $\boldsymbol{x} \in \prod_{k=1}^{d}(t_{k,j_k-q_k}, t_{k,j_k})$ and is zero for $\boldsymbol{x} \notin \prod_{k=1}^{d}[t_{k,j_k-q_k}, t_{k,j_k}]$.
3. For a given $x_k \in [t_{k,l-1}, t_{k,l}]$, only q_k adjacent B-splines $(B_{l,q_k}(x_k), \ldots, B_{l+q_k-1,q_k}(x_k))^T$ are nonzero (positive) for $k = 1, \ldots, d$.
4. Partition of unity: $\sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} B_{j_1,\ldots,j_d}(\boldsymbol{x}) = 1$ for any $\boldsymbol{x} \in (0,1)^d$.

Unless otherwise noted, we now describe the assumption on f_0 that will be used throughout this chapter.

Assumption 2. Under the true distribution P_0, we assume $Y_i = f_0(\boldsymbol{X}_i) + \varepsilon_i$, such that ε_i are i.i.d. with mean 0, variance σ_0^2 and $\|\varepsilon_i\|_{\psi_2} < \infty$ for $i = 1, \ldots, n$. Also, f_0 belongs to the isotropic Hölder space $\mathcal{H}^\alpha((0,1)^d)$ of order $\alpha > \max\{d/2, 2\}$. To avoid boundary effects, f_0 can be extended to a slightly bigger set $(-\epsilon, 1+\epsilon)^d$ for some $\epsilon > 0$.

Let p_0 be the density of P_0 with respect to the Lebesgue measure. Let $\mathrm{E}_0(\cdot)$ and $\mathrm{Var}_0(\cdot)$ be the expectation and variance operators taken with respect to P_0 respectively. For notational conciseness, write $\boldsymbol{F}_0 = (f_0(\boldsymbol{X}_1), \ldots, f_0(\boldsymbol{X}_n))^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$.

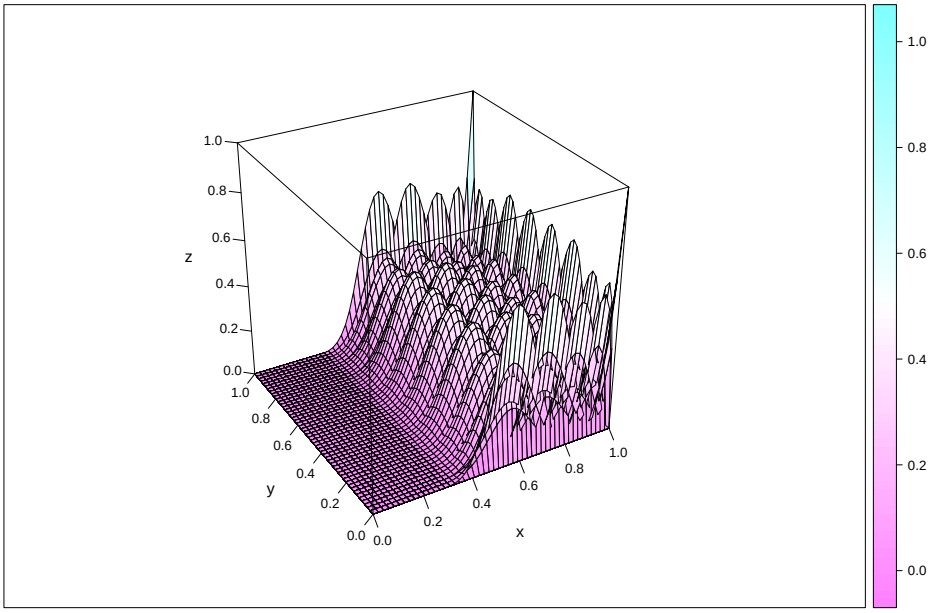


Figure 3.1: A set of 100 bivariate cubic B-splines. To show the inner structure, only the last 50 is plotted.

3.2 Prior and posterior conjugacy

For any $\boldsymbol{x} = (x_1, \ldots, x_d)^T$, we represent $f(\boldsymbol{x})$ as

$$f(\boldsymbol{x}) = \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1,\ldots,j_d} B_{j_1,q_1}(x_1) \cdots B_{j_d,q_d}(x_d) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})^T \boldsymbol{\theta}, \tag{3.2}$$

where $\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x}) = \{B_{j_1,q_1}(x_1) \cdots B_{j_d,q_d}(x_d)\}_{j_1=1,\ldots,j_d=1}^{J_1,\ldots,J_d}$ is a vector of $J = \prod_{k=1}^d J_k$ tensor-product B-splines, and $\boldsymbol{\theta} = \{\theta_{j_1,\ldots,j_d}\}_{j_1=1,\ldots,j_d=1}^{J_1,\ldots,J_d}$ are the corresponding basis coefficients. Each element in $\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})$ is a product of d univariate B-splines of fixed orders $\boldsymbol{q} = (q_1, \ldots, q_d)^T$, such that $q_k \geq \alpha$ for $k = 1, \ldots, d$, with number of basis $\boldsymbol{J} = (J_1, \ldots, J_d)^T$ for each B-spline component. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{X} = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T)^T$ and $\boldsymbol{B} = (\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{X}_1), \ldots, \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{X}_n))^T$. Then the model for (3.1) can be written compactly as

$$\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{\theta}, \sigma^2 \sim \mathrm{N}_n(\boldsymbol{B}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}_n). \tag{3.3}$$

In this chapter, we treat each component of $\boldsymbol{J}$ as deterministic and allow it to depend on n, d and α. At each $k = 1, \ldots, d$, the kth B-spline system has knot sequence $0 = t_{k,-q_k} = t_{k,1-q_k} = \cdots = t_{k,0} < t_{k,1} < \cdots < t_{k,N_k} < t_{k,N_k+1} = t_{k,N_k+2} = \cdots = t_{k,N_k+q} = 1$. The q_k duplicate knots

at each boundary point $\{0,1\}$ are needed, because the recursive formulas for B-spline and its derivatives (see equations (4.22) and (4.23) of Schumaker, 2007) entail using knot points with index exceeding N_k or have negative index. We note that the knot points and the number of interior knots N_k may depend on n. We then have $J_k = q_k + N_k$, where N_k and thus J_k are increasing with n. Moreover, we assume that $\prod_{k=1}^{d} J_k \leq n$.

At each $k = 1, \ldots, d$, define $\delta_{k,l} = t_{k,l} - t_{k,l-1}$ to be the one-step knot increment, and let $\Delta_k = \max_{1 \leq l \leq N_k} \delta_{k,l}$ be the mesh size. We assume that the knot sequence for each direction is quasi-uniform (see Definition 6.4 of Schumaker, 2007), that is,

$$\frac{\Delta_k}{\min_{1 \leq l \leq N_k} \delta_{k,l}} \leq C \tag{3.4}$$

for some $C > 0$. This assumption is reasonable in view of Lemma 6.17 from Schumaker (2007) which says that we can always choose a subset of knots from any given knot sequence to form a quasi-uniform sequence with $C = 3$. In addition, this class of partition is general enough for most applications and includes the uniform and nested uniform partitions as special cases (see Examples 6.6 and 6.7 of Schumaker, 2007 respectively).

If the design points $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^T$ for $i = 1, \ldots, n$ are deterministic, assume that there exists a cumulative distribution function $G(\boldsymbol{x})$ having support on $[0,1]^d$, with positive and continuous density $g(\boldsymbol{x})$ such that

$$\sup_{\boldsymbol{x} \in [0,1]^d} |G_n(\boldsymbol{x}) - G(\boldsymbol{x})| = o\left(\prod_{k=1}^{d} N_k^{-1} \right), \tag{3.5}$$

where $G_n(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\prod_{k=1}^{d} [X_{ik}, \infty)}(\boldsymbol{x})$ is the empirical distribution of $\{\boldsymbol{X}_i : i = 1, \ldots, n\}$. For example, let $n = m^d$ for some $m \in \mathbb{N}$, the discrete uniform design $\boldsymbol{X}_i \in \{(j-1)/(m-1) : j = 1, \ldots, m\}^d$ for $i = 1, \ldots, n$ satisfies (3.5) with G being the uniform distribution on $[0,1]^d$ and $N_k \lesssim n^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$.

For random design points, we assume that $\boldsymbol{X}_i \overset{i.i.d.}{\sim} G(\boldsymbol{x})$, and they are mutually independent of ε_i for $i = 1, \ldots, n$. Here, $G(\boldsymbol{x}) = P(\boldsymbol{X}_i \leq \boldsymbol{x})$ is a cumulative distribution function with continuous density $g(\boldsymbol{x})$ on $[0,1]^d$. By Donsker's theorem, $\|G_n - G\|_\infty = O_P(1/\sqrt{n})$. If $N_k \lesssim n^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$ and $\alpha > d/2$, then

$$\|G_n - G\|_\infty = O_P\left(\frac{1}{\sqrt{n}} \right) = o_P\left(\prod_{k=1}^{d} N_k^{-1} \right). \tag{3.6}$$

In this dissertation, we will prove results on posterior convergence rates and credible sets based on deterministic design points. The established theorems will then translate directly to the random case by conditioning $\boldsymbol{X}$ in the posterior distribution, i.e., by considering $\Pi(\cdot|\boldsymbol{Y}, \boldsymbol{X})$.

Note that $\boldsymbol{B}^T\boldsymbol{B}$ is indexed by d-dimensional multi-indices. That is, for $\boldsymbol{u} = (u_1, \ldots, u_d)^T$ and $\boldsymbol{v} = (v_1, \ldots, v_d)^T$ with $1 \leq u_k, v_k \leq J_k, k = 1, \ldots, d$, the $(\boldsymbol{u}, \boldsymbol{v})$th entry is $(\boldsymbol{B}^T\boldsymbol{B})_{\boldsymbol{u},\boldsymbol{v}} = \sum_{i=1}^n \prod_{k=1}^d B_{u_k,q_k}(X_{ik})B_{v_k,q_k}(X_{ik})$. Unlike in the one-dimensional case, $\boldsymbol{B}^T\boldsymbol{B}$ is not banded in the usual sense, regardless of any permutations of the columns of $\boldsymbol{B}$. In order to proceed, we extend the concept of banded matrix to general array-indexed matrix in the following definition:

Definition 3.1. (Multidimensional banding). Let $\boldsymbol{A}$ be a $2d$-dimensional square array of order $\boldsymbol{J} \times \boldsymbol{J}$ for $\boldsymbol{J} = (J_1, \ldots, J_d)^T$, such that each element of $\boldsymbol{A}$ is $a_{\boldsymbol{u},\boldsymbol{v}}$ for $\boldsymbol{u} = (u_1, \ldots, u_d)^T$ and $\boldsymbol{v} = (v_1, \ldots, v_d)^T$ with $1 \leq u_k, v_k \leq J_k, k = 1, \ldots, d$. We represent $\boldsymbol{A}$ as a matrix with rows and columns indexed by d-dimensional multi-indices $\boldsymbol{1}_d \leq \boldsymbol{u}, \boldsymbol{v} \leq \boldsymbol{J}$ respectively, where arrangement of the elements can be arbitrary, and the corresponding $(\boldsymbol{u}, \boldsymbol{v})$th element of this matrix denoted as $a_{\boldsymbol{u},\boldsymbol{v}}$. We say that $\boldsymbol{A}$ is $\boldsymbol{h} = (h_1, \ldots, h_d)^T$ banded if $a_{\boldsymbol{u},\boldsymbol{v}} = 0$ when $|u_k - v_k| > h_k$ for some $k = 1, \ldots, d$.

Recall that the $(\boldsymbol{u}, \boldsymbol{v})$th entry of $\boldsymbol{B}^T\boldsymbol{B}$ is $\sum_{i=1}^n \prod_{k=1}^d B_{u_k,q_k}(X_{ik})B_{v_k,q_k}(X_{ik})$. Given some $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^T$ for $i = 1, \ldots, n$ such that $X_{ik} \in [t_{k,l-1}, t_{k,l}]$, only q_k adjacent basis functions $(B_{l,q_k}(X_{ik}), \ldots, B_{l+q_k-1,q_k}(X_{ik}))^T$ will be nonzero for $k = 1, \ldots, d$. Hence if $|u_m - v_m| > q_m$ for some $m = 1, \ldots, d$, then $B_{u_m,q_m}(X_{im})B_{v_m,q_m}(X_{im}) = 0$, and we conclude that $(\boldsymbol{B}^T\boldsymbol{B})_{\boldsymbol{u},\boldsymbol{v}} = 0$. It then follows that $\boldsymbol{B}^T\boldsymbol{B}$ is $\boldsymbol{q}$-banded in view of Definition 3.1.

On the basis coefficients, we assign

$$\boldsymbol{\theta}|\sigma^2 \sim \mathrm{N}_J(\boldsymbol{\eta}, \sigma^2\boldsymbol{\Omega}). \tag{3.7}$$

We assume that $\|\boldsymbol{\eta}\|_\infty < \infty$ and the entries of $\boldsymbol{\Omega}$ do not depend on n. The entries of $\boldsymbol{\Omega}$ are indexed using d-dimensional multi-indices in a similar manner as for $\boldsymbol{B}^T\boldsymbol{B}$ described above. We further assume that $\boldsymbol{\Omega}^{-1}$ is a $\boldsymbol{v} = (v_1, \ldots, v_d)^T$ banded matrix, where each element of $\boldsymbol{v}$ does not depend on n. The bandedness of $\boldsymbol{\Omega}^{-1}$ ensures that the posterior precision matrix is also banded, which allows upper bound on the bias of the posterior mean to be established. We note that $\boldsymbol{\Omega}$ depends on n only through its dimension, which is $J \times J$. Furthermore, as $n \to \infty$, we assume that there exists constants $0 < c_1 \leq c_2 < \infty$ not depending on n such that

$$c_1\boldsymbol{I}_J \leq \boldsymbol{\Omega} \leq c_2\boldsymbol{I}_J. \tag{3.8}$$

We see that a scaled identity matrix satisfies the multidimensional banding assumption and

(3.8) for $\boldsymbol{v} = \mathbf{1}_d$. The prior support of f is given by the closure of its reproducing kernel Hilbert space (van der Vaart and van Zanten, 2008a), which is the J-dimensional space of polynomial splines spanned by elements of $\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\cdot)$. Since $q_k \geq \alpha$, this implies that the prior support of f is a subspace of $\mathcal{H}^\alpha((0,1)^d)$.

Based on (3.3) and (3.7), the conditional posterior of $\boldsymbol{\theta}$ is

$$\Pi(\boldsymbol{\theta}|\boldsymbol{Y}, \sigma^2) \sim \mathrm{N}_J\left[\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\left(\boldsymbol{B}^T\boldsymbol{Y} + \boldsymbol{\Omega}^{-1}\boldsymbol{\eta}\right), \sigma^2\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\right]. \tag{3.9}$$

It then follows that the induced posterior for f is $\Pi(f|\boldsymbol{Y}, \sigma^2) \sim \mathrm{GP}(\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}\boldsymbol{\eta}, \sigma^2\Sigma)$, where $\boldsymbol{A}$ and $\boldsymbol{c}$ are bounded linear operators mapping $\mathbb{R}^n$ and $\mathbb{R}^J$ respectively to $\mathcal{H}^\alpha((0,1)^d)$, and Σ is the covariance function defined on $(0,1)^d \times (0,1)^d$ such that for any $\boldsymbol{x}, \boldsymbol{y} \in (0,1)^d$,

$$\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})^T\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\boldsymbol{B}^T, \tag{3.10}$$

$$\boldsymbol{c}(\boldsymbol{x}) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})^T\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\boldsymbol{\Omega}^{-1}, \tag{3.11}$$

$$\Sigma(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})^T\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{y}). \tag{3.12}$$

Note that since the posterior mean can be written as $\sum_{i=1}^n a_i(\boldsymbol{x})Y_i + c(\boldsymbol{x})\boldsymbol{\eta}$ at a fixed $\boldsymbol{x}$, with $a_i(\boldsymbol{x})$ being the ith element of the row vector $\boldsymbol{A}(\boldsymbol{x})$, Assumption 2 with Properties 1 and 2 of Appendix C imply that $\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}\boldsymbol{\eta}$ is a sub-Gaussian process under P_0.

To deal with unknown σ^2, we will consider two approaches. The first is empirical Bayes estimation. Based on (3.3) and (3.7), it follows that $\boldsymbol{Y}|\sigma^2 \sim \mathrm{N}_n[\boldsymbol{B}\boldsymbol{\eta}, \sigma^2(\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^T + \boldsymbol{I}_n)]$. The corresponding marginal log-likelihood function is

$$\begin{aligned} l(\sigma^2) = &-\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\log\{\det\left(\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^T + \boldsymbol{I}_n\right)\} \\ &- \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\eta})^T(\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^T + \boldsymbol{I}_n)^{-1}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\eta}), \end{aligned}$$

with score function

$$\frac{dl(\sigma^2)}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\eta})^T(\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^T + \boldsymbol{I}_n)^{-1}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\eta}).$$

Setting the above to zero and solving for σ^2, we obtain

$$\widehat{\sigma}_n^2 = \frac{(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\eta})^T(\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^T + \boldsymbol{I}_n)^{-1}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\eta})}{n}. \tag{3.13}$$

Note that $P_0(\widehat{\sigma}_n^2 > 0) = 1$ since $P_0(\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{\eta}) = 0$. By the second order condition,

$$\left.\frac{d^2 l(\sigma^2)}{d(\sigma^2)^2}\right|_{\sigma^2=\widehat{\sigma}_n^2} = -\frac{n}{2\widehat{\sigma}_n^4} < 0.$$

Hence with P_0 almost surely, $\widehat{\sigma}_n^2$ is a local maximum. Since $\widehat{\sigma}_n^2$ is the unique critical point, we conclude that $\widehat{\sigma}_n^2$ is the maximum marginal likelihood estimator for σ^2. Empirical Bayes then entails substituting σ^2 for $\widehat{\sigma}_n^2$ in the conditional posterior of f, i.e.,

$$\Pi(f|\boldsymbol{Y}, \sigma^2)|_{\sigma^2=\widehat{\sigma}_n^2} = \Pi_{\widehat{\sigma}_n}(f|\boldsymbol{Y}) \sim \mathrm{GP}(\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}\boldsymbol{\eta}, \widehat{\sigma}_n^2\Sigma). \tag{3.14}$$

We then write $\mathrm{E}_{\widehat{\sigma}_n}(\cdot|\boldsymbol{Y})$ and $\mathrm{Var}_{\widehat{\sigma}_n}(\cdot|\boldsymbol{Y})$ for the expectation and variance operators taken with respect to this empirical posterior. For the second approach, we will consider full hierarchical Bayes by further endowing σ^2 with a conjugate inverse-gamma (IG) prior

$$\sigma^2 \sim \mathrm{IG}\left(\frac{\beta}{2}, \frac{\gamma}{2}\right) \tag{3.15}$$

for some hyperparameters $\beta > 4$ and $\gamma > 0$. Here, $\beta > 4$ ensures that the prior mean and variance of σ^2 exist. By direct calculations, the posterior of σ^2 is

$$\sigma^2|\boldsymbol{Y} \sim \mathrm{IG}\left(\frac{\beta + n}{2}, \frac{\gamma + n\widehat{\sigma}_n^2}{2}\right). \tag{3.16}$$

Under assumptions (3.4) and (3.5), Lemma B.7 concludes that there exist constants $0 < C_1 \leq C_2 < \infty$ not depending on n such that as $n \to \infty$,

$$C_1 n \left(\prod_{k=1}^{d} J_k^{-1}\right) \boldsymbol{I}_J \leq \boldsymbol{B}^T\boldsymbol{B} \leq C_2 n \left(\prod_{k=1}^{d} J_k^{-1}\right) \boldsymbol{I}_J. \tag{3.17}$$

In particular, $\|\boldsymbol{B}^T\boldsymbol{B}\|_{(2,2)} \asymp n\prod_{k=1}^{d} J_k^{-1}$. Combining the above with (3.8), it follows that

$$\begin{aligned}\left(C_1 n \prod_{k=1}^{d} J_k^{-1} + \frac{1}{c_2}\right) &\leq \lambda_{\min}\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right) \\ &\leq \lambda_{\max}\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right) \leq \left(C_2 n \prod_{k=1}^{d} J_k^{-1} + \frac{1}{c_1}\right).\end{aligned} \tag{3.18}$$

By Theorem 22 in de Boor (2001) Chapter XII, for $f_0 \in \mathcal{H}^{\alpha}((0,1)^d)$, there exists a $\boldsymbol{\theta}_{\infty} \in \mathbb{R}^J$

such that for constant $C_{\alpha,d,\boldsymbol{q}} > 0$ depending only on α, d and $\boldsymbol{q}$ with $R > 0$, we have

$$\sup_{\|f_0\|_{\mathcal{H}^\alpha} \le R} \|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\cdot)^T \boldsymbol{\theta}_\infty - f_0\|_\infty \le \sup_{\|f_0\|_{\mathcal{H}^\alpha} \le R} C_{\alpha,d,\boldsymbol{q}} \sum_{k=1}^{d} J_k^{-\alpha} \|f_0\|_{\mathcal{H}^\alpha} \lesssim \sum_{k=1}^{d} J_k^{-\alpha}. \tag{3.19}$$

Since $\|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\cdot)^T \boldsymbol{\theta}_\infty\|_\infty \le \|f_0\|_{\mathcal{H}^\alpha} + C_{\alpha,d,\boldsymbol{q}} \sum_{k=1}^{d} J_k^{-\alpha} \|f_0\|_{\mathcal{H}^\alpha} \le R + d$ uniformly in $\|f_0\|_{\mathcal{H}^\alpha} \le R$, we have

$$\sup_{\|f_0\|_{\mathcal{H}^\alpha} \le R} \|\boldsymbol{\theta}_\infty\|_\infty \lesssim \sup_{\|f_0\|_{\mathcal{H}^\alpha} \le R} \|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\cdot)^T \boldsymbol{\theta}_\infty\|_\infty = O(1) \tag{3.20}$$

by (12.25) of Theorem 12.5 in Schumaker (2007). The next theorem shows that $\widehat{\sigma}_n^2$ derived in (3.13) is a $\sqrt{n}$-consistent estimator of σ_0^2, uniformly over $\|f_0\|_{\mathcal{H}^\alpha} \le R$.

Theorem 3.1. *If $J_k \asymp n^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$ and $\alpha > d/2$, we have for $R > 0$,*

$$\sup_{\|f_0\|_{\mathcal{H}^\alpha} \le R} \mathrm{E}_0(\widehat{\sigma}_n^2 - \sigma_0^2)^2 = O\left(\frac{1}{n}\right). \tag{3.21}$$

Proof. Let $\boldsymbol{U} = (\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^T + \boldsymbol{I}_n)^{-1}$ and $J = \prod_{k=1}^{d} J_k$. By equation (33) of page 355 in Searle (1982), the absolute bias $|\mathrm{E}_0(\widehat{\sigma}_n^2) - \sigma_0^2|$ is

$$\begin{aligned}
&\left| \frac{\sigma_0^2}{n} \mathrm{tr}(\boldsymbol{U}) - \sigma_0^2 \right| + \frac{1}{n} (\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\eta})^T \boldsymbol{U} (\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\eta}) \\
&\lesssim \frac{1}{n} [\mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{U}) + (\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty)^T \boldsymbol{U} (\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty) + (\boldsymbol{B}\boldsymbol{\theta}_\infty - \boldsymbol{B}\boldsymbol{\eta})^T \boldsymbol{U} (\boldsymbol{B}\boldsymbol{\theta}_\infty - \boldsymbol{B}\boldsymbol{\eta})],
\end{aligned} \tag{3.22}$$

where we used $(\boldsymbol{x} + \boldsymbol{y})^T \boldsymbol{D} (\boldsymbol{x} + \boldsymbol{y}) \le 2\boldsymbol{x}^T \boldsymbol{D} \boldsymbol{x} + 2\boldsymbol{y}^T \boldsymbol{D} \boldsymbol{y}$ for any $\boldsymbol{D} \ge \boldsymbol{0}$ in the last line. Let $\boldsymbol{P}_{\boldsymbol{B}} = \boldsymbol{B}(\boldsymbol{B}^T \boldsymbol{B})^{-1} \boldsymbol{B}^T$. Suppose that $\boldsymbol{A}$ is an $m \times m$ matrix, $\boldsymbol{G}$ an $m \times r$ matrix, $\boldsymbol{T}$ an $r \times r$ matrix, and $\boldsymbol{W}$ an $r \times m$ matrix. Assume that $\boldsymbol{A}$ and $\boldsymbol{T}$ are invertible. Then, the binomial inverse theorem (see Theorem 18.2.8 of Harville, 1997 for a proof) says

$$(\boldsymbol{A} + \boldsymbol{G}\boldsymbol{T}\boldsymbol{W})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1} \boldsymbol{G} (\boldsymbol{T}^{-1} + \boldsymbol{W}\boldsymbol{A}^{-1}\boldsymbol{G})^{-1} \boldsymbol{W}\boldsymbol{A}^{-1}. \tag{3.23}$$

Therefore, two applications of (3.23) to $\boldsymbol{U}$ yield

$$\boldsymbol{U} = (\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^T + \boldsymbol{I}_n)^{-1} = \boldsymbol{I}_n - \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1})^{-1} \boldsymbol{B}^T = \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{B}} + \boldsymbol{V}, \tag{3.24}$$

where $\boldsymbol{V} = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1} [\boldsymbol{\Omega} + (\boldsymbol{B}^T\boldsymbol{B})^{-1}]^{-1} (\boldsymbol{B}^T\boldsymbol{B})^{-1} \boldsymbol{B}^T \ge \boldsymbol{0}$. Hence the first term in (3.22) is

$$\frac{1}{n} \mathrm{tr}(\boldsymbol{P}_{\boldsymbol{B}} - \boldsymbol{V}) \le \frac{1}{n} \mathrm{tr}(\boldsymbol{P}_{\boldsymbol{B}}) = \frac{J}{n}. \tag{3.25}$$

Note that $\boldsymbol{U} \leq \boldsymbol{I}_n$ since $\boldsymbol{B\Omega B}^T \geq \boldsymbol{0}$, and the second term in (3.22) is bounded by

$$\frac{1}{n}\|\boldsymbol{U}\|_{(2,2)}\|\boldsymbol{F}_0 - \boldsymbol{B\theta}_\infty\|^2 \leq \|\boldsymbol{F}_0 - \boldsymbol{B\theta}_\infty\|_\infty^2 \lesssim \sum_{k=1}^{d} J_k^{-2\alpha}, \tag{3.26}$$

in view of (3.19). By (3.24) and the fact that $(\boldsymbol{I} - \boldsymbol{P_B})\boldsymbol{B} = \boldsymbol{0}$, the last term in (3.22) is

$$\frac{1}{n}(\boldsymbol{\theta}_\infty - \boldsymbol{\eta})^T[\boldsymbol{\Omega} + (\boldsymbol{B}^T\boldsymbol{B})^{-1}]^{-1}(\boldsymbol{\theta}_\infty - \boldsymbol{\eta}) \leq \frac{1}{n}\left(c_1 + \frac{1}{C_2}\frac{J}{n}\right)^{-1} J\|\boldsymbol{\theta}_\infty - \boldsymbol{\eta}\|_\infty^2 \lesssim \frac{J}{n}, \tag{3.27}$$

where we used (3.8) and (3.17) to bound the maximum eigenvalue of $[\boldsymbol{\Omega} + (\boldsymbol{B}^T\boldsymbol{B})^{-1}]^{-1}$. By (3.20) and assumption on the prior, $\|\boldsymbol{\theta}_\infty - \boldsymbol{\eta}\|_\infty^2 = O(1)$. Combining the bounds in (3.25), (3.26) and (3.27) into (3.22), we obtain $|\mathrm{E}_0(\widehat{\sigma}_n^2) - \sigma_0^2| \lesssim J/n + \sum_{k=1}^{d} J_k^{-2\alpha}$.

Let $\boldsymbol{Y} = \boldsymbol{F}_0 + \boldsymbol{\varepsilon}$ and write $n\widehat{\sigma}_n^2 = (\boldsymbol{F}_0 - \boldsymbol{B\eta})^T\boldsymbol{U}(\boldsymbol{F}_0 - \boldsymbol{B\eta}) + 2(\boldsymbol{F}_0 - \boldsymbol{B\eta})^T\boldsymbol{U\varepsilon} + \boldsymbol{\varepsilon}^T\boldsymbol{U\varepsilon}$. Using the fact $\mathrm{Var}(T_1 + T_2) \leq 2\mathrm{Var}(T_1) + 2\mathrm{Var}(T_2)$, it follows that $\mathrm{Var}_0(\widehat{\sigma}_n^2)$ is bounded up to a constant multiple by

$$\frac{1}{n^2}[(\boldsymbol{F}_0 - \boldsymbol{B\theta}_\infty)^T\boldsymbol{U}^2(\boldsymbol{F}_0 - \boldsymbol{B\theta}_\infty) + (\boldsymbol{B\theta}_\infty - \boldsymbol{B\eta})^T\boldsymbol{U}^2(\boldsymbol{B\theta}_\infty - \boldsymbol{B\eta}) + \mathrm{Var}_0(\boldsymbol{\varepsilon}^T\boldsymbol{U\varepsilon})]. \tag{3.28}$$

In view of (3.19) and $\boldsymbol{U} \leq \boldsymbol{I}_n$, the first term above is bounded by

$$\frac{1}{n^2}\|\boldsymbol{U}\|_{(2,2)}^2\|\boldsymbol{F}_0 - \boldsymbol{B\theta}_\infty\|^2 \leq \frac{1}{n}\|\boldsymbol{F}_0 - \boldsymbol{B\theta}_\infty\|_\infty^2 \lesssim \frac{1}{n}\sum_{k=1}^{d} J_k^{-2\alpha}. \tag{3.29}$$

By repeated applications of (3.8) and (3.17),

$$\begin{aligned}\|\boldsymbol{V}\|_{(2,2)} &= \|\boldsymbol{B}^T\boldsymbol{B}\|_{(2,2)}\|(\boldsymbol{B}^T\boldsymbol{B})^{-1}\|_{(2,2)}\|[\boldsymbol{\Omega} + (\boldsymbol{B}^T\boldsymbol{B})^{-1}]^{-1}\|_{(2,2)} \\ &\lesssim \left(\frac{n}{J}\right)\left(\frac{J}{n}\right)^2\left(c_1 + \frac{1}{C_2}\frac{J}{n}\right)^{-1} \lesssim \frac{J}{n}.\end{aligned} \tag{3.30}$$

Using (3.24), idempotency of $\boldsymbol{I}_n - \boldsymbol{P_B}$ and $(\boldsymbol{I}_n - \boldsymbol{P_B})\boldsymbol{B} = \boldsymbol{0}$, the second term in (3.28) can be written as

$$\frac{1}{n^2}(\boldsymbol{\theta}_\infty - \boldsymbol{\eta})^T\boldsymbol{B}^T(\boldsymbol{I}_n - \boldsymbol{P_B} + \boldsymbol{V})^2\boldsymbol{B}(\boldsymbol{\theta}_\infty - \boldsymbol{\eta}) = \frac{1}{n^2}(\boldsymbol{\theta}_\infty - \boldsymbol{\eta})^T\boldsymbol{B}^T\boldsymbol{V}^2\boldsymbol{B}(\boldsymbol{\theta}_\infty - \boldsymbol{\eta}).$$

In view of (3.30) and (3.17), the right hand side above is bounded by

$$\frac{1}{n^2}\|\boldsymbol{V}\|_{(2,2)}^2\|\boldsymbol{B}(\boldsymbol{\theta}_\infty - \boldsymbol{\eta})\|^2 \lesssim \frac{1}{n^2}\left(\frac{J}{n}\right)^2\|\boldsymbol{B}^T\boldsymbol{B}\|_{(2,2)}J\|\boldsymbol{\theta}_\infty - \boldsymbol{\eta}\|_\infty^2 \lesssim \frac{1}{n}\left(\frac{J}{n}\right)^2, \tag{3.31}$$

where $\|\boldsymbol{\theta}_\infty - \boldsymbol{\eta}\|_\infty^2$ is bounded using (3.20) and assumption on the prior. By Lemma D.3, the last term in (3.28) is $O(1/n)$. Combining this with the bounds established in (3.29) and (3.31) into (3.28), we obtain $\mathrm{Var}_0(\widehat{\sigma}_n^2) \lesssim 1/n$. By taking $J_k \asymp n^{1/(2\alpha+d)} \ll n$ for $k = 1, \ldots, n$ and $\alpha > d/2$, the mean square error is

$$\mathrm{E}_0(\widehat{\sigma}_n^2 - \sigma_0^2)^2 \lesssim \frac{1}{n} + \left(\frac{J}{n}\right)^2 + \sum_{k=1}^{d} J_k^{-4\alpha} \lesssim \frac{1}{n}. \quad \square$$

A Bayesian analog of Theorem 3.1 above is given below, which shows that the posterior of σ^2 in (3.16) converges to σ_0^2 at the parametric rate $1/\sqrt{n}$.

Theorem 3.2. *If $J_k \asymp n^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, and $\alpha > d/2$, we have for $R > 0$ and any $M_n \to \infty$,*

$$\sup_{\|f_0\|_{\mathcal{H}^\alpha} \le R} \mathrm{E}_0 \Pi\left(|\sigma^2 - \sigma_0^2| > \frac{M_n}{\sqrt{n}} \middle| \boldsymbol{Y}\right) \to 0. \tag{3.32}$$

Proof. By Markov's inequality, it suffices to show that $\mathrm{E}_0\mathrm{E}[(\sigma^2 - \sigma_0^2)^2|\boldsymbol{Y}] = O(n^{-1})$. The posterior mean and variance of σ^2 are

$$\begin{aligned} \mathrm{E}(\sigma^2|\boldsymbol{Y}) &= \frac{\gamma}{\beta+n-2} + \frac{n}{\beta+n-2}\widehat{\sigma}_n^2, \\ \mathrm{Var}(\sigma^2|\boldsymbol{Y}) &= \frac{4}{\beta+n-4}\left(\frac{\gamma}{\beta+n-2} + \frac{n}{\beta+n-2}\widehat{\sigma}_n^2\right)^2. \end{aligned} \tag{3.33}$$

We have $\mathrm{E}[(\sigma^2 - \sigma_0^2)^2|\boldsymbol{Y}] = [\mathrm{E}(\sigma^2|\boldsymbol{Y}) - \sigma_0^2]^2 + \mathrm{Var}(\sigma^2|\boldsymbol{Y})$. Theorem 3.1 implies that $\mathrm{E}_0(\widehat{\sigma}_n^2) = \sigma_0^2 + O(1/\sqrt{n})$, $\mathrm{Var}_0(\widehat{\sigma}_n^2) = O(1/n)$ and $\mathrm{E}_0(\widehat{\sigma}_n^2)^2 = \sigma_0^4 + O(1/n)$. Hence, $\mathrm{E}_0[\mathrm{E}(\sigma^2|\boldsymbol{Y}) - \sigma_0^2)]^2$ is

$$\begin{aligned} &[\mathrm{E}_0\mathrm{E}(\sigma^2|\boldsymbol{Y}) - \sigma_0^2]^2 + \mathrm{Var}_0[\mathrm{E}(\sigma^2|\boldsymbol{Y})] \\ &= \left\{O\left(\frac{1}{n}\right) + \left[1 + O\left(\frac{1}{n}\right)\right]\left[\sigma_0^2 + O\left(\frac{1}{\sqrt{n}}\right)\right] - \sigma_0^2\right\}^2 + \left[1 + O\left(\frac{1}{n}\right)\right]^2 O\left(\frac{1}{n}\right) \\ &= O\left(\frac{1}{n}\right), \end{aligned}$$

while $\mathrm{E}_0\mathrm{Var}(\sigma^2|\boldsymbol{Y})$ is

$$O\left(\frac{1}{n}\right)\left\{O\left(\frac{1}{n^2}\right) + \left[1 + O\left(\frac{1}{n}\right)\right]^2\left[\sigma_0^4 + O\left(\frac{1}{n}\right)\right]\right\} = O\left(\frac{1}{n}\right). \quad \square$$

3.3 Posterior convergence rates for f

Since the explicit expression for the posterior of f given $(\boldsymbol{Y}, \sigma^2)$ is available due to normal-normal conjugacy, we will derive posterior convergence rates by directly analyzing the posterior distribution. This is done by decomposing the posterior mean square error into three parts corresponding to: posterior variance, variance and bias of the posterior mean. We then bound these three terms separately and use Markov's inequality to arrive at our desired results. Theorem 3.3 below is for pointwise posterior convergence and is a crucial en-route step in deriving Theorems 3.4 and 3.5. For notational simplicity, we write $\sigma^2 \in \mathcal{U}_n$ if $|\sigma^2 - \sigma_0^2| \leq \zeta_n/\sqrt{n}$ for any $\zeta_n = o(\sqrt{n})$.

Theorem 3.3. *If $J_k \asymp n^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, we have for $R > 0$, every fixed $\boldsymbol{x} \in (0,1)^d$ and any $M_n \to \infty$,*

$$\textit{Empirical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(|f(\boldsymbol{x}) - f_0(\boldsymbol{x})| > M_n n^{-\alpha/(2\alpha+d)} | \boldsymbol{Y}) \to 0.$$

$$\textit{Hierarchical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \Pi(|f(\boldsymbol{x}) - f_0(\boldsymbol{x})| > M_n n^{-\alpha/(2\alpha+d)} | \boldsymbol{Y}) \to 0.$$

Proof. At any $\boldsymbol{x} \in (0,1)^d$, $\Pi(f(\boldsymbol{x})|\boldsymbol{Y}, \sigma^2) \sim \mathrm{N}(\boldsymbol{A}(\boldsymbol{x})\boldsymbol{Y} + \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\eta}, \sigma^2\Sigma(\boldsymbol{x}, \boldsymbol{x}))$ where $\boldsymbol{A}(\boldsymbol{x}), \boldsymbol{c}(\boldsymbol{x})$ and $\Sigma(\boldsymbol{x}, \boldsymbol{x})$ were given by (3.10), (3.11) and (3.12) respectively. Under P_0, $\boldsymbol{A}(\boldsymbol{x})\boldsymbol{Y} + \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\eta}$ is a sub-Gaussian random variable with mean $\boldsymbol{A}(\boldsymbol{x})\boldsymbol{F}_0 + \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\eta}$ and variance $\sigma_0^2\Psi(\boldsymbol{x}, \boldsymbol{x})$, where $\sigma_0^2\Psi$ is the covariance function of the sub-Gaussian process $\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}\boldsymbol{\eta}$ under P_0, such that for $\boldsymbol{x}, \boldsymbol{y} \in (0,1)^d$,

$$\Psi(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})^T \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \boldsymbol{B}^T\boldsymbol{B} \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{y}). \tag{3.34}$$

Note that the posterior variance $\sigma^2\Sigma(\boldsymbol{x}, \boldsymbol{x})$ does not depend on $\boldsymbol{Y}$ and f_0, while the posterior mean $\boldsymbol{A}(\boldsymbol{x})\boldsymbol{Y} + \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\eta}$ does not depend on σ^2. Therefore,

$$\begin{aligned}
&\sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}\{[f(\boldsymbol{x}) - f_0(\boldsymbol{x})]^2 | \boldsymbol{Y}, \sigma^2\} \\
&\leq \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}\{[f(\boldsymbol{x}) - \boldsymbol{A}(\boldsymbol{x})\boldsymbol{Y} - \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\eta}]^2 | \sigma^2\} + \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \mathrm{E}\{[\boldsymbol{A}(\boldsymbol{x})\boldsymbol{Y} + \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\eta} - f_0(\boldsymbol{x})]^2 | \boldsymbol{Y}\} \\
&= \sup_{\sigma^2 \in \mathcal{U}_n} \sigma^2\Sigma(\boldsymbol{x}, \boldsymbol{x}) + \sigma_0^2\Psi(\boldsymbol{x}, \boldsymbol{x}) + \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} [\boldsymbol{A}(\boldsymbol{x})\boldsymbol{F}_0 + \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\eta} - f_0(\boldsymbol{x})]^2.
\end{aligned} \tag{3.35}$$

To bound the posterior variance $\sigma^2\Sigma(\boldsymbol{x},\boldsymbol{x})$ uniformly over $\sigma^2 \in \mathcal{U}_n$, first observe that

$$\begin{aligned}\|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})\|^2 &= \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \prod_{k=1}^{d} B_{j_k,q_k}^2(x_k) \\ &\leq \prod_{k=1}^{d} \max_{1\leq j_k \leq J_k} B_{j_k,q_k}(x_k) \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} B_{j_k,q_k}(x_k) \leq 1. \end{aligned} \tag{3.36}$$

Cauchy-Schwarz inequality and the definition of spectral norm then imply

$$\begin{aligned}\sup_{\sigma^2\in\mathcal{U}_n} \sigma^2\Sigma(\boldsymbol{x},\boldsymbol{x}) &\leq \left(\sigma_0 + \frac{\zeta_n}{\sqrt{n}}\right) \left\| \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \right\|_{(2,2)} \\ &\lesssim \left(C_1 n \prod_{k=1}^{d} J_k^{-1} + \frac{1}{c_2} \right)^{-1} = O\left(\frac{1}{n} \prod_{k=1}^{d} J_k \right), \end{aligned} \tag{3.37}$$

where the last line follows from (3.18). In view of (3.34), the variance of the posterior mean, i.e., $\sigma_0^2\Psi(\boldsymbol{x},\boldsymbol{x})$ can be bounded above by

$$\begin{aligned}\sigma_0^2\Psi(\boldsymbol{x},\boldsymbol{x}) &\leq \sigma_0^2 \left\| \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \right\|_{(2,2)}^2 \|\boldsymbol{B}^T\boldsymbol{B}\|_{(2,2)} \|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})\|^2 \\ &= O\left(\frac{1}{n} \prod_{k=1}^{d} J_k \right)^2 O\left(n \prod_{k=1}^{d} J_k^{-1} \right) = O\left(\frac{1}{n} \prod_{k=1}^{d} J_k \right), \end{aligned} \tag{3.38}$$

where we have used (3.18), (3.17) and (3.36) to bound the three norms respectively. To bound the last squared bias term in (3.35), we will appeal (3.19) and the fact that $\|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})\|_1 = \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \prod_{k=1}^{d} B_{j_k,q_k}(x_k) = 1$. Hölder's inequality implies uniformly in $\|f_0\|_{\mathcal{H}^\alpha} \leq R$,

$$\begin{aligned}&|\boldsymbol{A}(\boldsymbol{x})\boldsymbol{F}_0 + \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\eta} - f_0(\boldsymbol{x})| \\ &= \left| \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})^T \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \left(\boldsymbol{B}^T\boldsymbol{F}_0 + \boldsymbol{\Omega}^{-1}\boldsymbol{\eta}\right) - f_0(\boldsymbol{x}) \right| \\ &\leq \left| \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})^T \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \left(\boldsymbol{B}^T\boldsymbol{F}_0 + \boldsymbol{\Omega}^{-1}\boldsymbol{\eta}\right) - \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})^T\boldsymbol{\theta}_\infty \right| + |f_0(\boldsymbol{x}) - \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})^T\boldsymbol{\theta}_\infty| \\ &\leq \left| \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})^T \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \left[\boldsymbol{B}^T(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty) + \boldsymbol{\Omega}^{-1}(\boldsymbol{\eta} - \boldsymbol{\theta}_\infty)\right] \right| + O\left(\sum_{k=1}^{d} J_k^{-\alpha} \right) \\ &\leq \left\| \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \right\|_{(\infty,\infty)} \left\{ \|\boldsymbol{B}^T(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty)\|_\infty + \left\|\boldsymbol{\Omega}^{-1}\right\|_{(\infty,\infty)} (\|\boldsymbol{\theta}_\infty\|_\infty + \|\boldsymbol{\eta}\|_\infty) \right\} \\ &\quad + O\left(\sum_{k=1}^{d} J_k^{-\alpha} \right). \end{aligned} \tag{3.39}$$

By assumption, $\boldsymbol{\Omega}^{-1}$ is $\boldsymbol{v}$-banded with fixed $\boldsymbol{v}$ and uniformly bounded entries. Therefore,

$\|\boldsymbol{\Omega}^{-1}\|_{(\infty,\infty)} = O(1)$. Since $\boldsymbol{B}^T\boldsymbol{B}$ is $\boldsymbol{q}$-banded, Lemma D.2 and equation (3.18) imply that $\|(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1})^{-1}\|_{(\infty,\infty)} = O(n^{-1}\prod_{k=1}^d J_k)$. Also, $\sup_{\|f_0\|_{\mathcal{H}^\alpha}\leq R}\|\boldsymbol{\theta}_\infty\|_\infty$ and $\|\boldsymbol{\eta}\|_\infty$ are both $O(1)$ by (3.20) and the assumption on the prior. To bound $\sup_{\|f_0\|_{\mathcal{H}^\alpha}\leq R}\|\boldsymbol{B}^T(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty)\|_\infty$, we use the non-negativity of B-splines, (3.19) and Lemma B.2 to write uniformly in $\|f_0\|_{\mathcal{H}^\alpha} \leq R$,

$$\begin{aligned}\|\boldsymbol{B}^T(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty)\|_\infty &\leq \max_{1\leq j_k\leq J_k, k=1,\ldots,d}\sum_{i=1}^n\prod_{k=1}^d B_{j_k,q_k}(X_{ik})|f_0(\boldsymbol{X}_i) - \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{X}_i)^T\boldsymbol{\theta}_\infty| \\ &\lesssim \sum_{k=1}^d J_k^{-\alpha}\max_{1\leq j_k\leq J_k, k=1,\ldots,d}\sum_{i=1}^n\prod_{k=1}^d B_{j_k,q_k}(X_{ik}) \\ &\lesssim n\sum_{k=1}^d J_k^{-\alpha}\left(\prod_{k=1}^d J_k^{-1}\right).\end{aligned}$$

Therefore, combining the bounds obtained and squaring (3.39), we have for any $\boldsymbol{x} \in (0,1)^d$,

$$\sup_{\|f_0\|_{\mathcal{H}^\alpha}\leq R}[\boldsymbol{A}(\boldsymbol{x})\boldsymbol{Y} + \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\eta} - f_0(\boldsymbol{x})]^2 \lesssim \frac{1}{n^2}\prod_{k=1}^d J_k^2 + \sum_{k=1}^d J_k^{-2\alpha}. \tag{3.40}$$

Combining all bounds from (3.37), (3.38) and (3.40) into (3.35), we obtain

$$\begin{aligned}\sup_{\|f_0\|_{\mathcal{H}^\alpha}\leq R}\mathrm{E}_0\sup_{\sigma^2\in\mathcal{U}_n}\mathrm{E}\left[(f(\boldsymbol{x}) - f_0(\boldsymbol{x}))^2|\boldsymbol{Y},\sigma^2\right] &\lesssim \frac{1}{n}\prod_{k=1}^d J_k + \frac{1}{n^2}\prod_{k=1}^d J_k^2 + \sum_{k=1}^d J_k^{-2\alpha} \\ &\lesssim \frac{1}{n}\prod_{k=1}^d J_k + \sum_{k=1}^d J_k^{-2\alpha},\end{aligned} \tag{3.41}$$

where the last line follows since $\prod_{k=1}^d J_k \leq n$ by assumption. To balance the orders of the two terms on the right, let $J_k = J$ for $k = 1,\ldots,d$. Then the right hand side of (3.41) reduces to

$$O\left(\frac{J^d}{n}\right) + O\left(J^{-2\alpha}\right). \tag{3.42}$$

Hence, they will have the same order when $J_k = J \asymp n^{1/(2\alpha+d)}$ for $k = 1,\ldots,d$. This implies that the corresponding posterior mean square error will have order $n^{-2\alpha/(2\alpha+d)}$.

Observe that by Markov's inequality, Theorem 3.1 implies for $R > 0$ and any $M_n \to \infty$,

$$\inf_{\|f_0\|_{\mathcal{H}^\alpha}\leq R}P_0\left(\sigma_0^2 - \frac{M_n}{\sqrt{n}} \leq \widehat{\sigma}_n^2 \leq \sigma_0^2 + \frac{M_n}{\sqrt{n}}\right) \to 1. \tag{3.43}$$

For the empirical Bayes version, by Markov's inequality, (3.41) and (3.43) above, we have for

$\epsilon_n = n^{-\alpha/(2\alpha+d)}$ and any $M_n \to \infty$,

$$\mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(|f(\boldsymbol{x}) - f_0(\boldsymbol{x})| > M_n \epsilon_n | \boldsymbol{Y}) \leq \frac{\mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}([f(\boldsymbol{x}) - f_0(\boldsymbol{x})]^2 | \boldsymbol{Y}, \sigma^2)}{M_n^2 \epsilon_n^2} + o(1) \to 0, \quad (3.44)$$

uniformly over $\|f_0\|_{\mathcal{H}^\alpha} \leq R$. For the hierarchical Bayes procedure, we have for $\epsilon_n = n^{-\alpha/(2\alpha+d)}$ and any $M_n \to \infty$,

$$\begin{aligned} &\mathrm{E}_0 \Pi(|f(\boldsymbol{x}) - f_0(\boldsymbol{x})| > M_n \epsilon_n | \boldsymbol{Y}) \\ &\leq \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \Pi(|f(\boldsymbol{x}) - f_0(\boldsymbol{x})| > M_n \epsilon_n | \boldsymbol{Y}, \sigma^2) + \mathrm{E}_0 \Pi\left(|\sigma^2 - \sigma_0^2| > \frac{\zeta_n}{\sqrt{n}} \middle| \boldsymbol{Y} \right), \end{aligned} \quad (3.45)$$

uniformly over $\|f_0\|_{\mathcal{H}^\alpha} \leq R$. The first term on the right hand side is $o(1)$ by (3.41), while the second term goes to zero by Theorem 3.2. □

Theorem 3.4. *If $J_k \asymp n^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, we have for $R > 0$ and every $M_n \to \infty$,*

$$\begin{aligned} \textit{Empirical Bayes:} \quad & \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(f : \|f - f_0\|_2 > M_n n^{-\alpha/(2\alpha+d)} | \boldsymbol{Y}) \to 0. \\ \textit{Hierarchical Bayes:} \quad & \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \Pi(f : \|f - f_0\|_2 > M_n n^{-\alpha/(2\alpha+d)} | \boldsymbol{Y}) \to 0. \end{aligned}$$

Proof. By Fubini's theorem,

$$\begin{aligned} \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}(\|f - f_0\|_2^2 | \boldsymbol{Y}, \sigma^2) &\leq \int_{(0,1)^d} \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}(|f(\boldsymbol{x}) - f_0(\boldsymbol{x})|^2 | \boldsymbol{Y}, \sigma^2) d\boldsymbol{x} \\ &\lesssim \frac{1}{n} \prod_{k=1}^d J_k + \sum_{k=1}^d J_k^{-2\alpha}, \end{aligned} \quad (3.46)$$

where the last line follows from (3.41). Setting $J_k \asymp n^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$ will balance the orders of these two terms in view of (3.65). The empirical and hierarchical Bayes posterior convergence rates then follow from (3.44) and (3.45) respectively with absolute values replaced by L_2-norms. □

Theorem 3.5. *If $J_k \asymp (n/\log n)^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, then for $R > 0$ and any $M_n \to \infty$,*

$$\begin{aligned} \textit{Empirical Bayes:} \quad & \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(\|f - f_0\|_\infty > M_n (\log n/n)^{\alpha/(2\alpha+d)} | \boldsymbol{Y}) \to 0. \\ \textit{Hierarchical Bayes:} \quad & \sup_{\|f_0\|_{\mathcal{H}^\alpha} \leq R} \mathrm{E}_0 \Pi(\|f - f_0\|_\infty > M_n (\log n/n)^{\alpha/(2\alpha+d)} | \boldsymbol{Y}) \to 0. \end{aligned}$$

Proof. Recall that $\Pi(f | \boldsymbol{Y}, \sigma^2) \sim \mathrm{GP}(\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}\boldsymbol{\eta}, \sigma^2 \Sigma)$, with $\boldsymbol{A}, \boldsymbol{c}$ and Σ defined in (3.10), (3.11)

and (3.12) respectively. Let $Z_n \sim \mathrm{GP}(0, \Sigma)$. Under the true distribution P_0, $\boldsymbol{AY} + \boldsymbol{c\eta}$ is a sub-Gaussian process with mean function $\boldsymbol{AF}_0 + \boldsymbol{c\eta}$ and covariance function $\sigma_0^2\Psi$, where Ψ was defined in (3.34). Let Q_n be a sub-Gaussian process with mean function 0 and covariance function $\sigma_0^2\Psi$. Note that Z_n does not depend on $\boldsymbol{Y}$ and f_0, while the posterior mean $\boldsymbol{AY} + \boldsymbol{c\eta}$ does not depend on σ^2. Therefore,

$$\begin{aligned}
&\sup_{\|f_0\|_{\mathcal{H}^\alpha} \le R} \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}(\|f - f_0\|_\infty^2 | \boldsymbol{Y}, \sigma^2) \\
&\le \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}(\|f - \boldsymbol{AY} - \boldsymbol{c\eta}\|_\infty^2 | \sigma^2) + \sup_{\|f_0\|_{\mathcal{H}^\alpha} \le R} \mathrm{E}_0\{\mathrm{E}(\|\boldsymbol{AY} + \boldsymbol{c\eta} - f_0\|_\infty^2 | \boldsymbol{Y})\} \\
&= \sup_{\sigma^2 \in \mathcal{U}_n} \sigma^2 \mathrm{E}\|Z_n\|_\infty^2 + \mathrm{E}\|Q_n\|_\infty^2 + \sup_{\|f_0\|_{\mathcal{H}^\alpha} \le R} \|\boldsymbol{AF}_0 + \boldsymbol{c\eta} - f_0\|_\infty^2. \qquad (3.47)
\end{aligned}$$

Since $Q_n = \boldsymbol{A\varepsilon}$, then by Property 2 of Appendix C and Assumption 2, Q_n is sub-Gaussian with respect to the semi-metric $d(\boldsymbol{t}, \boldsymbol{s}) = \sqrt{\mathrm{Var}(Q_n(\boldsymbol{t}) - Q_n(\boldsymbol{s}))}$. By Lemma B.4 with $\boldsymbol{r} = \boldsymbol{0}$, Z_n and Q_n satisfy the condition for Lemma C.1. Applying Lemma C.1 with $p = 2$, we see that for any $0 < \delta_n < 1$ and $T_n = N(\delta_n, (0,1)^d, \|\cdot\|)$,

$$\begin{aligned}
\sup_{\sigma^2 \in \mathcal{U}_n} \sigma^2 \mathrm{E}\|Z_n\|_\infty^2 &\lesssim \sigma_0^2 \log(1/\delta_n) \left\{ n\delta_n^2 + \max_{1 \le i \le T_n} \mathrm{E}[Z_n(\boldsymbol{u}_i)^2] \right\} \\
&\lesssim \log(1/\delta_n) \left(n\delta_n^2 + \frac{1}{n} \prod_{k=1}^d J_k \right), \qquad (3.48)
\end{aligned}$$

where the last line follows from (3.37). Similarly,

$$\mathrm{E}\|Q_n\|_\infty^2 \lesssim \log(1/\delta_n) \left\{ n\delta_n^2 + \max_{1 \le i \le T_n} \mathrm{E}[Q_n(\boldsymbol{u}_i)^2] \right\} \lesssim \log(1/\delta_n) \left(n\delta_n^2 + \frac{1}{n} \prod_{k=1}^d J_k \right), \qquad (3.49)$$

where the last line follows from (3.38). Setting $\delta_n \asymp n^{-1} \prod_{k=1}^d J_k^{1/2}$ for both (3.48) and (3.49), we have

$$\sup_{\sigma^2 \in \mathcal{U}_n} \sigma^2 \mathrm{E}\|Z_n\|_\infty^2 \lesssim \frac{\log n}{n} \prod_{k=1}^d J_k, \qquad \mathrm{E}\|Q_n\|_\infty^2 \lesssim \frac{\log n}{n} \prod_{k=1}^d J_k. \qquad (3.50)$$

Since the bound for (3.40) is uniform for all $\boldsymbol{x} \in (0,1)^d$ and all $\|f_0\|_{\mathcal{H}^\alpha} \le R$, the last deterministic squared bias term in (3.47) is

$$\sup_{\|f_0\|_{\mathcal{H}^\alpha} \le R} \|\boldsymbol{AF}_0 + \boldsymbol{c\eta} - f_0\|_\infty^2 = O\left(\frac{1}{n^2} \prod_{k=1}^d J_k^2 \right) + O\left(\sum_{k=1}^d J_k^{-2\alpha} \right). \qquad (3.51)$$

Therefore, combining the bounds for (3.50) and (3.51) into (3.47), we will obtain

$$\begin{aligned}\sup_{\|f_0\|_{\mathcal{H}^\alpha}\leq R} \mathrm{E}_0 \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}(\|f-f_0\|_\infty^2|\boldsymbol{Y},\sigma^2) &\lesssim \frac{\log n}{n}\prod_{k=1}^d J_k + \frac{1}{n^2}\prod_{k=1}^d J_k^2 + \sum_{k=1}^d J_k^{-2\alpha} \\ &\lesssim \frac{\log n}{n}\prod_{k=1}^d J_k + \sum_{k=1}^d J_k^{-2\alpha}, \end{aligned} \tag{3.52}$$

where the last line follows since $\prod_{k=1}^d J_k \leq n$ by assumption. To balance the orders of the two terms on the right, let $J_k = J$ for $k = 1, \ldots, d$. Then the right hand side of (3.52) reduces to

$$O\left(\frac{J^d}{n}\log n\right) + O\left(J^{-2\alpha}\right).$$

Hence, they will have the same order when $J_k = J \asymp (n/\log n)^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, yielding the rate $(\log n/n)^{2\alpha/(2\alpha+d)}$ for the posterior mean square error in (3.52). The empirical and hierarchical Bayes posterior convergence rates then follow from (3.44) and (3.45) respectively with absolute values replaced by sup-norms. □

Note that we will incur an extra logarithmic factor in the rate once we move from pointwise to supremum norm convergence. The pointwise, L_2-and L_∞-rates established above are the minimax rates (see Stone, 1980 for pointwise and Theorem 1 of Stone, 1982 for general L_r-norms, with $1 \leq r \leq \infty$). By examining the proofs of Theorems 3.3, 3.4 and 3.5 above, it also follows that the point estimator of $f_0(\boldsymbol{x})$, i.e., the posterior mean $\boldsymbol{A}(\boldsymbol{x})\boldsymbol{Y} + \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\eta}$ converges to $f_0(\boldsymbol{x})$ at the same $n^{-\alpha/(2\alpha+d)}$ rate. Similarly, the posterior mean of f, i.e., $\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}\boldsymbol{\eta}$ converges to f_0 under L_2-norm at the rate $n^{-\alpha/(2\alpha+d)}$, while in L_∞-norm, it is $(\log n/n)^{\alpha/(2\alpha+d)}$.

3.4 Posterior convergence rates for mixed partial derivatives

Recall that our prior is $f(\boldsymbol{x}) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})^T\boldsymbol{\theta}$ and $\boldsymbol{\theta}|\sigma^2 \sim \mathrm{N}(\boldsymbol{\eta}, \sigma^2\boldsymbol{\Omega})$. The $\boldsymbol{r} = (r_1, \ldots, r_d)^T$ mixed partial derivative of f with respect to $\boldsymbol{x}$ is

$$D^{\boldsymbol{r}} f(\boldsymbol{x}) = \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1,\ldots,j_d} \prod_{k=1}^d \frac{\partial^{r_k}}{\partial x_k^{r_k}} B_{j_k,q_k}(x_k).$$

We define $|\boldsymbol{r}| = \sum_{k=1}^d r_k$ for notational simplicity, and require that $|\boldsymbol{r}| < \alpha$. Let $\mathfrak{D}_{j_k}^{r_k}$ be the r_k forward difference quotient operator with respect to the j_kth index, such that for $r_k = 0$,

$\mathfrak{D}_{j_k}^0 \theta_{j_1,\ldots,j_d} = \theta_{j_1,\ldots,j_d}$, and for $r_k \geq 1$,

$$\mathfrak{D}_{j_k}^{r_k} \theta_{j_1,\ldots,j_d} = (q_k - r_k) \frac{\mathfrak{D}_{j_k}^{r_k-1} \theta_{j_1,\ldots,j_{k-1},j_k+1,j_{k+1},\ldots,j_d} - \mathfrak{D}_{j_k}^{r_k-1} \theta_{j_1,\ldots,j_{k-1},j_k,j_{k+1},\ldots,j_d}}{(t_{k,j_k} - t_{k,j_k-q_k+1})}. \tag{3.53}$$

Furthermore, let $(\prod_{k=1}^d \mathfrak{D}_{j_k}^{r_k}) \theta_{j_1,\ldots,j_d}$ be the application of $\mathfrak{D}_{j_k}^{r_k}$ to $\theta_{j_1,\ldots,j_d}$ for all direction $k = 1, \ldots, d$, i.e., $(\prod_{k=1}^d \mathfrak{D}_{j_k}^{r_k}) \theta_{j_1,\ldots,j_d} = \mathfrak{D}_{j_1}^{r_1} \ldots \mathfrak{D}_{j_d}^{r_d} \theta_{j_1,\ldots,j_d}$. Using equations (15) and (16) of Chapter X from de Boor, 2001, $D^{\boldsymbol{r}} f$ can also be written as

$$D^{\boldsymbol{r}} f(\boldsymbol{x}) = \sum_{j_1=1}^{J_1-r_1} \cdots \sum_{j_d=1}^{J_d-r_d} \left(\prod_{k=1}^d \mathfrak{D}_{j_k}^{r_k} \right) \theta_{j_1,\ldots,j_d} B_{j_k,q_k-r_k}(x_k) \tag{3.54}$$

where $B_{j_k,q_k-r_k}(\cdot)$ is the corresponding $q_k - r_k$ order B-spline.

Let $\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x}) = \{\prod_{k=1}^d B_{j_k,q_k-r_k}(x_k)\}_{j_1=1,\ldots,j_d=1}^{J_1-r_1,\ldots,J_d-r_d}$. Then $D^{\boldsymbol{r}} f(\boldsymbol{x}) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})^T \boldsymbol{W}_{\boldsymbol{r}} \boldsymbol{\theta}$, where $\boldsymbol{W}_{\boldsymbol{r}}$ is a $\prod_{k=1}^d (J_k - r_k) \times J$ matrix. The structure and entries of $\boldsymbol{W}_{\boldsymbol{r}}$ depend on the order of elements in $\boldsymbol{\theta}$, which does not have natural ordering since its d-dimensional indices do not form a totally ordered set. Instead, we impose a lexicographic or dictionary ordering to its elements. Each row of $\boldsymbol{W}_{\boldsymbol{r}}$ has only $\prod_{k=1}^d (r_k + 1)$ nonzero entries and their arrangement is analogues to a banded matrix, namely the position of the $\prod_{k=1}^d (r_k + 1)$ nonzero entries in the current row is a shift of one entry to the right of the $\prod_{k=1}^d (r_k + 1)$ nonzero entries' position in the previous row. Also, each column of $\boldsymbol{W}_{\boldsymbol{r}}$ has at most $\prod_{k=1}^d (r_k + 1)$ nonzero entries. We index the rows and columns of $\boldsymbol{W}_{\boldsymbol{r}}$ using d-dimensional indices as in Definition 3.1.

Define $\tilde{r} = \prod_{k=1}^d (r_k + 1) - 2$, and let $\mathcal{G} = \{\boldsymbol{u} = (u_1, \ldots, u_d) : \boldsymbol{0} \leq \boldsymbol{u} \leq \boldsymbol{r}, \boldsymbol{u} \neq \boldsymbol{r}, \boldsymbol{u} \neq \boldsymbol{0}\}$. By ordering the elements in $\mathcal{G}$ lexicographically, we can enumerate its elements by $\mathcal{G} = \{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{\tilde{r}}\}$. Furthermore, define sets $\mathcal{I} = \{(i_1, \ldots, i_d) : 1 \leq i_k \leq J_k - r_k, k = 1, \ldots, d\}$ and $\mathcal{J} = \{(j_1, \ldots, j_d) : 1 \leq j_k \leq J_k, k = 1, \ldots, d\}$, where we order their elements lexicographically. Let $w_{\boldsymbol{i},\boldsymbol{j}}^{(\boldsymbol{r})}$ denote the $(\boldsymbol{i}, \boldsymbol{j})$th element of $\boldsymbol{W}_{\boldsymbol{r}}$ such that $\boldsymbol{i} \in \mathcal{I}$ and $\boldsymbol{j} \in \mathcal{J}$. Also, let $\boldsymbol{\theta} = \{\theta_{\boldsymbol{j}} : \boldsymbol{j} \in \mathcal{J}\}$. The expressions for the nonzero entries of $\boldsymbol{W}_{\boldsymbol{r}}$ can be described as follows: for each row $\boldsymbol{i} \in \mathcal{I}$, the first and last nonzero entries are given by

$$\begin{aligned} w_{\boldsymbol{i},\boldsymbol{i}}^{(\boldsymbol{r})} &= (-1)^{\sum_{k=1}^d r_k} \prod_{k=1}^d \prod_{l=1}^{r_k} \frac{q_k - l}{t_{k,i_k} - t_{k,i_k-q_k+l}}, \\ w_{\boldsymbol{i},\boldsymbol{i}+\boldsymbol{r}}^{(\boldsymbol{r})} &= \prod_{k=1}^d \prod_{l=1}^{r_k} \frac{q_k - l}{t_{k,i_k+l-1} - t_{k,i_k-q_k+r_k}}. \end{aligned} \tag{3.55}$$

If $\tilde{r}$ is odd, we partition $\mathcal{G} = \mathcal{G}_1 \cup \{\boldsymbol{g}_{(\tilde{r}+1)/2}\} \cup \mathcal{G}_2$ such that $\mathcal{G}_1 = \{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{(\tilde{r}-1)/2}\}$ and

$\mathcal{G}_2 = \{\boldsymbol{g}_{(\tilde{r}+3)/2}, \ldots, \boldsymbol{g}_{\tilde{r}}\}$. The intermediate nonzero entries for $\boldsymbol{h} = (h_1, \ldots, h_d)^T \in \mathcal{G}_1$ are

$$w^{(\boldsymbol{r})}_{\boldsymbol{i},\boldsymbol{i}+\boldsymbol{h}} = (-1)^{\sum_{k=1}^d (r_k - h_k)} w^{(\boldsymbol{r})}_{\boldsymbol{i},\boldsymbol{i}} \left[1 + \prod_{k=1}^d \sum_{t=1}^{\binom{r_k}{h_k}-1} \prod_{s=1}^t \frac{t_{k,i_k} - t_{k,i_k - q_k + s}}{t_{k,i_k+1} - t_{k,i_k+1-q_k+s}} \right], \tag{3.56}$$

while for $\boldsymbol{h} \in \mathcal{G}_2$, it is

$$w^{(\boldsymbol{r})}_{\boldsymbol{i},\boldsymbol{i}+\boldsymbol{h}} = (-1)^{\sum_{k=1}^d (r_k - h_k)} w^{(\boldsymbol{r})}_{\boldsymbol{i},\boldsymbol{i}+\boldsymbol{r}} \left[1 + \prod_{k=1}^d \sum_{t=1}^{\binom{r_k}{h_k}-1} \prod_{s=1}^t \frac{t_{k,i_k+r_k-s} - t_{k,i_k+r_k-q_k}}{t_{k,i_k+r_k-1-s} - t_{k,i+r_k-1-q_k}} \right]. \tag{3.57}$$

When $\boldsymbol{h} = \boldsymbol{g}_{(\tilde{r}+1)/2}$, then $w^{(\boldsymbol{r})}_{\boldsymbol{i},\boldsymbol{i}+\boldsymbol{h}}$ is

$$\begin{aligned} & w^{(\boldsymbol{r})}_{\boldsymbol{i},\boldsymbol{i}} \prod_{k=1}^d \left(\frac{t_{k,i_k} - t_{k,i_k-q_k+1}}{t_{k,i_k+1} - t_{k,i_k+2-q_k}} \right) \left[1 + \prod_{k=1}^d \sum_{t=1}^{\binom{r_k-1}{r_k/2-1}-1} \prod_{s=0}^{t-1} \frac{t_{k,i_k+s} - t_{k,i_k+2-q_k}}{t_{k,i_k+1+s} - t_{k,i_k+3-q_k}} \right] \\ & \quad + w^{(\boldsymbol{r})}_{\boldsymbol{i},\boldsymbol{i}+\boldsymbol{r}} \prod_{k=1}^d \left(\frac{t_{k,i_k+r_k-1} - t_{k,i_k+r_k-q_k}}{t_{k,i_k+r_k-2} - t_{k,i_k+r_k-1-q_k}} \right) \\ & \quad \times \left[1 + \prod_{k=1}^d \sum_{t=1}^{\binom{r_k-1}{r_k/2}-1} \prod_{s=0}^{t-1} \frac{t_{k,i_k+r_k-2} - t_{k,i_k+r_k-q_k-s}}{t_{k,i_k+r_k-3} - t_{k,i_k+r_k-1-q_k-s}} \right]. \end{aligned} \tag{3.58}$$

If $\tilde{r}$ is even, we partition $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$ where $\mathcal{G}_1 = \{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{\tilde{r}/2}\}$ and $\mathcal{G}_2 = \{\boldsymbol{g}_{\tilde{r}/2+1}, \ldots, \boldsymbol{g}_{\tilde{r}}\}$. Then the expression for $w^{(m)}_{\boldsymbol{i},\boldsymbol{i}+\boldsymbol{h}}$ is (3.56) for $\boldsymbol{h} \in \mathcal{G}_1$ and is (3.57) for $\boldsymbol{h} \in \mathcal{G}_2$.

The conditional posterior distribution for $D^{\boldsymbol{r}} f$ is $\Pi(D^{\boldsymbol{r}} f | \boldsymbol{Y}, \sigma^2) \sim \mathrm{GP}(\boldsymbol{A}_{\boldsymbol{r}} \boldsymbol{Y} + \boldsymbol{c}_{\boldsymbol{r}} \boldsymbol{\eta}, \sigma^2 \Sigma_{\boldsymbol{r}})$, such that for $\boldsymbol{x}, \boldsymbol{y} \in (0,1)^d$,

$$\boldsymbol{A}_{\boldsymbol{r}}(\boldsymbol{x}) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})^T \boldsymbol{W}_{\boldsymbol{r}} \left(\boldsymbol{B}^T \boldsymbol{B} + \boldsymbol{\Omega}^{-1} \right)^{-1} \boldsymbol{B}^T, \tag{3.59}$$

$$\boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x}) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})^T \boldsymbol{W}_{\boldsymbol{r}} \left(\boldsymbol{B}^T \boldsymbol{B} + \boldsymbol{\Omega}^{-1} \right)^{-1} \boldsymbol{\Omega}^{-1}, \tag{3.60}$$

$$\Sigma_{\boldsymbol{r}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})^T \boldsymbol{W}_{\boldsymbol{r}} \left(\boldsymbol{B}^T \boldsymbol{B} + \boldsymbol{\Omega}^{-1} \right)^{-1} \boldsymbol{W}_{\boldsymbol{r}}^T \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{y}). \tag{3.61}$$

Note that if $\boldsymbol{r} = \boldsymbol{0}$ and defining $\boldsymbol{W}_{\boldsymbol{0}} = \boldsymbol{I}_J$, the three operators (3.59), (3.60) and (3.61) above reduce to that defined in (3.10), (3.11) and (3.12) respectively. Using the same reasoning as before, we see that under Assumption 2 with Properties 1 and 2 of Appendix C, $\boldsymbol{A}_{\boldsymbol{r}} \boldsymbol{Y} + \boldsymbol{c}_{\boldsymbol{r}} \boldsymbol{\eta}$ is a sub-Gaussian process under P_0. Analytical expression for the conditional posterior of $D^{\boldsymbol{r}} f$ with $\boldsymbol{r} \in \mathbb{N}_0^d$ and $|\boldsymbol{r}| < \alpha$ is available due to normal-normal conjugacy, and we derive its pointwise,

L_2- and L_∞-posterior convergence rates by directly bounding the posterior mean square error. Note that Theorem 3.6 below is an important en-route step in deriving Theorems 3.7 and 3.8.

Theorem 3.6. *If $J_k \asymp n^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, then for $R > 0$, any $\boldsymbol{x} \in (0,1)^d$ and any $M_n \to \infty$, we have uniformly over $\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \leq R$,*

$$\begin{aligned} \textit{Empirical Bayes:} \quad & \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(|D^{\boldsymbol{r}} f(\boldsymbol{x}) - D^{\boldsymbol{r}} f_0(\boldsymbol{x})| > M_n n^{-(\alpha-|\boldsymbol{r}|)/(2\alpha+d)} | \boldsymbol{Y}) \to 0. \\ \textit{Hierarchical Bayes:} \quad & \mathrm{E}_0 \Pi(|D^{\boldsymbol{r}} f(\boldsymbol{x}) - D^{\boldsymbol{r}} f_0(\boldsymbol{x})| > M_n n^{-(\alpha-|\boldsymbol{r}|)/(2\alpha+d)} | \boldsymbol{Y}) \to 0. \end{aligned}$$

Proof. The posterior rate calculation proceeds similarly as illustrated in the proof of Theorem 3.3. That is, by bounding the corresponding posterior variance, variance of the posterior mean and the squared bias as in (3.35), and making the substitutions $\boldsymbol{A} = \boldsymbol{A}_{\boldsymbol{r}}$, $\boldsymbol{c} = \boldsymbol{c}_{\boldsymbol{r}}$ and $\Sigma = \Sigma_{\boldsymbol{r}}$, where $\boldsymbol{A}_{\boldsymbol{r}}, \boldsymbol{c}_{\boldsymbol{r}}$ and $\Sigma_{\boldsymbol{r}}$ were defined in (3.59), (3.60) and (3.61) respectively. The only difference is the extra matrix factor $\boldsymbol{W}_{\boldsymbol{r}}$ in these calculations.

Recall that each row of $\boldsymbol{W}_{\boldsymbol{r}}$ has $\prod_{k=1}^d (r_k + 1)$ nonzero entries and each column has at most $\prod_{k=1}^d (r_k + 1)$ nonzero entries. From the explicit expressions of these nonzero entries given in (3.55)–(3.58), assumption (3.4) and Lemma B.1 imply that each of these nonzero entries is of the order $\prod_{k=1}^d \Delta_k^{-r_k} \asymp \prod_{k=1}^d J_k^{r_k}$. This further implies that

$$\begin{aligned} \|\boldsymbol{W}_{\boldsymbol{r}}\|_{(\infty,\infty)} &= \max_{1 \leq i \leq \prod_{k=1}^d (J_k - r_k)} \sum_{j=1}^{J} |(\boldsymbol{W}_{\boldsymbol{r}})_{ij}| = O\left(\prod_{k=1}^d \frac{1}{\Delta_k^{r_k}} \right) = O\left(\prod_{k=1}^d J_k^{r_k} \right), \\ \|\boldsymbol{W}_{\boldsymbol{r}}^T\|_{(\infty,\infty)} &= \max_{1 \leq j \leq J} \sum_{i=1}^{\prod_{k=1}^d (J_k - r_k)} |(\boldsymbol{W}_{\boldsymbol{r}})_{ij}| = O\left(\prod_{k=1}^d \frac{1}{\Delta_k^{r_k}} \right) = O\left(\prod_{k=1}^d J_k^{r_k} \right) \end{aligned}$$

and

$$\|\boldsymbol{W}_{\boldsymbol{r}}^T \boldsymbol{W}_{\boldsymbol{r}}\|_{(2,2)} \leq \|\boldsymbol{W}_{\boldsymbol{r}}^T \boldsymbol{W}_{\boldsymbol{r}}\|_{(\infty,\infty)} \leq \|\boldsymbol{W}_{\boldsymbol{r}}\|_{(\infty,\infty)} \|\boldsymbol{W}_{\boldsymbol{r}}^T\|_{(\infty,\infty)} = O\left(\prod_{k=1}^d J_k^{2r_k} \right). \tag{3.62}$$

It follows that the posterior mean square error is inflated by the introduction of a $\prod_{k=1}^d J_k^{2r_k}$ factor in the bounds for the posterior variance and variance of the posterior mean, and a $\prod_{k=1}^d J_k^{r_k}$ in the bias. Also, the bound on tensor-product B-splines approximation error in (3.19) but for $D^{\boldsymbol{r}} f_0$ is of the order $\sum_{k=1}^d J_k^{-(\alpha-r_k)}$ uniformly in $\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \leq R$, as a result of applying Corollary 6.21 of Schumaker, 2007 (or Theorem 26, Chapter XII of de Boor, 2001).

Therefore, the derivative analogs of (3.40) and (3.41) for any $\boldsymbol{x} \in (0,1)^d$ are given by

$$\sup_{\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R} |\boldsymbol{A}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{Y} + \boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{\eta} - D^{\boldsymbol{r}} f_0(\boldsymbol{x})|^2 \lesssim \prod_{k=1}^d J_k^{2r_k} \left(\frac{1}{n^2} \prod_{k=1}^d J_k^2 + \sum_{k=1}^d J_k^{-2\alpha} \right), \quad (3.63)$$

and uniformly over $\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R$,

$$\begin{aligned} \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}\left[(D^{\boldsymbol{r}} f(\boldsymbol{x}) - D^{\boldsymbol{r}} f_0(\boldsymbol{x}))^2 | \boldsymbol{Y} \right] &\lesssim \left(\frac{1}{n} \prod_{k=1}^d J_k^{2r_k+1} \right) + \prod_{k=1}^d J_k^{2r_k} \left(\frac{1}{n^2} \prod_{k=1}^d J_k^2 + \sum_{l=1}^d J_k^{-2\alpha} \right) \\ &\lesssim \prod_{k=1}^d J_k^{2r_k} \left(\frac{1}{n} \prod_{k=1}^d J_k + \sum_{k=1}^d J_k^{-2\alpha} \right) \end{aligned} \quad (3.64)$$

respectively, where the last line above follows from $\prod_{k=1}^d J_k \le n$. To balance the orders of the two terms on the right, let $J_k = J$, $k = 1, \ldots, d$. Then the right hand side of (3.64) reduces to

$$O\left(\frac{J^{2\sum_{k=1}^d r_k + d}}{n} \right) + O\left(J^{-2(\alpha - \sum_{k=1}^d r_k)} \right). \quad (3.65)$$

Hence they will have the same order when $J_k = J \asymp n^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$ and all $\boldsymbol{r} \in \mathbb{N}_0^d$ such that $\sum_{k=1}^d r_k < \alpha$. This implies that the posterior mean square error in (3.64) will have order $n^{-2(\alpha - \sum_{k=1}^d r_k)/(2\alpha+d)}$. The empirical and hierarchical Bayes posterior convergence rates then follow from (3.44) and (3.45) respectively with f and f_0 replaced by $D^{\boldsymbol{r}} f$ and $D^{\boldsymbol{r}} f_0$. □

Theorem 3.7. *If $J_k \asymp n^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, then for $R > 0$ and every $M_n \to \infty$, we have uniformly over $\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R$,*

$$\begin{aligned} \textit{Empirical Bayes:} \quad & \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(f : \|D^{\boldsymbol{r}} f - D^{\boldsymbol{r}} f_0\|_2 > M_n n^{-(\alpha-|\boldsymbol{r}|)/(2\alpha+d)} | \boldsymbol{Y}) \to 0. \\ \textit{Hierarchical Bayes:} \quad & \mathrm{E}_0 \Pi(f : \|D^{\boldsymbol{r}} f - D^{\boldsymbol{r}} f_0\|_2 > M_n n^{-(\alpha-|\boldsymbol{r}|)/(2\alpha+d)} | \boldsymbol{Y}) \to 0. \end{aligned}$$

Proof. Following the steps in (3.46) with f and f_0 replaced by $D^{\boldsymbol{r}} f$ and $D^{\boldsymbol{r}} f_0$, the bound is given by (3.64). Set $J_k \asymp n^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, to balance the orders of the two terms. The result then follows from Markov's inequality. □

Theorem 3.8. *If $J_k \asymp (n/\log n)^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, then for $R > 0$ and any $M_n \to \infty$, we have uniformly over $\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R$,*

$$\begin{aligned} \textit{Empirical Bayes:} \quad & \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(\|D^{\boldsymbol{r}} f - D^{\boldsymbol{r}} f_0\|_\infty > M_n (\log n/n)^{(\alpha-|\boldsymbol{r}|)/(2\alpha+d)} | \boldsymbol{Y}) \to 0. \\ \textit{Hierarchical Bayes:} \quad & \mathrm{E}_0 \Pi(\|D^{\boldsymbol{r}} f - D^{\boldsymbol{r}} f_0\|_\infty > M_n (\log n/n)^{(\alpha-|\boldsymbol{r}|)/(2\alpha+d)} | \boldsymbol{Y}) \to 0. \end{aligned}$$

Proof. The calculations for finding the sup-norm rate as described in Theorem 3.5 apply with slight modifications. The only difference is the $\boldsymbol{W_r}$ matrix appearing in $\boldsymbol{A_r}, \boldsymbol{c_r}$ and $\Sigma_{\boldsymbol{r}}$ from (3.59), (3.60) and (3.61) respectively. Recall that $\Pi(D^{\boldsymbol{r}} f|\boldsymbol{Y}, \sigma^2) \sim \text{GP}(\boldsymbol{A_r Y} + \boldsymbol{c_r}\boldsymbol{\eta}, \sigma^2\Sigma_{\boldsymbol{r}})$. Define $Z_{n,\boldsymbol{r}} \sim \text{GP}(0, \Sigma_{\boldsymbol{r}})$. Under P_0, $\boldsymbol{A_r Y} + \boldsymbol{c_r}\boldsymbol{\eta}$ is a sub-Gaussian process with mean function $\boldsymbol{A_r F}_0 + \boldsymbol{c_r}\boldsymbol{\eta}$ and covariance function $\sigma_0^2\Psi_{\boldsymbol{r}}$, such that for any $\boldsymbol{x}, \boldsymbol{y} \in (0,1)^d$,

$$\Psi_{\boldsymbol{r}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})^T \boldsymbol{W_r} \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \boldsymbol{B}^T\boldsymbol{B} \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \boldsymbol{W}_{\boldsymbol{r}}^T \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{y}). \tag{3.66}$$

Define $Q_{n,\boldsymbol{r}}$ to be a sub-Gaussian process with mean function 0 and covariance function $\sigma_0^2\Psi_{\boldsymbol{r}}$. From (3.62), we see that the $\|\cdot\|_{(2,2)}$ matrix norm of $\boldsymbol{W_r}$ is of the order $\prod_{k=1}^d J_k^{r_k}$. Observe that for any $\boldsymbol{x} \in (0,1)^d$,

$$\begin{aligned}
\Sigma_{\boldsymbol{r}}(\boldsymbol{x}, \boldsymbol{x}) &= \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(x)^T \boldsymbol{W_r} \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \boldsymbol{W}_{\boldsymbol{r}}^T \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x}) \\
&\leq \|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})\|^2 \|\boldsymbol{W}_{\boldsymbol{r}}^T\boldsymbol{W_r}\|_{(2,2)} \| \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \|_{(2,2)} \\
&\lesssim \frac{1}{n} \prod_{k=1}^d J_k^{2r_k+1},
\end{aligned} \tag{3.67}$$

where we have used (3.36), (3.62) and (3.18) to bound the norms in line 2 respectively. By similar reasoning, we have for any $\boldsymbol{x} \in (0,1)^d$,

$$\begin{aligned}
\Psi_{\boldsymbol{r}}(\boldsymbol{x}, \boldsymbol{x}) &\leq \|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})\|^2 \|\boldsymbol{W}_{\boldsymbol{r}}^T\boldsymbol{W_r}\|_{(2,2)} \|\boldsymbol{B}^T\boldsymbol{B}\|_{(2,2)} \left\| \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \right\|_{(2,2)} \\
&\lesssim \frac{1}{n} \prod_{k=1}^d J_k^{2r_k+1},
\end{aligned} \tag{3.68}$$

where we have used (3.36), (3.62), (3.17) and (3.18) to bound the norms in the first line respectively. Since $Q_{n,\boldsymbol{r}} = \boldsymbol{A_r}\boldsymbol{\varepsilon}$, then by Property 2 of Appendix C and Assumption 2, $Q_{n,\boldsymbol{r}}$ is sub-Gaussian with respect to the semi-metric $d(\boldsymbol{t}, \boldsymbol{s}) = \sqrt{\text{Var}(Q_{n,\boldsymbol{r}}(\boldsymbol{t}) - Q_{n,\boldsymbol{r}}(\boldsymbol{s}))}$. By Lemma B.4, both $Z_{n,\boldsymbol{r}}$ and $Q_{n,\boldsymbol{r}}$ satisfy the condition for Lemma C.1. Therefore, Lemma C.1 with $p = 2$ and the two upper bounds from (3.67), (3.68) imply that for any $0 < \delta_n < 1$ and $T_n = N(\delta_n, (0,1)^d, \|\cdot\|)$,

$$\begin{aligned}
\sup_{\sigma^2 \in \mathcal{U}_n} \sigma^2 \text{E}(\|Z_{n,\boldsymbol{r}}\|_\infty^2) &\lesssim \sigma_0^2 \log(1/\delta_n) \left[n\delta_n^2 + \max_{1 \leq i \leq T_n} \Sigma_{\boldsymbol{r}}(\boldsymbol{u}_i, \boldsymbol{u}_i) \right] \\
&\lesssim \log(1/\delta_n) \left(n\delta_n^2 + \frac{1}{n} \prod_{k=1}^d J_k^{2r_k+1} \right)
\end{aligned} \tag{3.69}$$

and

$$\begin{aligned}\mathrm{E}(\|Q_{n,\boldsymbol{r}}\|_\infty^2) &\lesssim \log(1/\delta_n)\left[n\delta_n^2 + \max_{1\le i\le T_n}\Psi_{\boldsymbol{r}}(\boldsymbol{u}_i,\boldsymbol{u}_i)\right]\\ &\lesssim \log(1/\delta_n)\left(n\delta_n^2 + \frac{1}{n}\prod_{k=1}^d J_k^{2r_k+1}\right). \end{aligned} \tag{3.70}$$

By setting $\delta_n^2 \asymp n^{-2}\prod_{k=1}^d J_k^{2r_k+1}$ for both (3.69) and (3.70), we have

$$\sup_{\sigma^2\in\mathcal{U}_n}\sigma^2\mathrm{E}(\|Z_{n,\boldsymbol{r}}\|_\infty^2) \lesssim \frac{\log n}{n}\prod_{k=1}^d J_k^{2r_k+1}, \qquad \mathrm{E}(\|Q_{n,\boldsymbol{r}}\|_\infty^2) \lesssim \frac{\log n}{n}\prod_{k=1}^d J_k^{2r_k+1}. \tag{3.71}$$

Since the bound for (3.63) is uniform for all $\boldsymbol{x}\in(0,1)^d$ and all $\|D^{\boldsymbol{r}}f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}}\le R$, we have

$$\sup_{\|D^{\boldsymbol{r}}f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}}\le R}\|\boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{F}_0 + \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta} - D^{\boldsymbol{r}}f_0\|_\infty^2 \lesssim \prod_{k=1}^d J_k^{2r_k}\left(\frac{1}{n^2}\prod_{k=1}^d J_k^2 + \sum_{k=1}^d J_k^{-2\alpha}\right). \tag{3.72}$$

Therefore, the derivative analog of (3.52) with f and f_0 replaced by $D^{\boldsymbol{r}}f$ and $D^{\boldsymbol{r}}f_0$ is

$$\sup_{\|D^{\boldsymbol{r}}f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}}\le R}\mathrm{E}_0\sup_{\sigma^2\in\mathcal{U}_n}\mathrm{E}(\|D^{\boldsymbol{r}}f - D^{\boldsymbol{r}}f_0\|_\infty^2|\boldsymbol{Y}) \lesssim \prod_{k=1}^d J_k^{2r_k}\left(\frac{\log n}{n}\prod_{k=1}^d J_k + \sum_{k=1}^d J_k^{-2\alpha}\right). \tag{3.73}$$

To balance the orders of the two terms on the right, let $J_k = J$ for $k = 1,\ldots,d$. Then the right hand side of (3.73) reduces to

$$O\left(\frac{J^{2\sum_{k=1}^d r_k + d}}{n}\log n\right) + O\left(J^{-2(\alpha-\sum_{k=1}^d r_k)}\right).$$

They will have the same order when $J_k = J \asymp (n/\log n)^{1/(2\alpha+d)}$ for $k = 1,\ldots,d$, and all $\sum_{k=1}^d r_k < \alpha$ including $\boldsymbol{r} = \boldsymbol{0}$, yielding the rate $(\log n/n)^{2(\alpha-\sum_{k=1}^d r_k)/(2\alpha+d)}$ for the posterior mean square error in (3.73). The empirical and hierarchical Bayes posterior convergence rates then follow from (3.44) and (3.45) respectively with absolute values replaced by sup-norms and f, f_0 replaced by $D^{\boldsymbol{r}}f$, $D^{\boldsymbol{r}}f_0$. □

Given any mode of convergence, i.e., under pointwise, L_2-or L_∞-norms, the same optimal J_k for $k = 1,\ldots,d$, applies to both f and all its mixed partial derivatives with $\boldsymbol{r}\in\mathbb{N}_0^d$ such that $|\boldsymbol{r}| < \alpha$. If we view $D^{\boldsymbol{0}}f \equiv f$, then posterior convergence rates established in Section 3.3 are special cases of Section 3.4 by setting $\boldsymbol{r} = \boldsymbol{0}$. These rates coincide with the minimax rates for estimating $D^{\boldsymbol{r}}f(\boldsymbol{x})$ at any $\boldsymbol{x}\in(0,1)^d$ (Stone, 1980), $D^{\boldsymbol{r}}f$ in L_2-and L_∞-norms (Theorem 1 of Stone, 1982). It also follows from the proof that the posterior mean of $D^{\boldsymbol{r}}f(\boldsymbol{x})$ converges

to $D^{\boldsymbol{r}} f_0(\boldsymbol{x})$ at any $\boldsymbol{x} \in (0,1)^d$ at the rate given in Theorem 3.6. By the same reasoning, the posterior mean of $D^{\boldsymbol{r}} f$ converges to $D^{\boldsymbol{r}} f_0$ under L_2-and L_∞-norms at the rates given by Theorems 3.7 and 3.8 respectively.

3.5 Extension to anisotropic Hölder space

In this section, we will investigate the effect of anisotropy on posterior rates of convergence. A function f is anisotropic if it has different smoothness at different dimensions. Formally, if f is anisotropic with order $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^T$, then f has partial derivatives up to order α_k at direction k for $k = 1, \ldots, d$.

Let $\boldsymbol{r} \in \mathbb{N}_0^d$ such that $\sum_{k=1}^d r_k/\alpha_k < 1$. We define the anisotropic Hölder norm as

$$\|D^{\boldsymbol{r}} f\|_{\boldsymbol{\alpha},\infty} = \|D^{\boldsymbol{r}} f\|_\infty + \sum_{k=1}^d \left\| D^{(\alpha_k - r_k)\boldsymbol{e}_k} D^{\boldsymbol{r}} f \right\|_\infty . \tag{3.74}$$

Note that the norm above includes the case where $D^{\boldsymbol{0}} f \equiv f$. We introduce the anisotropic Hölder space $\mathcal{H}^{\boldsymbol{\alpha}}((0,1)^d)$ of order $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^T \in \mathbb{N}^d$ with domain $(0,1)^d$, consisting of functions $f : (0,1)^d \to \mathbb{R}$ such that $\|f\|_{\boldsymbol{\alpha},\infty} < \infty$. Therefore, if f belongs to this space, then $\partial^{\alpha_k} f / \partial x_k^{\alpha_k}$ for $k = 1, \ldots, d$ are uniformly bounded. Here, we assume $\boldsymbol{\alpha} \in \mathbb{N}^d$ because approximation results for anisotropic tensor-product B-splines assume integer smoothness (see Chapter 12, Section 3 of Schumaker, 2007). Hence strictly speaking, this section is not a generalization of the isotropic case discussed in previous sections, where α was not restricted to be positive integers. The anisotropic analog of Assumption 2 is as follows:

Assumption 3. Under the true distribution P_0, we assume $Y_i = f_0(\boldsymbol{X}_i) + \varepsilon_i$, such that ε_i are i.i.d. with mean 0, variance σ_0^2 and $\|\varepsilon_i\|_{\psi_2} < \infty$ for $i = 1, \ldots, n$. Also, f_0 belongs to the anisotropic Hölder space $\mathcal{H}^{\boldsymbol{\alpha}}((0,1)^d)$ of order $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^T \in \mathbb{N}^d$, such that $\alpha_k > \max\{d/2, 2\}$ for $k = 1, \ldots, d$. To avoid boundary effects, f_0 can be extended to a slightly bigger set $(-\epsilon, 1+\epsilon)^d$ for some $\epsilon > 0$.

Define α^* to be the harmonic mean of $(\alpha_1, \ldots, \alpha_d)^T$, i.e., $\alpha^{*-1} = d^{-1} \sum_{k=1}^d \alpha_k^{-1}$. Note that if $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)^T$, then the anisotropic Hölder space $\mathcal{H}^{\boldsymbol{\alpha}}((0,1)^d)$ reduces to its isotropic counterpart $\mathcal{H}^{\alpha}((0,1)^d)$. As in the isotropic case, f_0 has a unique continuous extension onto the closure of $(0,1)^d$. With the exception that $\boldsymbol{\alpha} \in \mathbb{N}^d$, all prior specifications and assumption on the knots for each B-spline component are identical to Section 3.2. For random covariates, $J_k \lesssim n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, and $\alpha^* \geq d/2$ ensure (3.5), while the deterministic case is identical.

Hence, we will only highlight any major differences with the isotropic case when discussing anisotropic analogs of pointwise, L_2-and L_∞-posterior convergence results from previous sections. Here, the order of each B-spline component at every direction is assumed to be $q_k \geq \alpha_k$ for $k = 1, \ldots, d$. Also, the prior support of f in this case is a subspace of $\mathcal{H}^{\boldsymbol{\alpha}}((0,1)^d)$.

By equation (12.37) of Theorem 12.7 in Schumaker (2007), for $f_0 \in \mathcal{H}^{\alpha}((0,1)^d)$, there exists a $\boldsymbol{\theta}_\infty \in \mathbb{R}^J$ such that for constant $C_{\boldsymbol{\alpha},\boldsymbol{q},d} > 0$ depending only on $\boldsymbol{\alpha}$, $\boldsymbol{q}$ and d with $R > 0$,

$$\sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \leq R} \|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\cdot)^T \boldsymbol{\theta}_\infty - f_0\|_\infty \leq \sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \leq R} C_{\boldsymbol{\alpha},\boldsymbol{q},d} \sum_{k=1}^{d} J_k^{-\alpha_k} \left\| \frac{\partial^{\alpha_k}}{\partial x_k^{\alpha_k}} f_0 \right\|_\infty \lesssim \sum_{k=1}^{d} J_k^{-\alpha_k}. \quad (3.75)$$

Since $\|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\cdot)^T \boldsymbol{\theta}_\infty\|_\infty \leq \|f_0\|_{\boldsymbol{\alpha},\infty} + C_{\boldsymbol{\alpha},\boldsymbol{q},d} \sum_{k=1}^{d} J_k^{-\alpha_k} \|f_0\|_{\boldsymbol{\alpha},\infty} \leq R + d$ uniformly in $\|f_0\|_{\boldsymbol{\alpha},\infty} \leq R$,

$$\sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \leq R} \|\boldsymbol{\theta}_\infty\|_\infty \lesssim \sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \leq R} \|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\cdot)^T \boldsymbol{\theta}_\infty\|_\infty = O(1) \quad (3.76)$$

by (12.25) of Theorem 12.5 in Schumaker (2007). The next two theorems are anisotropic analogs of Theorems 3.1 and 3.2 respectively, and their proofs are almost identical.

Theorem 3.9. *If $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, and $\alpha^* > d/2$, we have for $R > 0$,*

$$\sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \leq R} \mathrm{E}_0(\widehat{\sigma}_n^2 - \sigma_0^2)^2 = O\left(\frac{1}{n}\right). \quad (3.77)$$

Theorem 3.10. *If $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$ and $\alpha^* > d/2$, then for $R > 0$ and any $M_n \to \infty$,*

$$\sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \leq R} \mathrm{E}_0 \Pi\left(|\sigma^2 - \sigma_0^2| > \frac{M_n}{\sqrt{n}} \middle| \boldsymbol{Y}\right) \to 0. \quad (3.78)$$

The three results below are anisotropic analogs of Theorems 3.3, 3.4 and 3.5 respectively.

Theorem 3.11. *If $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, we have for $R > 0$, any $\boldsymbol{x} \in (0,1)^d$ and any $M_n \to \infty$,*

$$\textit{Empirical Bayes:} \quad \sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \leq R} \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(|f(\boldsymbol{x}) - f_0(\boldsymbol{x})| > M_n n^{-\alpha^*/(2\alpha^*+d)} | \boldsymbol{Y}) \to 0.$$

$$\textit{Hierarchical Bayes:} \quad \sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \leq R} \mathrm{E}_0 \Pi(|f(\boldsymbol{x}) - f_0(\boldsymbol{x})| > M_n n^{-\alpha^*/(2\alpha^*+d)} | \boldsymbol{Y}) \to 0.$$

Proof. Aside from the added technicalities of working with different smoothness, the posterior rate calculations follow closely with the steps detailed in the proof of Theorem 3.3, but with

$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^T \in \mathbb{N}^d$. At $\boldsymbol{x} \in (0,1)^d$, the expressions for both $\sup_{\sigma^2 \in \mathcal{U}_n} \sigma^2 \Sigma(\boldsymbol{x}, \boldsymbol{x})$ and $\sigma_0^2 \Psi(\boldsymbol{x}, \boldsymbol{x})$ in (3.37) and (3.38) will be $O(n^{-1} \prod_{k=1}^d J_k)$ respectively. To bound the squared bias $|\boldsymbol{A}(\boldsymbol{x})\boldsymbol{Y} + \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\eta} - f_0(\boldsymbol{x})|^2$ in (3.35) uniformly in $\|f_0\|_{\boldsymbol{\alpha},\infty} \leq R$, we use instead (3.75) to write

$$\begin{aligned}
&|\boldsymbol{A}(\boldsymbol{x})\boldsymbol{F}_0 + \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\eta} - f_0(\boldsymbol{x})| \\
&\leq \left\| \left(\boldsymbol{B}^T \boldsymbol{B} + \boldsymbol{\Omega}^{-1} \right)^{-1} \right\|_{(\infty,\infty)} \left\{ \|\boldsymbol{B}^T (\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_\infty)\|_\infty + \left\| \boldsymbol{\Omega}^{-1} \right\|_{(\infty,\infty)} (\|\boldsymbol{\theta}_\infty\|_\infty + \|\boldsymbol{\eta}\|_\infty) \right\} \\
&\quad + O\left(\sum_{k=1}^d J_k^{-\alpha_k} \right). \qquad (3.79)
\end{aligned}$$

By following the same argument of bounding the norms above, the anisotropic analogs of (3.40) and (3.41) will be

$$\sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \leq R} |\boldsymbol{A}(\boldsymbol{x})\boldsymbol{Y} + \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\eta} - f_0(\boldsymbol{x})|^2 \lesssim \frac{1}{n^2} \prod_{k=1}^d J_k^2 + \sum_{k=1}^d J_k^{-2\alpha_k} \qquad (3.80)$$

and

$$\begin{aligned}
\sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \leq R} \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}\left[(f(\boldsymbol{x}) - f_0(\boldsymbol{x}))^2 | \boldsymbol{Y} \right] &\lesssim \frac{1}{n} \prod_{k=1}^d J_k + \frac{1}{n^2} \prod_{k=1}^d J_k^2 + \sum_{k=1}^d J_k^{-2\alpha_k} \\
&\lesssim \frac{1}{n} \prod_{k=1}^d J_k + \sum_{k=1}^d J_k^{-2\alpha_k} \qquad (3.81)
\end{aligned}$$

respectively, where the last line follows since $\prod_{k=1}^d J_k \leq n$ by assumption. To balance the orders of the two terms on the right, let $J_k = J^{1/\alpha_k}$ for $k = 1, \ldots, d$. Then the right hand side of (3.81) reduces to

$$O\left(\frac{J^{\sum_{k=1}^d (1/\alpha_k)}}{n} \right) + O(J^{-2}). \qquad (3.82)$$

They will have the same order if $J \asymp n^{\alpha^*/(2\alpha^*+d)}$, and $J_k = J^{1/\alpha_k} \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$. This implies that the posterior mean square error in (3.81) will have order $n^{-2\alpha^*/(2\alpha^*+d)}$. The empirical and hierarchical Bayes posterior convergence rates then follow from (3.44) and (3.45) respectively. □

Theorem 3.12. *If $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, we have for $R > 0$ and any $M_n \to \infty$,*

$$\textit{Empirical Bayes:} \quad \sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \le R} \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(\|f - f_0\|_2 > M_n n^{-\alpha^*/(2\alpha^*+d)} | \boldsymbol{Y}) \to 0.$$

$$\textit{Hierarchical Bayes:} \quad \sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \le R} \mathrm{E}_0 \Pi(\|f - f_0\|_2 > M_n n^{-\alpha^*/(2\alpha^*+d)} | \boldsymbol{Y}) \to 0.$$

Proof. Following the steps in (3.46), the bound is given by (3.81). Set $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, to balance the orders of the two terms. The result then follows from Markov's inequality. □

Theorem 3.13. *If $J_k \asymp (n/\log n)^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, we have for $R > 0$ and any $M_n \to \infty$,*

$$\textit{Empirical Bayes:} \quad \sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \le R} \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(\|f - f_0\|_\infty > M_n (\log n/n)^{\alpha^*/(2\alpha^*+d)} | \boldsymbol{Y}) \to 0.$$

$$\textit{Hierarchical Bayes:} \quad \sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \le R} \mathrm{E}_0 \Pi(\|f - f_0\|_\infty > M_n (\log n/n)^{\alpha^*/(2\alpha^*+d)} | \boldsymbol{Y}) \to 0.$$

Proof. The proof is identical to the proof of Theorem 3.5, with the exception that the last squared bias term in (3.47) is bounded using (3.80), i.e.,

$$\sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \le R} \|\boldsymbol{A}\boldsymbol{F}_0 + \boldsymbol{c}\boldsymbol{\eta} - f_0\|_\infty^2 \lesssim \frac{1}{n^2} \prod_{k=1}^d J_k^2 + \sum_{k=1}^d J_k^{-2\alpha_k}. \tag{3.83}$$

By Lemma B.4 with $\boldsymbol{r} = \boldsymbol{0}$ (see the associated remarks), Z_n and Q_n satisfy the condition for Lemma C.1. Then the anisotropic analog of (3.52) is

$$\sup_{\|f_0\|_{\boldsymbol{\alpha},\infty} \le R} \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}(\|f - f_0\|_\infty^2 | \boldsymbol{Y}) \lesssim \frac{\log n}{n} \prod_{k=1}^d J_k + \sum_{k=1}^d J_k^{-2\alpha_k}. \tag{3.84}$$

To balance the orders of the two terms on the right, let $J_k = J^{1/\alpha_k}$ for $k = 1, \ldots, d$. Then the right hand side of (3.84) reduces to

$$O\left(\frac{J^{\sum_{k=1}^d (1/\alpha_k)}}{n} \log n \right) + O(J^{-2}).$$

Both terms will have the same order if $J \asymp (n/\log n)^{\alpha^*/(2\alpha^*+d)}$. This implies that $J_k = J^{1/\alpha_k} \asymp (n/\log n)^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, yielding the rate $(\log n/n)^{2\alpha^*/(2\alpha^*+d)}$ for the posterior mean square error in (3.84). The empirical and hierarchical Bayes posterior convergence rates then follow from (3.44) and (3.45) respectively with absolute values replaced by sup-norms. □

3.5.1 Anisotropic posterior convergence rates for mixed partial derivatives

The basic setup is identical to Section 3.4 with the exception that $\sum_{k=1}^{d}(r_k/\alpha_k) < 1$. The posterior convergence rates below are uniform in $\|D^{\boldsymbol{r}} f_0\|_{\boldsymbol{\alpha},\infty} \leq R$ for $R > 0$.

Theorem 3.14. *If $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, we have for any $\boldsymbol{x} \in (0,1)^d$ and any $M_n \to \infty$,*

$$\textit{Empirical Bayes:} \quad \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(|D^{\boldsymbol{r}} f(\boldsymbol{x}) - D^{\boldsymbol{r}} f_0(\boldsymbol{x})| > M_n n^{-\alpha^*\{1-\sum_{k=1}^{d}(r_k/\alpha_k)\}/(2\alpha^*+d)} | \boldsymbol{Y}) \to 0.$$

$$\textit{Hierarchical Bayes:} \quad \mathrm{E}_0 \Pi(|D^{\boldsymbol{r}} f(\boldsymbol{x}) - D^{\boldsymbol{r}} f_0(\boldsymbol{x})| > M_n n^{-\alpha^*\{1-\sum_{k=1}^{d}(r_k/\alpha_k)\}/(2\alpha^*+d)} | \boldsymbol{Y}) \to 0.$$

Proof. The proof is identical to the proof of Theorem 3.6, with the exception that the tensor B-spline approximation error in (3.75) but for $D^{\boldsymbol{r}} f_0$ is of the order $\sum_{k=1}^{d} J_k^{-(\alpha_k - r_k)}$, as a result of applying Lemma B.8. Therefore, the derivative analogs of (3.80) and (3.81) are given by

$$\sup_{\|D^{\boldsymbol{r}} f_0\|_{\boldsymbol{\alpha},\infty} \leq R} |\boldsymbol{A}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{Y} + \boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{\eta} - D^{\boldsymbol{r}} f_0(\boldsymbol{x})|^2 \lesssim \prod_{k=1}^{d} J_k^{2r_k} \left(\frac{1}{n^2} \prod_{k=1}^{d} J_k^2 + \sum_{k=1}^{d} J_k^{-2\alpha_k} \right)$$

and uniformly over $\|D^{\boldsymbol{r}} f_0\|_{\boldsymbol{\alpha},\infty} \leq R$,

$$\begin{aligned} \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}\left[(D^{\boldsymbol{r}} f(\boldsymbol{x}) - D^{\boldsymbol{r}} f_0(\boldsymbol{x}))^2 | \boldsymbol{Y} \right] &\lesssim \left(\frac{1}{n} \prod_{k=1}^{d} J_k^{2r_k+1} \right) + \prod_{k=1}^{d} J_k^{2r_k} \left(\frac{1}{n^2} \prod_{k=1}^{d} J_k^2 + \sum_{l=1}^{d} J_k^{-2\alpha_k} \right) \\ &\lesssim \prod_{k=1}^{d} J_k^{2r_k} \left(\frac{1}{n} \prod_{k=1}^{d} J_k + \sum_{k=1}^{d} J_k^{-2\alpha_k} \right) \end{aligned} \tag{3.85}$$

respectively, where the last line above follows from $\prod_{k=1}^{d} J_k \leq n$ by assumption. To balance the orders of the two terms on the right, let $J_k = J^{1/\alpha_k}$ for $k = 1, \ldots, d$. Then the right hand side of (3.85) reduces to

$$O\left(\frac{J^{\sum_{k=1}^{d}(2r_k+1)/\alpha_k}}{n} \right) + O(J^{2(\sum_{k=1}^{d} r_k/\alpha_k - 1)}). \tag{3.86}$$

They will have the same order if $J \asymp n^{\alpha^*/(2\alpha^*+d)}$, and $J_k = J^{1/\alpha_k} \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, and all $\boldsymbol{r} \in \mathbb{N}_0^d$ such that $\sum_{k=1}^{d}(r_k/\alpha_k) < 1$. Hence, the posterior mean square error in (3.85) will have order $n^{-2\alpha^*\{1-\sum_{k=1}^{d}(r_k/\alpha_k)\}/(2\alpha^*+d)}$. The empirical and hierarchical Bayes posterior convergence rates then follow from (3.44) and (3.45) respectively with f and f_0 replaced by $D^{\boldsymbol{r}} f$ and $D^{\boldsymbol{r}} f_0$. □

Theorem 3.15. *If $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, we have for any $M_n \to \infty$,*

$$\textit{Empirical Bayes:} \quad \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(\|D^{\boldsymbol{r}} f - D^{\boldsymbol{r}} f_0\|_2 > M_n n^{-\alpha^*\{1-\sum_{k=1}^d (r_k/\alpha_k)\}/(2\alpha^*+d)} | \boldsymbol{Y}) \to 0.$$

$$\textit{Hierarchical Bayes:} \quad \mathrm{E}_0 \Pi(\|D^{\boldsymbol{r}} f - D^{\boldsymbol{r}} f_0\|_2 > M_n n^{-\alpha^*\{1-\sum_{k=1}^d (r_k/\alpha_k)\}/(2\alpha^*+d)} | \boldsymbol{Y}) \to 0.$$

Proof. Following the steps in (3.46) with f and f_0 replaced by $D^{\boldsymbol{r}} f$ and $D^{\boldsymbol{r}} f_0$, the bound is given by (3.85). Set $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, to balance the orders of the two terms. The result then follows from Markov's inequality. □

Theorem 3.16. *If $J_k \asymp (n/\log n)^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, we have for any $M_n \to \infty$,*

$$\textit{Empirical Bayes:} \ \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(\|D^{\boldsymbol{r}} f - D^{\boldsymbol{r}} f_0\|_\infty > M_n (\log n/n)^{\alpha^*\{1-\sum_{k=1}^d (r_k/\alpha_k)\}/(2\alpha^*+d)} | \boldsymbol{Y}) \to 0.$$

$$\textit{Hierarchical Bayes:} \ \mathrm{E}_0 \Pi(\|D^{\boldsymbol{r}} f - D^{\boldsymbol{r}} f_0\|_\infty > M_n (\log n/n)^{\alpha^*\{1-\sum_{k=1}^d (r_k/\alpha_k)\}/(2\alpha^*+d)} | \boldsymbol{Y}) \to 0.$$

Proof. The proof is identical to the proof of Theorem 3.8 with the exception that we use (3.80) to bound the squared bias, i.e.,

$$\sup_{\|D^{\boldsymbol{r}} f_0\|_{\boldsymbol{\alpha},\infty} \le R} \|\boldsymbol{A}_{\boldsymbol{r}} \boldsymbol{F}_0 + \boldsymbol{c}_{\boldsymbol{r}} \boldsymbol{\eta} - D^{\boldsymbol{r}} f_0\|_\infty^2 \lesssim \prod_{k=1}^d J_k^{2r_k} \left(n^{-2} \prod_{k=1}^d J_k^2 + \sum_{k=1}^d J_k^{-2\alpha_k} \right).$$

By Lemma B.4 (see the associated remarks) with $\boldsymbol{r} \in \mathbb{N}_0^d$ and $\sum_{k=1}^d (r_k/\alpha_k) < 1$, $Z_{n,\boldsymbol{r}}$ and $Q_{n,\boldsymbol{r}}$ satisfy the condition for Lemma C.1. Then the anisotropic analog of (3.73) is given by

$$\sup_{\|D^{\boldsymbol{r}} f_0\|_{\boldsymbol{\alpha},\infty} \le R} \mathrm{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathrm{E}(\|D^{\boldsymbol{r}} f - D^{\boldsymbol{r}} f_0\|_\infty^2 | \boldsymbol{Y}) \lesssim \prod_{k=1}^d J_k^{2r_k} \left(\frac{\log n}{n} \prod_{k=1}^d J_k + \sum_{k=1}^d J_k^{-2\alpha_k} \right). \quad (3.87)$$

To balance the orders of the two terms on the right, let $J_k = J^{1/\alpha_k}$ for $k = 1, \ldots, d$. Then the right hand side of (3.87) reduces to

$$O\left(\frac{J^{\sum_{k=1}^d (2r_k+1)/\alpha_k}}{n} \log n \right) + O(J^{2(\sum_{k=1}^d r_k/\alpha_k - 1)}).$$

Both terms will have the same order if $J \asymp (n/\log n)^{\alpha^*/(2\alpha^*+d)}$. This implies that $J_k = J^{1/\alpha_k} \asymp (n/\log n)^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$ and all $\boldsymbol{r} \in \mathbb{N}_0^d$ such that $\sum_{k=1}^d (r_k/\alpha_k) < 1$, yielding the rate $(\log n/n)^{2\alpha^*\{1-\sum_{k=1}^d (r_k/\alpha_k)\}/(2\alpha^*+d)}$ for the posterior mean square error in (3.87). The empirical and hierarchical Bayes posterior convergence rates then follow from (3.44) and (3.45) respectively with absolute values replaced by sup-norms and f, f_0 replaced by $D^{\boldsymbol{r}} f$, $D^{\boldsymbol{r}} f_0$. □

If $\alpha_k = \alpha$ for $k = 1, \ldots, d$, then $\alpha^* = \alpha$. This implies that the pointwise, L_2 and L_∞-

posterior convergence rates established in this section will reduce to its isotropic counterparts of Sections 3.3 and 3.4.

3.6 Credible sets for f and its mixed partial derivatives

In this section, we will treat f and all its mixed partial derivatives in a unified framework by viewing f as $D^{\mathbf{0}}f$. Then if we define $\boldsymbol{W}_{\mathbf{0}} = \boldsymbol{I}_J$, we will have $\boldsymbol{A}_{\mathbf{0}} = \boldsymbol{A}, \boldsymbol{c}_{\mathbf{0}} = \boldsymbol{c}$ and $\Sigma_{\mathbf{0}} = \Sigma$ where $\boldsymbol{A}_{\boldsymbol{r}}, \boldsymbol{c}_{\boldsymbol{r}}$ and $\Sigma_{\boldsymbol{r}}$ were given by (3.59), (3.60) and (3.61) respectively. We begin by constructing pointwise credible set for $D^{\boldsymbol{r}}f(\boldsymbol{x})$ at $\boldsymbol{x} \in (0,1)^d$. Let $\omega_n \in [0,1]$ be a sequence such that $\omega_n \to 0$ as $n \to \infty$. Define z_δ to be the upper δ-quantile of a standard normal. Using the fact that $P(Z > z) \le z^{-1}\exp(-z^2/2)$ for $Z \sim \mathrm{N}(0,1)$ and $z > 0$, we then have

$$z_{\omega_n/2} = O(\sqrt{-\log \omega_n}). \tag{3.88}$$

In particular, if $\omega_n \gtrsim 1/n$, then $z_{\omega_n/2} = O(\sqrt{\log n})$; and if $\omega_n \gtrsim 1/\log n$, then $z_{\omega_n/2} = O(\sqrt{\log\log n})$. Since $\Pi(D^{\boldsymbol{r}}f(\boldsymbol{x})|\boldsymbol{Y}, \sigma^2) \sim \mathrm{N}(\boldsymbol{A}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{Y} + \boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{\eta}, \sigma^2\Sigma_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{x}))$, we can construct a $1-\omega_n$ pointwise credible interval for $D^{\boldsymbol{r}}f(\boldsymbol{x})$ at any $\boldsymbol{x} \in (0,1)^d$ as

$$\Pi(f : |D^{\boldsymbol{r}}f(\boldsymbol{x}) - \boldsymbol{A}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{\eta}| \le z_{\omega_n/2}\sigma\sqrt{\Sigma_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{x})}|\boldsymbol{Y}, \sigma^2) = 1 - \omega_n. \tag{3.89}$$

Since σ^2 is unknown, we will consider the empirical Bayes approach by substituting σ^2 for $\widehat{\sigma}_n^2$ derived in (3.13), leading to the following empirical credible set:

$$\widehat{\mathcal{C}}(\boldsymbol{x}) = \{f : |D^{\boldsymbol{r}}f(\boldsymbol{x}) - \boldsymbol{A}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{\eta}| \le z_{\omega_n/2}\widehat{\sigma}_n\sqrt{\Sigma_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{x})}\}, \tag{3.90}$$

where by construction, $\Pi_{\widehat{\sigma}_n}(\widehat{\mathcal{C}}(\boldsymbol{x})|\boldsymbol{Y}) = 1 - \omega_n$. We shall study the frequentist coverage of $\widehat{\mathcal{C}}(\boldsymbol{x})$ at any $\boldsymbol{x} \in (0,1)^d$ as $n \to \infty$, which is given by

$$\inf_{\|D^{\boldsymbol{r}}f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R} P_0(f_0 \in \widehat{\mathcal{C}}(\boldsymbol{x})). \tag{3.91}$$

Theorem 3.17. *If $J_k \asymp n^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, the asymptotic coverage of* (3.91) *is* 1 *and the radius of $\widehat{\mathcal{C}}(\boldsymbol{x})$ is $O_{P_0}(n^{-(\alpha-|\boldsymbol{r}|)/(2\alpha+d)}\sqrt{-\log\omega_n})$, uniformly over $\|D^{\boldsymbol{r}}f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R$ at any $\boldsymbol{x} \in (0,1)^d$, $R > 0$ and for all $|\boldsymbol{r}| < \alpha$.*

Proof. For any sequence $\zeta_n = o(\sqrt{n})$, define $t_{\boldsymbol{r}}(\boldsymbol{x}) = \inf_{|\sigma^2 - \sigma_0^2| \le \zeta_n/\sqrt{n}} z_{\omega_n/2}\sigma\sqrt{\Sigma_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{x})}$. By (3.43), to show (3.91) has asymptotic coverage of 1, it suffices to show that

$$\sup_{\|D^{\boldsymbol{r}}f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R} P_0\left(|D^{\boldsymbol{r}}f_0(\boldsymbol{x}) - \boldsymbol{A}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{\eta}| > t_{\boldsymbol{r}}(\boldsymbol{x})\right) \to 0. \tag{3.92}$$

Since $z_{\omega_n/2} \to \infty$ as $\omega_n \to 0$ and $\inf_{|\sigma^2-\sigma_0^2|\le\zeta_n/\sqrt{n}} \sigma^2 = \sigma_0^2 + o(1)$, we have $t_{\boldsymbol{r}}(\boldsymbol{x})^2 \gg \Sigma_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{x})$. By (3.18), $\Sigma_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{x}) \ge \lambda_{\min}\{(\boldsymbol{B}^T\boldsymbol{B}+\boldsymbol{\Omega}^{-1})^{-1}\}\|\boldsymbol{W}_{\boldsymbol{r}}^T\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})\|^2 \gtrsim n^{-1}\prod_{k=1}^d J_k\|\boldsymbol{W}_{\boldsymbol{r}}^T\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})\|^2$. For any $\boldsymbol{x} = (x_1,\ldots,x_d)^T \in (0,1)^d$, let i_{x_k} be a positive integer such that $x_k \in [t_{k,i_{x_k}-1}, t_{k,i_{x_k}}]$, then it follows that only $(B_{i_{x_k},q_k-r_k}(x_k),\ldots,B_{i_{x_k}+q_k-r_k-1,q_k-r_k}(x_k))^T$ are nonzero at each $k = 1,\ldots,d$. By (2.60)–(2.63), $(\partial^{r_k}/\partial x_k^{r_k})B_{j_k,q_k}(x_k)$ can be written as a linear combination of $(B_{j_k,q_k-r_k}(x_k),\ldots,B_{j_k+r_k,q_k-r_k}(x_k))^T$ for any $1 \le j_k \le J_k$ with $k = 1,\ldots,d$. Choose $j_k = i_{x_k}+q_k-r_k-1$ for $k = 1,\ldots,d$, and we have

$$\begin{aligned}
\|\boldsymbol{W}_{\boldsymbol{r}}^T\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})\|^2 &= \sum_{j_1=1}^{J_1}\cdots\sum_{j_d=1}^{J_d}\prod_{k=1}^d\left(\frac{\partial^{r_k}}{\partial x_k^{r_k}}B_{j_k,q_k}(x_k)\right)^2 \\
&\ge \prod_{k=1}^d\prod_{u=1}^{r_k}\frac{(q_k-u)^2}{(t_{k,i_{x_k}+q_k-r_k-1}-t_{k,i_{x_k}-r_k-1+u})^2}B_{i_{x_k}+q_k-r_k-1,q_k-r_k}(x_k)^2 \\
&\ge \prod_{k=1}^d\frac{1}{\Delta_k^{2r_k}}\left(\frac{q_k-r_k}{q_k}\right)^2 \gtrsim \prod_{k=1}^d J_k^{2r_k}, \qquad (3.93)
\end{aligned}$$

where the last line follows since (3.4) implies that $t_{k,i_{x_k}+q_k-r_k-1}-t_{k,i_{x_k}-r_k-1+u} \le (q_k-u+1)\Delta_k$ for $k = 1,\ldots,d$, and we used Lemma B.1 to obtain the final lower bound. Consequently, $t_{\boldsymbol{r}}(\boldsymbol{x})^2 \gg n^{-1}\prod_{k=1}^d J_k^{2r_k+1}$.

Under the true distribution P_0, $\boldsymbol{A}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{Y}+\boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{\eta}$ is a sub-Gaussian random variable with mean $\boldsymbol{A}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{F}_0+\boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{\eta}$ and variance $\sigma_0^2\Psi_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{x})$, where $\Psi_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{x})$ at $\boldsymbol{x} \in (0,1)^d$ is given by (3.66). Define $Q_{n,\boldsymbol{r}}(\boldsymbol{x})$ to be a sub-Gaussian random variable of mean 0 and variance $\sigma_0^2\Psi_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{x})$. Then uniformly in $\|D^{\boldsymbol{r}}f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R$,

$$\begin{aligned}
\mathrm{E}_0|D^{\boldsymbol{r}}f_0(\boldsymbol{x})-\boldsymbol{A}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{Y}-\boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{\eta}|^2 &= \mathrm{E}[\boldsymbol{A}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{F}_0+\boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{\eta}-D^{\boldsymbol{r}}f_0(\boldsymbol{x})+Q_{n,\boldsymbol{r}}(\boldsymbol{x})]^2 \\
&\le 2\mathrm{E}_0[Q_{n,\boldsymbol{r}}(\boldsymbol{x})^2]+2|\boldsymbol{A}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{F}_0+\boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{\eta}-D^{\boldsymbol{r}}f_0(\boldsymbol{x})|^2 \\
&\lesssim \frac{1}{n}\prod_{k=1}^d J_k^{2r_k+1}+\sum_{k=1}^d J_k^{-2(\alpha-r_k)},
\end{aligned}$$

where the last line follows from (3.68) and (3.63) respectively. By Markov's inequality, we have uniformly over $\|D^{\boldsymbol{r}}f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R$ that

$$P_0(|D^{\boldsymbol{r}}f_0(\boldsymbol{x})-\boldsymbol{A}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{Y}-\boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{\eta}| > t_{\boldsymbol{r}}(\boldsymbol{x})) \lesssim \frac{n^{-1}\prod_{k=1}^d J_k^{2r_k+1}+\sum_{k=1}^d J_k^{-2(\alpha-r_k)}}{t_{\boldsymbol{r}}(\boldsymbol{x})^2}.$$

Since $t_{\boldsymbol{r}}(\boldsymbol{x})^2 \gg n^{-1}\prod_{k=1}^d J_k^{2r_k+1}$, we have $t_{\boldsymbol{r}}(\boldsymbol{x})^2 \gg \sum_{k=1}^d J_k^{-2(\alpha-r_k)}$ if $J_k \asymp n^{1/(2\alpha+d)}$ for $k = 1,\ldots,d$. Hence the right hand side above will go to zero as $n \to \infty$. We conclude

that $\inf_{\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R} P_0(|D^{\boldsymbol{r}} f_0(\boldsymbol{x}) - \boldsymbol{A}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{\eta}| \le t_{\boldsymbol{r}}(\boldsymbol{x})) \to 1$ if $J_k \asymp n^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$. Let $h_{\boldsymbol{r}}(\boldsymbol{x}) = z_{\omega_n/2}\sqrt{\Sigma_{\boldsymbol{r}}(\boldsymbol{x}, \boldsymbol{x})}$ at $\boldsymbol{x} \in (0,1)^d$. By (3.88), we have $z_{\omega_n/2} = O(\sqrt{-\log \omega_n})$. Since $\Sigma_{\boldsymbol{r}}(\boldsymbol{x}, \boldsymbol{x}) \lesssim n^{-1} \prod_{k=1}^d J_k^{2r_k+1}$ by (3.67), we have by (3.43) that uniformly over $\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R$, $\widehat{\sigma}_n h_{\boldsymbol{r}}(\boldsymbol{x}) = O_{P_0}(n^{-(\alpha-|\boldsymbol{r}|)/(2\alpha+d)}\sqrt{-\log \omega_n})$ at any $\boldsymbol{x} \in (0,1)^d$. □

Remark For the anisotropic case, assuming that f_0 satisfies Assumption 3, then if $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, we have uniformly over $\|D^{\boldsymbol{r}} f_0\|_{\boldsymbol{\alpha},\infty} \le R$ that the asymptotic coverage of $\widehat{\mathcal{C}}(\boldsymbol{x})$ is 1, and its radius is $O_{P_0}(n^{-\alpha^*\{1-\sum_{k=1}^d (r_k/\alpha_k)\}/(2\alpha^*+d)}\sqrt{-\log \omega_n})$ for any $\boldsymbol{x} \in (0,1)^d$.

The conditional posterior distribution of $D^{\boldsymbol{r}} f$ is $\Pi(D^{\boldsymbol{r}} f|\boldsymbol{Y}, \sigma^2) \sim \mathrm{GP}(\boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} + \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}, \sigma^2\Sigma_{\boldsymbol{r}})$. Note that the conditional distribution of $(D^{\boldsymbol{r}} f - \boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta})/\sigma$ given $(\boldsymbol{Y}, \sigma^2)$ is $\mathrm{GP}(0, \Sigma_{\boldsymbol{r}})$, and hence does not depend on $\boldsymbol{Y}$ and σ^2. This enables us to construct L_2-credible ball for $D^{\boldsymbol{r}} f$, centered at its posterior mean with radius $\sigma h_{\boldsymbol{r}}$, where $h_{\boldsymbol{r}}$ is chosen such that its posterior mass is exactly $1 - \omega_n$. That is,

$$\Pi(f : \|D^{\boldsymbol{r}} f - \boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}\|_2 \le \sigma h_{\boldsymbol{r}}|\boldsymbol{Y}, \sigma^2) = 1 - \omega_n, \tag{3.94}$$

where $h_{\boldsymbol{r}} \to \infty$ as $\omega_n \to 0$ and $h_{\boldsymbol{r}}$ does not depend on $\boldsymbol{Y}$. The corresponding empirical L_2-credible set is

$$\widehat{\mathcal{C}}_2 = \{f : \|D^{\boldsymbol{r}} f - \boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}\|_2 \le \widehat{\sigma}_n h_{\boldsymbol{r}}\} \tag{3.95}$$

where by construction, $\Pi_{\widehat{\sigma}_n}(\widehat{\mathcal{C}}_2|\boldsymbol{Y}) = 1 - \omega_n$. We shall study the frequentist coverage of $\widehat{\mathcal{C}}_2$ as $n \to \infty$, which is given by

$$\inf_{\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R} P_0(f_0 \in \widehat{\mathcal{C}}_2). \tag{3.96}$$

Theorem 3.18. *If $J_k \asymp n^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, the asymptotic coverage of* (3.96) *is* 1 *and the radius of $\widehat{\mathcal{C}}_2$ is $O_{P_0}(n^{-(\alpha-|\boldsymbol{r}|)/(2\alpha+d)}\sqrt{-\log \omega_n})$, uniformly over $\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R$ for $R > 0$ and all $|\boldsymbol{r}| < \alpha$.*

Proof. For any sequence $\zeta_n = o(\sqrt{n})$, define $t_{\boldsymbol{r}} = \inf_{|\sigma^2 - \sigma_0^2| \le \zeta_n/\sqrt{n}} \sigma h_{\boldsymbol{r}}$. By (3.43), to show (3.96) has asymptotic coverage of 1, it suffices to show

$$\sup_{\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R} P_0\left(\|D^{\boldsymbol{r}} f_0 - \boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}\|_2 > t_{\boldsymbol{r}}\right) \to 0. \tag{3.97}$$

Let $Z_{n,\boldsymbol{r}} \sim \mathrm{GP}(0, \Sigma_{\boldsymbol{r}})$. Now $\Pi_{\widehat{\sigma}_n}(\widehat{\mathcal{C}}_2|\boldsymbol{Y}) = 1 - \omega_n$ implies that $P(\|Z_{n,\boldsymbol{r}}\|_2 > h_{\boldsymbol{r}}) \to 0$. Observe that $\int_{(0,1)^d} Z_{n,\boldsymbol{r}}(\boldsymbol{x})d\boldsymbol{x}$ is normally distributed, and by Fubini's theorem, its mean is 0 and vari-

ance is $\int_{(0,1)^d}\int_{(0,1)^d}\Sigma_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}$. Now since $\int_{(0,1)^d} Z_{n,\boldsymbol{r}}(\boldsymbol{x})d\boldsymbol{x} \le \|Z_{n,\boldsymbol{r}}\|_1 \le \|Z_{n,\boldsymbol{r}}\|_2$ a.s., we have

$$1-\Phi\left(\frac{h_{\boldsymbol{r}}}{\sqrt{\int_{(0,1)^d}\int_{(0,1)^d}\Sigma_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}}}\right) \le P(\|Z_{n,\boldsymbol{r}}\|_2 > h_{\boldsymbol{r}}) \to 0. \tag{3.98}$$

Hence, $h_{\boldsymbol{r}}^2 \gg \int_{(0,1)^d}\int_{(0,1)^d}\Sigma_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}$. To derive a lower bound for $\int_{(0,1)^d}\int_{(0,1)^d}\Sigma_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}$, we first need to establish lower bounds for integrals of tensor-product B-splines, and this is done by extending the last two equations in the proof of Lemma 6.7 in Zhou et al. (1998) to higher dimensions. Given $\boldsymbol{x} = (x_1,\ldots,x_d)^T$ such that $x_k \in [t_{k,j_k-1}, t_{j_k}]$ for some $j_k = 1,\ldots,N_k+1$, with $k = 1,\ldots,d$, we have by repeated applications of the B-spline recursive formula at each dimension (see equation (4.22) of Schumaker, 2007),

$$\begin{aligned}\prod_{k=1}^d B_{j_k,q_k}(x_k) &= \prod_{k=1}^d (t_{k,j_k}-x_k)^{q_k-1}\prod_{l_k=1}^{q_k-1}(t_{k,j_k}-t_{k,j_k+l_k-q_k+1})^{-1}\\ &\ge \prod_{k=1}^d \frac{\Delta_k^{-q_k+1}(t_{k,j_k}-x_k)^{q_k-1}}{(q_k-1)!},\end{aligned}$$

where $\Delta_k = \max_{1\le l\le N_k}(t_{k,l}-t_{k,l-1})$ is the maximum knot increment at direction $k = 1,\ldots,d$. Therefore, by the quasi-uniform knot assumption in (3.4),

$$\begin{aligned}\int_{(0,1)^d}\prod_{k=1}^d B_{j_k,q_k}(x_k)d\boldsymbol{x} &= \sum_{j_d=1}^{N_d+1}\cdots\sum_{j_1=1}^{N_1+1}\int_{t_{d,j_d-1}}^{t_{d,j_d}}\cdots\int_{t_{1,j_1-1}}^{t_{1,j_1}}\prod_{k=1}^d B_{j_k,q_k}(x_k)dx_1\cdots dx_d\\ &\ge \prod_{k=1}^d N_k\left(\frac{\Delta_k}{C}\right)^{q_k}\frac{\Delta_k^{-q_k+1}}{q_k!}\\ &\ge \left(\prod_{k=1}^d C^{q_k}q_k!\right)^{-1},\end{aligned} \tag{3.99}$$

where the constant $C > 0$ is from (3.4). Using the expression for $\Sigma_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{y})$ at $\boldsymbol{x},\boldsymbol{y}\in(0,1)^d$ given

in (3.61),

$$\begin{aligned}
&\int_{(0,1)^d}\int_{(0,1)^d} \Sigma_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} \\
&\quad= \left(\int_{(0,1)^d} \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})d\boldsymbol{x}\right)^T \boldsymbol{W}_{\boldsymbol{r}}(\boldsymbol{B}^T\boldsymbol{B}+\boldsymbol{\Omega}^{-1})^{-1}\boldsymbol{W}_{\boldsymbol{r}}^T \left(\int_{(0,1)^d} \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})d\boldsymbol{x}\right) \\
&\quad\geq \lambda_{\min}\left\{(\boldsymbol{B}^T\boldsymbol{B}+\boldsymbol{\Omega}^{-1})^{-1}\right\}\left\|\boldsymbol{W}_{\boldsymbol{r}}^T\left(\int_{(0,1)^d} \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})d\boldsymbol{x}\right)\right\|^2 \\
&\quad\gtrsim \frac{1}{n}\prod_{k=1}^d J_k \left\|\boldsymbol{W}_{\boldsymbol{r}}^T\left(\int_{(0,1)^d} \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})d\boldsymbol{x}\right)\right\|^2, \qquad (3.100)
\end{aligned}$$

where the last line follows from (3.18). Now using expressions for the entries of $\boldsymbol{W}_{\boldsymbol{r}}$ described in (3.55)–(3.58) and the lower bound in (3.99),

$$\begin{aligned}
\left\|\boldsymbol{W}_{\boldsymbol{r}}^T\left(\int_{(0,1)^d} \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{x})d\boldsymbol{x}\right)\right\|^2 &\geq \prod_{k=1}^d\prod_{l=1}^{r_k}\left(\frac{q_k-l}{t_{k,1}-t_{k,1-q_k+l}}\right)^2\left(\int_{(0,1)^d}\prod_{k=1}^d B_{1,q_k-r_k}(x_k)d\boldsymbol{x}\right)^2 \\
&\geq \left(\prod_{k=1}^d C^{q_k}q_k!\right)^{-2}\prod_{k=1}^d\left\{\left(\frac{q_k-r_k}{q_k}\right)^2\frac{1}{\Delta^{2r_k}}\right\} \\
&\gtrsim \prod_{k=1}^d N_k^{2r_k} \asymp \prod_{k=1}^d J_k^{2r_k}.
\end{aligned}$$

The second last line follows since (3.4) implies that $t_{k,1}-t_{k,1-q_k+l} \leq (q_k-l+1)\Delta_k$ for $l=1,\ldots,r_k$ with $k=1,\ldots,d$, and we applied Lemma B.1 to obtain the final lower bound. In view of (3.100), we then have $\int_{(0,1)^d}\int_{(0,1)^d}\Sigma_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} \gtrsim n^{-1}\prod_{k=1}^d J_k^{2r_k+1}$. Therefore, $t_{\boldsymbol{r}}^2 \gtrsim \sigma_0^2 h_{\boldsymbol{r}}^2 \gg n^{-1}\prod_{k=1}^d J_k^{2r_k+1}$.

Under the true distribution P_0, $\boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y}+\boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}$ is a sub-Gaussian process with mean function $\boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{F}_0+\boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}$ and covariance function $\sigma_0^2\Psi_{\boldsymbol{r}}$, where the expression for $\Psi_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{y})$ at $\boldsymbol{x},\boldsymbol{y}\in(0,1)^d$ is given by (3.66). Define $Q_{n,\boldsymbol{r}}$ to be a sub-Gaussian process with mean function 0 and covariance

function $\sigma_0^2\Psi_{\boldsymbol{r}}$. Then uniformly in $\|D^{\boldsymbol{r}}f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R$,

$$\begin{aligned}
\mathrm{E}_0\|D^{\boldsymbol{r}}f_0 - \boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}\|_2^2 &= \mathrm{E}\|\boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{F}_0 + \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta} - D^{\boldsymbol{r}}f_0 + Q_{n,\boldsymbol{r}}\|_2^2 \\
&\le 2\mathrm{E}\|Q_{n,\boldsymbol{r}}\|_2^2 + 2\|\boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{F}_0 + \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta} - D^{\boldsymbol{r}}f_0\|_2^2 \\
&= 2\int_{(0,1)^d} \left[\Psi_{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{x}) + |\boldsymbol{A}_{\boldsymbol{r}}(x)\boldsymbol{F}_0 + \boldsymbol{c}_{\boldsymbol{r}}(\boldsymbol{x})\boldsymbol{\eta} - D^{\boldsymbol{r}}f_0(\boldsymbol{x})|^2\right] d\boldsymbol{x} \\
&\lesssim \frac{1}{n}\prod_{k=1}^d J_k^{2r_k+1} + \sum_{k=1}^d J_k^{-2(\alpha-r_k)},
\end{aligned}$$

where the last line follows from (3.68) and (3.63) respectively. By Markov's inequality,

$$\sup_{\|D^{\boldsymbol{r}}f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}}\le R} P_0(\|D^{\boldsymbol{r}}f_0 - \boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}\|_2 > t_{\boldsymbol{r}}) \lesssim \frac{n^{-1}\prod_{k=1}^d J_k^{2r_k+1} + \sum_{k=1}^d J_k^{-2(\alpha-r_k)}}{t_{\boldsymbol{r}}^2}.$$

Since $t_{\boldsymbol{r}}^2 \gg n^{-1}\prod_{k=1}^d J_k^{2r_k+1}$, we have $t_{\boldsymbol{r}}^2 \gg \sum_{k=1}^d J_k^{-2(\alpha-r_k)}$ if $J_k \asymp n^{1/(2\alpha+d)}$ for $k = 1,\ldots,d$. Hence the right hand side will go to zero as $n \to \infty$. Consequently, $\inf_{\|D^{\boldsymbol{r}}f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}}\le R} P_0(\|D^{\boldsymbol{r}}f_0 - \boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}\|_2 \le t_{\boldsymbol{r}}) \to 1$ if $J_k \asymp n^{1/(2\alpha+d)}$ for $k = 1,\ldots,d$. Applying (3.5) from Ledoux and Talagrand (1991) to $Z_{n,\boldsymbol{r}}$ with $\|\cdot\|$ as the L_2-norm, we have

$$\omega_n = P(\|Z_{n,\boldsymbol{r}}\|_2 > h_{\boldsymbol{r}}) \le 4\exp\left(-\frac{h_{\boldsymbol{r}}^2}{8\mathrm{E}\|Z_{n,\boldsymbol{r}}\|_2^2}\right).$$

It then follows that $h_{\boldsymbol{r}} \lesssim (\mathrm{E}\|Z_{n,\boldsymbol{r}}\|_2^2)^{1/2}\sqrt{-\log\omega_n}$. By Fubini's theorem and (3.37), we have $\mathrm{E}\|Z_{n,\boldsymbol{r}}\|_2^2 = \int_{(0,1)^d}\Sigma(\boldsymbol{x},\boldsymbol{x})d\boldsymbol{x} \lesssim n^{-1}\prod_{k=1}^d J_k^{2r_k+1}$. In view of (3.43), we have uniformly over $\|D^{\boldsymbol{r}}f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R$ that the radius of $\widehat{\mathcal{C}}_2$ is $\widehat{\sigma}_n h_{\boldsymbol{r}} = O_{P_0}(n^{-(\alpha-|\boldsymbol{r}|)/(2\alpha+d)}\sqrt{-\log\omega_n})$. □

Remark For the anisotropic case, assuming that f_0 satisfies Assumption 3, then if $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1,\ldots,d$, we have uniformly in $\|D^{\boldsymbol{r}}f_0\|_{\boldsymbol{\alpha},\infty} \le R$ that the asymptotic coverage of $\widehat{\mathcal{C}}_2$ is 1 and its radius is $O_{P_0}(n^{-\alpha^*\{1-\sum_{k=1}^d(r_k/\alpha_k)\}/(2\alpha^*+d)}\sqrt{-\log\omega_n})$.

We will use a slightly different approach to construct L_∞-credible region for $D^{\boldsymbol{r}}f$ of size $1-\omega$. Note that here $\omega \in [0,1]$ does not depend on n and we assume that $\omega < 1/2$. We choose a sequence $h_{\infty,\boldsymbol{r}}$ such that

$$\Pi_{\widehat{\sigma}_n}(f : \|D^{\boldsymbol{r}}f - \boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}\|_\infty \le \widehat{\sigma}_n h_{\infty,\boldsymbol{r}}|\boldsymbol{Y}) = 1-\omega, \tag{3.101}$$

where $h_{\infty,\boldsymbol{r}}$ does not depend on $\boldsymbol{Y}$. As we shall see in Theorem 3.19 below, in order to ensure frequentist coverage of 1 asymptotically, we need to introduce an inflation factor ρ_n in the radius, such that $\rho_n \to \infty$ as $n \to \infty$. That is, we will consider the following inflated credible

region

$$\widehat{\mathcal{C}}_\infty = \{f : \|D^{\boldsymbol{r}} f - \boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}\|_\infty \le \rho_n \widehat{\sigma}_n h_{\infty,\boldsymbol{r}}\}, \tag{3.102}$$

where by construction, $\Pi_{\widehat{\sigma}_n}(\widehat{\mathcal{C}}_\infty|\boldsymbol{Y}) \ge 1-\omega$ if $\rho_n \ge 1$. We shall study the frequentist coverage of $\widehat{\mathcal{C}}_\infty$, which is given by

$$\inf_{\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R} P_0(f_0 \in \widehat{\mathcal{C}}_\infty). \tag{3.103}$$

Theorem 3.19. *If $J_k \asymp (n/\log n)^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, the asymptotic coverage of* (3.103) *is* 1 *and the radius of $\widehat{\mathcal{C}}_\infty$ is $O_{P_0}((\log n/n)^{(\alpha-|\boldsymbol{r}|)/(2\alpha+d)}\rho_n)$, uniformly in $\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R$ for $R > 0$ and any $|\boldsymbol{r}| < \alpha$.*

Remark To simultaneously control the size of $\widehat{\mathcal{C}}_\infty$ and ensure guaranteed frequentist coverage, we can take ρ_n to be a slowly varying factor tending to infinity. For example, $\rho_n = (\log\log n)^{1/3}$ gives excellent results (see the simulation experiment in Section 2.6).

Proof. For any $\zeta_n = o(\sqrt{n})$, define $t_{\infty,\boldsymbol{r}} = \inf_{|\sigma^2-\sigma_0^2| \le \zeta_n/\sqrt{n}} \sigma h_{\infty,\boldsymbol{r}}$. By (3.43), to show (3.103) has asymptotic coverage of 1, it suffices to show

$$\sup_{\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R} P_0(\|D^{\boldsymbol{r}} f_0 - \boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}\|_\infty > \rho_n t_{\infty,\boldsymbol{r}}) \to 0. \tag{3.104}$$

Let $Z_{n,\boldsymbol{r}} \sim \mathrm{GP}(0, \Sigma_{\boldsymbol{r}})$. Define $M(Z_{n,\boldsymbol{r}})$ to be the median of $\|Z_{n,\boldsymbol{r}}\|_\infty$, i.e., $M(Z_{n,\boldsymbol{r}})$ is a number satisfying both $P(\|Z_{n,\boldsymbol{r}}\|_\infty \le M(Z_{n,\boldsymbol{r}})) \ge 1/2$ and $P(\|Z_{n,\boldsymbol{r}}\|_\infty \ge M(Z_{n,\boldsymbol{r}})) \ge 1/2$. In addition, we define $\sigma_Z^2 = \sup_{\boldsymbol{x}\in(0,1)^d} \mathrm{Var}(Z_{n,\boldsymbol{r}}(\boldsymbol{x}))$ and note that by (3.67), $\sigma_Z^2 \lesssim n^{-1}\prod_{k=1}^d J_k^{2r_k+1}$. Using the fact that $\sigma_Z \le 2M(Z_{n,\boldsymbol{r}})$ and $|\mathrm{E}\|Z_{n,\boldsymbol{r}}\|_\infty - M(Z_{n,\boldsymbol{r}})| \le \sigma_Z(\pi/2)^{1/2}$ (see Pages 52 and 54 of Ledoux and Talagrand, 1991), we have $\mathrm{E}\|Z_{n,\boldsymbol{r}}\|_\infty \le (1+2(\pi/2)^{1/2})M(Z_{n,\boldsymbol{r}})$. Since $\sigma_Z^2 = o(1)$ for $J_k \asymp (n/\log n)^{1/(2\alpha+1)}$ with $k = 1, \ldots, d$, we have $\mathrm{E}\|Z_{n,\boldsymbol{r}}\|_\infty \ge M(Z_{n,\boldsymbol{r}}) + o(1)$. Consequently, $\mathrm{E}\|Z_{n,\boldsymbol{r}}\|_\infty \asymp M(Z_{n,\boldsymbol{r}})$.

Now, (3.101) implies that $P(\|Z_{n,\boldsymbol{r}}\|_\infty > h_{\infty,\boldsymbol{r}}) = \omega$. Since $\omega < 1/2$ by assumption, this implies that $h_{\infty,\boldsymbol{r}} \ge M(Z_{n,\boldsymbol{r}}) \asymp \mathrm{E}\|Z_{n,\boldsymbol{r}}\|_\infty$. Hence, we need a lower bound for $\mathrm{E}\|Z_{n,\boldsymbol{r}}\|_\infty$. Define $\mathcal{T}_k = \{t_{k,1}, \ldots, t_{k,N_k}\}$ to be the sequence of N_k interior knot points of the kth component B-spline for $k = 1, \ldots, d$. Let $\mathcal{T} = \prod_{k=1}^d \mathcal{T}_k$ be the d-fold Cartesian product of the sets $\mathcal{T}_k$ for $k = 1, \ldots, d$. Define $\mathcal{I} = \{(i_1, \ldots, i_d) : 1 \le i_k \le N_k, k = 1, \ldots, d\}$ and we arrange the elements of $\mathcal{I}$ lexicographically. Then, we can enumerate the $N = \prod_{k=1}^d N_k$ elements of $\mathcal{T}$ as $\{\boldsymbol{\tau}_{\boldsymbol{i}} : \boldsymbol{i} \in \mathcal{I}\}$, where $\boldsymbol{\tau}_{\boldsymbol{i}} = (t_{1,i_1}, \ldots, t_{d,i_d})$ with $(i_1, \ldots, i_d) \in \mathcal{I}$. Define $u(x_1, \ldots, x_d) = \prod_{k=1}^d (\partial^{r_k}/\partial x_k^{r_k}) B_{j_k,q_k}(x_k)$. Applying the multivariate mean value theorem to $u(x_1, \ldots, x_d)$ at

points $\boldsymbol{\tau_i}$ and $\boldsymbol{\tau_m}$, we have for some point $\boldsymbol{\tau}^* = (t_1^*, \ldots, t_d^*) = \lambda\boldsymbol{\tau_i} + (1-\lambda)\boldsymbol{\tau_m}$ with $\lambda \in [0,1]$ that

$$\begin{aligned}\sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} |u(\boldsymbol{\tau_i}) - u(\boldsymbol{\tau_m})|^2 &= \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} |\nabla u(\boldsymbol{\tau}^*)^T(\boldsymbol{\tau_i} - \boldsymbol{\tau_m})|^2 \\ &= \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \left| \sum_{\gamma=1}^{d} \left(\frac{\partial u}{\partial x_\gamma} \right) (t_{\gamma,i_\gamma} - t_{\gamma,m_\gamma}) \right|^2 . \end{aligned} \tag{3.105}$$

For any $\boldsymbol{x} = (x_1, \ldots, x_d)^T \in (0,1)^d$, let i_{x_k} be a positive integer such that $x_k \in [t_{i_{x_k}-1}, t_{i_{x_k}}]$ for $k = 1, \ldots, d$. Then it follows that only $(B_{i_{x_k},q_k-r_k}(x_k) \ldots, B_{i_{x_k}+q_k-r_k-1,q_k-r_k}(x_k))^T$ are nonzero for $k = 1, \ldots, d$. By (2.60)–(2.63), $(\partial^{r_k}/\partial x_k^{r_k}) B_{j_k,q_k}(x_k)$ can be written as a linear combination of $(B_{j_k,q_k-r_k}(x_k), \ldots, B_{j_k+r_k,q_k-r_k}(x_k))^T$ for any $1 \le j_k \le J_k$ with $k = 1, \ldots, d$. Choose $j_1 = i_{x_1} + q_1 - r_1 - 2$ and $j_k = i_{x_k} + q_k - r_k - 1$ for $k = 2, \ldots, d$, then only $\partial u/\partial x_1$ will be positive while $\partial u/\partial x_k$ will be zero for $k = 2, \ldots, d$. Therefore,

$$\begin{aligned}\sum_{\gamma=1}^{d} \left(\frac{\partial u}{\partial x_\gamma} \right) (t_{\gamma,i_\gamma} - t_{\gamma,m_\gamma}) &= \sum_{\gamma=1}^{d} \frac{\partial^{r_\gamma+1}}{\partial x_\gamma^{r_\gamma+1}} B_{j_\gamma,q_\gamma}(x_\gamma) \prod_{k\neq\gamma}^{d} \left(\frac{\partial^{r_k}}{\partial x_k^{r_k}} B_{j_k,q_k}(x_k) \right) (t_{\gamma,i_\gamma} - t_{\gamma,m_\gamma}) \\ &= \frac{\partial^{r_1+1}}{\partial x_1^{r_1+1}} B_{j_1,q_1}(x_1) \prod_{k=2}^{d} \left(\frac{\partial^{r_k}}{\partial x_k^{r_k}} B_{j_k,q_k}(x_k) \right) (t_{1,i_1} - t_{1,m_1}). \end{aligned} \tag{3.106}$$

By repeated applications of (2.84),

$$\begin{aligned}\sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} |u(\boldsymbol{\tau}_i) - u(\boldsymbol{\tau}_m)|^2 &\ge \left(\frac{\partial^{r_1+1}}{\partial x_1^{r_1+1}} B_{j_1,q_1}(x_1) \right)^2 \prod_{k=2}^{d} \left(\frac{\partial^{r_k}}{\partial x_k^{r_k}} B_{j_k,q_k}(x_k) \right)^2 (t_{1,i_1} - t_{1,m_1})^2 \\ &\gtrsim J_1^{2r_1+2} \prod_{k=2}^{d} J_k^{2r_k} \left(\min_{1\le l\le N_1} \delta_{1,l}^2 \right) \gtrsim \prod_{k=1}^{d} J_k^{2r_k}, \end{aligned} \tag{3.107}$$

where $\delta_{1,l} = t_{1,l} - t_{1,l-1}$ for $1 \le l \le N_1$, and the last line follows from assumption (3.4) and Lemma B.1. Define $V_{\boldsymbol{i}} = Z_{n,\boldsymbol{r}}(\boldsymbol{\tau_i})$ for $\mathbf{1}_d \le \boldsymbol{i} \le \boldsymbol{N}$ where $\boldsymbol{N} = (N_1, \ldots, N_d)$. Then for any $\mathbf{1}_d \le \boldsymbol{i}, \boldsymbol{m} \le \boldsymbol{N}$,

$$\begin{aligned}\mathrm{E}(V_{\boldsymbol{i}} - V_{\boldsymbol{m}})^2 &= (\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{\tau_i}) - \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{\tau_m}))^T \boldsymbol{W}_{\boldsymbol{r}} (\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1})^{-1} \boldsymbol{W}_{\boldsymbol{r}}^T (\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{\tau_i}) - \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{\tau_m})) \\ &\ge \lambda_{\min}\{(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1})^{-1}\} \|\boldsymbol{W}_{\boldsymbol{r}}^T (\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{\tau_i}) - \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{\tau_m}))\|^2 \\ &\gtrsim \left(\frac{1}{n} \prod_{k=1}^{d} J_k \right) \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} |u(\boldsymbol{\tau_i}) - u(\boldsymbol{\tau_m})|^2 \ge c \left(\frac{1}{n} \prod_{k=1}^{d} J_k^{2r_k+1} \right), \end{aligned} \tag{3.108}$$

where $c > 0$ is a universal constant not depending on n, and we used (3.18) to lower bound the minimum eigenvalue of $(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1})^{-1}$. Define $U_{\boldsymbol{i}} = \sqrt{(2n/c)}\prod_{k=1}^d J_k^{-(r_k+1/2)} V_{\boldsymbol{i}}$ and $H_{\boldsymbol{i}}$ i.i.d. $\mathrm{N}(0,1)$ with $\mathbf{1}_d \le \boldsymbol{i} \le \boldsymbol{N}$. By equation (3.14) of Ledoux and Talagrand (1991), we have $\mathrm{E}(\max_{\mathbf{1}_d \le \boldsymbol{i} \le \boldsymbol{N}} H_{\boldsymbol{i}}) \gtrsim \sqrt{\log N}$. Therefore,

$$\mathrm{E}(U_{\boldsymbol{i}} - U_{\boldsymbol{m}})^2 \ge 2 = \mathrm{E}(H_{\boldsymbol{i}} - H_{\boldsymbol{m}})^2 \tag{3.109}$$

and by Slepian's Lemma (Corollary 3.14 of Ledoux and Talagrand, 1991), we have

$$\mathrm{E}\left(\max_{\mathbf{1}_d \le \boldsymbol{i} \le \boldsymbol{N}} U_{\boldsymbol{i}}\right) \ge \mathrm{E}\left(\max_{\mathbf{1}_d \le \boldsymbol{i} \le \boldsymbol{N}} H_{\boldsymbol{i}}\right) \gtrsim \sqrt{\log N}. \tag{3.110}$$

Since $N = \prod_{k=1}^d N_k \sim \prod_{k=1}^d J_k$ by Lemma B.1, then $t_{\infty,\boldsymbol{r}}^2 \gtrsim \sigma_0^2 h_{\infty,\boldsymbol{r}}^2 \gtrsim (\log n/n)\prod_{k=1}^d J_k^{2r_k+1}$. Therefore, we have uniformly in $\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R$,

$$\begin{aligned}
\mathrm{E}_0(\|D^{\boldsymbol{r}} f_0 - \boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}\|_\infty^2) &= \mathrm{E}(\|\boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{F}_0 + \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta} - D^{\boldsymbol{r}} f_0 + Q_{n,\boldsymbol{r}}\|_\infty^2) \\
&\le 2\mathrm{E}(\|Q_{n,\boldsymbol{r}}\|_\infty^2) + 2\mathrm{E}(\|\boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{F}_0 + \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta} - D^{\boldsymbol{r}} f_0\|_\infty^2) \\
&\lesssim \frac{\log n}{n}\prod_{k=1}^d J_k^{2r_k+1} + \frac{1}{n^2}\prod_{k=1}^d J_k^{2r_k+2} + \sum_{k=1}^d J_k^{-2(\alpha-r_k)} \\
&\lesssim \frac{\log n}{n}\prod_{k=1}^d J_k^{2r_k+1} + \sum_{k=1}^d J_k^{-2(\alpha-r_k)},
\end{aligned} \tag{3.111}$$

where the last two lines follow from (3.71) and (3.63). By Markov's inequality, we have uniformly over $\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R$ that

$$P_0(\|D^{\boldsymbol{r}} f_0 - \boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}\|_\infty > \rho_n t_{\infty,\boldsymbol{r}}) \lesssim \frac{(\log n/n)\prod_{k=1}^d J_k^{2r_k+1} + \sum_{k=1}^d J_k^{-2(\alpha-r_k)}}{\rho_n^2 t_{\infty,\boldsymbol{r}}^2}. \tag{3.112}$$

Now, since $t_{\infty,\boldsymbol{r}}^2 \gtrsim (\log n/n)\prod_{k=1}^d J_k^{2r_k+1}$, we have $\rho_n^2 t_{\infty,\boldsymbol{r}}^2 \gg (\log n/n)\prod_{k=1}^d J_k^{2r_k+1}$. If $J_k \asymp (n/\log n)^{1/(2\alpha+d)}$ for $k = 1,\ldots,d$, we have $\rho_n^2 t_{\infty,\boldsymbol{r}}^2 \gg \sum_{k=1}^d J_k^{-2(\alpha-r_k)}$. Hence the fraction on the right hand side goes to 0 as $n \to \infty$. Therefore, $\inf_{\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R} P_0(\|D^{\boldsymbol{r}} f_0 - \boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}\|_\infty \le \rho_n t_{\infty,\boldsymbol{r}}) \to 1$ if $J_k \asymp (n/\log n)^{1/(2\alpha+d)}$ for $k = 1,\ldots,d$. Applying (3.5) from Ledoux and Talagrand (1991) to $Z_{n,\boldsymbol{r}}$ with $\|\cdot\|$ as the L_∞-norm, we have

$$\omega = P(\|Z_{n,\boldsymbol{r}}\|_\infty > h_{\infty,\boldsymbol{r}}) \le 4\exp\left(-\frac{h_{\infty,\boldsymbol{r}}^2}{8\mathrm{E}\|Z_{n,\boldsymbol{r}}\|_\infty^2}\right).$$

Then, $h_{\infty,\boldsymbol{r}} \lesssim (\mathrm{E}\|Z_{n,\boldsymbol{r}}\|_\infty^2)^{1/2}\sqrt{-\log\omega}$. By (3.71), we have $\mathrm{E}\|Z_{n,\boldsymbol{r}}\|_\infty^2 \lesssim (\log n/n)\prod_{k=1}^d J_k^{2r_k+1}$. In view of (3.43), we have uniformly over $\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \le R$ that the radius of $\widehat{\mathcal{C}}_\infty$ is $\rho_n\widehat{\sigma}_n h_{\infty,\boldsymbol{r}} =$

$O_{P_0}((\log n/n)^{(\alpha-|\boldsymbol{r}|)/(2\alpha+d)}\rho_n)$. □

Remark For the anisotropic case, assuming that f_0 satisfies Assumption 3, then if $J_k \asymp (n/\log n)^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, the asymptotic coverage of $\widehat{\mathcal{C}}_\infty$ is 1, and its radius is $O_{P_0}((\log n/n)^{\alpha^*\{1-\sum_{k=1}^d (r_k/\alpha_k)\}/(2\alpha^*+d)}\rho_n)$ uniformly over $\|D^{\boldsymbol{r}} f_0\|_{\boldsymbol{\alpha},\infty} \leq R$.

Remark If we allow $\omega = \omega_n \to 0$ without using ρ_n, as in the construction for pointwise and L_2-credible regions, we need $h^2_{\infty,\boldsymbol{r}} \gg (\log n/n)\prod_{k=1}^d J_k^{2r_k+1}$ to make the right hand side of (3.112) approach 0. We currently are unable to show this but conjecture that the lower bound for $h^2_{\infty,\boldsymbol{r}}$ established above, i.e., $(\log n/n)\prod_{k=1}^d J_k^{2r_k+1}$ is sharp regardless whether $\omega < 1/2$ is fixed or approaching 0. The introduction of an inflation factor is not unique to this problem, and it was used by Szabo et al. (2014) in the context of constructing adaptive L_2-credible regions.

Another way to ensure guaranteed frequentist coverage is to undersmooth the prior, i.e., using more B-splines in the expansion than the optimal number at each dimension. However, it turns out that we need two different priors to accomplish this task for L_∞-credible sets. Namely, we use the optimal J_k to estimate the center $\boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} + \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}$ and a larger J_k to estimate the size of the radius for each $k = 1, \ldots, d$. We write f as $f_{\boldsymbol{J}}$ to show that $\boldsymbol{J} = (J_1, \ldots, J_d)^T$ tensor B-splines have been used in the expansion of (3.2). In a similar manner, we write $\boldsymbol{A}^{\boldsymbol{J}}_{\boldsymbol{r}}$ for $\boldsymbol{A}_{\boldsymbol{r}}$, $\boldsymbol{c}^{\boldsymbol{J}}_{\boldsymbol{r}}$ for $\boldsymbol{c}_{\boldsymbol{r}}$, $\Sigma^{\boldsymbol{J}}_{\boldsymbol{r}}$ for $\Sigma_{\boldsymbol{r}}$ and $\widehat{\sigma}^2_{n,\boldsymbol{J}}$ for $\widehat{\sigma}^2_n$. Let $\boldsymbol{J}_n = (J_{n,1}, \ldots, J_{n,d})^T$ where $J_{n,k} \asymp (n/\log n)^{1/(2\alpha+d)}$ is the optimal number of basis at each dimension $k = 1, \ldots, d$. Using the same construction as in (3.102), consider the following empirical credible region for $f_{\boldsymbol{J}}$ with radius $\widehat{\sigma}_{n,\boldsymbol{J}} h_{\boldsymbol{J},\boldsymbol{r}}$,

$$\widehat{\mathcal{C}}^{\boldsymbol{J}}_\infty = \{f_{\boldsymbol{J}} : \|D^{\boldsymbol{r}} f_{\boldsymbol{J}} - \boldsymbol{A}^{\boldsymbol{J}}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}^{\boldsymbol{J}}_{\boldsymbol{r}}\boldsymbol{\eta}\|_\infty \leq \widehat{\sigma}_{n,\boldsymbol{J}} h_{\boldsymbol{J},\boldsymbol{r}}\}, \tag{3.113}$$

where $h_{\boldsymbol{J},\boldsymbol{r}}$ does not depend on $\boldsymbol{Y}$. We first choose $h_{\boldsymbol{J},\boldsymbol{r}}$ such that $\Pi_{\widehat{\sigma}_{n,\boldsymbol{J}}}(\widehat{\mathcal{C}}^{\boldsymbol{J}}_\infty|\boldsymbol{Y}) = 1 - \omega$ with $\omega < 1/2$, where the appropriate order for each component of $\boldsymbol{J}$ will be determined in Theorem 3.20 below. Using this $h_{\boldsymbol{J},\boldsymbol{r}}$, we then study the frequentist coverage of the following set:

$$\inf_{\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \leq R} P_0(\|D^{\boldsymbol{r}} f_0 - \boldsymbol{A}^{\boldsymbol{J}_n}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}^{\boldsymbol{J}_n}_{\boldsymbol{r}}\boldsymbol{\eta}\|_\infty \leq \widehat{\sigma}_{n,\boldsymbol{J}} h_{\boldsymbol{J},\boldsymbol{r}}). \tag{3.114}$$

Theorem 3.20. *Let $|\boldsymbol{r}| < \beta < \alpha$. If $J_k \asymp (n/\log n)^{1/(2\beta+d)}$ for $k = 1, \ldots, d$, then the asymptotic coverage of* (3.114) *is* 1*, and the radius of $\widehat{\mathcal{C}}^{\boldsymbol{J}}_\infty$ is $O_{P_0}((\log n/n)^{(\beta-|\boldsymbol{r}|)/(2\beta+d)})$ uniformly over $\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \leq R$ for $R > 0$ and any $|\boldsymbol{r}| < \beta$.*

Proof. Define $t_{\boldsymbol{J},\boldsymbol{r}} = \inf_{|\sigma^2-\sigma_0^2|\leq\zeta_n/\sqrt{n}} \sigma h_{\boldsymbol{J},\boldsymbol{r}}$. Following the same argument as in Theorem 3.19, we see that $t^2_{\boldsymbol{J},\boldsymbol{r}} \gtrsim (\log n/n)\prod_{k=1}^d J_k^{2r_k+1}$. Using the optimal $J_{n,k} \asymp (n/\log n)^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$ to estimate the posterior mean, $\mathrm{E}_0(\|D^{\boldsymbol{r}} f_0 - \boldsymbol{A}^{\boldsymbol{J}_n}_{\boldsymbol{r}}\boldsymbol{Y} - \boldsymbol{c}^{\boldsymbol{J}_n}_{\boldsymbol{r}}\boldsymbol{\eta}\|^2_\infty) \lesssim (\log n/n)\prod_{k=1}^d J_{n,k}^{2r_k+1} +$

$\sum_{k=1}^d J_{n,k}^{-2(\alpha-r_k)}$ by substituting J_k with $J_{n,k}$ for $k = 1, \ldots, d$ in (3.111) of Theorem 3.19. By Markov's inequality, we have uniformly over $\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \leq R$ that

$$P_0(\|D^{\boldsymbol{r}} f_0 - \boldsymbol{A}_{\boldsymbol{r}}^{\boldsymbol{J}_n} \boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}^{\boldsymbol{J}_n} \boldsymbol{\eta}\|_\infty > t_{\boldsymbol{J},\boldsymbol{r}}) \lesssim \frac{(\log n/n) \prod_{k=1}^d J_{n,k}^{2r_k+1} + \sum_{k=1}^d J_{n,k}^{-2(\alpha-r_k)}}{(\log n/n) \prod_{k=1}^d J_k^{2r_k+1}}. \quad (3.115)$$

Since $J_k \gg J_{n,k}$ for $k = 1, \ldots, d$, we have $(\log n/n) \prod_{k=1}^d J_k^{2r_k+1} \gg (\log n/n) \prod_{k=1}^d J_{n,k}^{2r_k+1}$ and $(\log n/n) \prod_{k=1}^d J_k^{2r_k+1} \gg \sum_{k=1}^d J_{n,k}^{-2(\alpha-r_k)}$, and the right hand side above will go to 0 as $n \to \infty$. We then conclude that $\inf_{\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \leq R} P_0(\|D^{\boldsymbol{r}} f_0 - \boldsymbol{A}_{\boldsymbol{r}}^{\boldsymbol{J}_n} \boldsymbol{Y} - \boldsymbol{c}_{\boldsymbol{r}}^{\boldsymbol{J}_n} \boldsymbol{\eta}\|_\infty \leq t_{\boldsymbol{J},\boldsymbol{r}}) \to 1$ with $J_k \asymp n^{1/(2\beta+d)}$ for $k = 1, \ldots, d$. As in the proof of Theorem 3.19, it follows from (3.5) of Ledoux and Talagrand (1991) that $h_{\boldsymbol{J},\boldsymbol{r}} \lesssim (\mathrm{E}\|Z_{n,\boldsymbol{r}}^{\boldsymbol{J}}\|_\infty^2)^{1/2} \sqrt{-\log \omega}$ for $Z_{n,\boldsymbol{r}}^{\boldsymbol{J}} \sim \mathrm{GP}(0, \Sigma_{\boldsymbol{r}}^{\boldsymbol{J}})$. By (3.71), we have $\mathrm{E}\|Z_{n,\boldsymbol{r}}^{\boldsymbol{J}}\|_\infty^2 \lesssim (\log n/n) \prod_{k=1}^d J_k^{2r_k+1}$. In view of (3.43), we have uniformly over $\|D^{\boldsymbol{r}} f_0\|_{\mathcal{H}^{\alpha-|\boldsymbol{r}|}} \leq R$ that the radius of $\widehat{\mathcal{C}}_\infty^{\boldsymbol{J}}$ is $\widehat{\sigma}_{n,\boldsymbol{J}} h_{\boldsymbol{J},\boldsymbol{r}} = O_{P_0}((\log n/n)^{(\beta-|\boldsymbol{r}|)/(2\beta+d)})$. □

Remark If we had used the optimal $\boldsymbol{J}_n$ to estimate both the radius and the posterior mean, then the right hand side of (3.115) does not approach 0, and hence the need for two different priors. This is in contrast to the undersmoothing method described in Knapik et al. (2011) where they considered L_2-credible region. In Theorem 4.2 of that paper, one prior is sufficient to do undersmoothing. The main difference is that unlike in the L_2 case, the expectation and standard deviation of $\|Z_{n,\boldsymbol{r}}^{\boldsymbol{J}}\|_\infty$ has the same order of magnitude, and their argument using asymptotically completely separated distributions (see page 22 of the same paper) does not apply here.

Remark For the anisotropic case, we assume that f_0 satisfies Assumption 3. Let $\beta_k < \alpha_k$ for all $k = 1, \ldots, d$ and $\sum_{k=1}^d (r_k/\beta_k) < 1$. If $J_k \asymp (n/\log n)^{\beta^*/\{\beta_k(2\beta^*+d)\}}$ for $k = 1, \ldots, d$, we have uniformly in $\|D^{\boldsymbol{r}} f_0\|_{\boldsymbol{\alpha},\infty} \leq R$ that the asymptotic coverage of $\widehat{\mathcal{C}}_\infty^{\boldsymbol{J}}$ is 1 and its radius is $O_{P_0}((\log n/n)^{\beta^*\{1-\sum_{k=1}^d (r_k/\beta_k)\}/(2\beta^*+d)})$.

Theorem 3.19 uses the optimal number of tensor B-spline basis to estimate both the center and radius, but inflates the width of the credible region by some factor to ensure frequentist coverage of 1 asymptotically. Theorem 3.20 achieves the same aim by using the optimal number to estimate the center but uses a larger $\boldsymbol{J}$ component-wise to estimate the radius. In terms of implementation, Theorem 3.19 enjoys two advantages over Theorem 3.20. The first advantage is that we do not have to use two priors, one with optimal $\boldsymbol{J}_n$ and the other with larger $\boldsymbol{J}$ component-wise for point estimation and credible region construction respectively. Since the inflation factor ρ_n does not depend on α, which is the underlying smoothness of f_0, it can be used in the adaptive context, i.e, in the case of unknown α and we put a prior on each component of $\boldsymbol{J}$. On the other hand, we cannot undersmooth if we use an adaptive prior on each

component of $\boldsymbol{J}$, since adaptive prior by definition smooths the right amount.

In view of Theorems 3.17–3.20, we see that there are three different types of credible region constructions that will guarantee frequentist coverage of 1 asymptotically. The first is to let the credible level $1 - \omega_n$ approach one as $n \to \infty$, and it was used to construct pointwise and L_2-credible regions (Theorems 3.17 and 3.18 respectively). If ω_n does not depend on n, then one could either inflate the radius by some slowly varying factor ρ_n (Theorem 3.19), or do undersmoothing by using a larger $\boldsymbol{J}$ component-wise in determining the radius (Theorem 3.20).

Chapter 4

Bayesian procedure for estimating location of maximum

Consider the regression problem in (3.1) with true function $f_0 : (0,1)^d \to \mathbb{R}$. Recall that Assumption 2 says f_0 belongs to the isotropic Hölder space $\mathcal{H}^\alpha((0,1)^d)$ for $\alpha > 2$. To estimate the location of maximum for f in (3.1), we make an additional assumption by requiring that f_0 has a unique maximum. That is, for positive constants τ, τ_1, δ and λ_0 such that $\tau_1 \leq \tau$, we have the following condition:

Assumption 4. There is a unique point $\boldsymbol{\mu}_0$ in $(0,1)^d$ that maximizes f_0. That is, $M_0 = \sup_{\boldsymbol{x}\in(0,1)^d} f_0(\boldsymbol{x}) = f_0(\boldsymbol{\mu}_0)$. Moreover, f_0 has a well-separated maximum, i.e., $f_0(\boldsymbol{\mu}_0) \geq f_0(\boldsymbol{x}) + \delta$ for all $\boldsymbol{x} \notin B(\boldsymbol{\mu}_0, \tau_1)$, and

$$\sup_{\boldsymbol{x}\in B(\boldsymbol{\mu}_0,\tau)} \lambda_{\max}(\boldsymbol{H}f_0(\boldsymbol{x})) \leq -\lambda_0. \tag{4.1}$$

Assumptions 2 and 4 imply that $\nabla f_0(\boldsymbol{\mu}_0) = \mathbf{0}$ and the Hessian $\boldsymbol{H}f_0(\boldsymbol{\mu}_0)$ is symmetric and negative definite. In addition, the Hessian matrix $\boldsymbol{H}f_0(\boldsymbol{x})$ is continuous. Observe that Assumption 4 is still valid under the anisotropic case, and both Assumptions 3 and 4 will yield the same implications. Let $\mathcal{W}^\alpha((0,1)^d)$ be a subspace of $\mathcal{H}^\alpha((0,1)^d)$ such that each of its elements satisfies Assumptions 2 and 4. We endow this subspace with the following norm, such that for $f \in \mathcal{W}^\alpha((0,1)^d)$,

$$\|f\|_{\mathcal{W}^\alpha} = \|f\|_{\mathcal{H}^\alpha} + \sum_{k=1}^{d} \|D_k f\|_{\mathcal{H}^{\alpha-1}},$$

where $\|\cdot\|_{\mathcal{H}^\alpha}$ is the isotropic Hölder norm defined in (1.9). For the anisotropic case, let $f \in \mathcal{W}^{\boldsymbol{\alpha}}((0,1)^d) \subseteq \mathcal{H}^{\boldsymbol{\alpha}}((0,1)^d)$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^T$, and we introduce the following norm:

$$\|f\|_{\mathcal{W}^{\boldsymbol{\alpha}},\infty} = \|f\|_{\boldsymbol{\alpha},\infty} + \sum_{k=1}^{d} \|D_k f\|_{\boldsymbol{\alpha},\infty},$$

where $\|\cdot\|_{\boldsymbol{\alpha},\infty}$ is the anisotropic Hölder norm defined in (3.74). In what follows, we will mainly concentrate on isotropic f_0, with brief remarks on its anisotropic version. If we do not pursuit uniformity over balls in $\mathcal{W}^{\alpha}((0,1)^d)$ for our results, then Assumption 4 can be simplified to:

Assumption 4.1. There is a unique point $\boldsymbol{\mu}_0$ in $(0,1)^d$ that maximizes f_0. Moreover, f_0 has a well-separated maximum and $\lambda_{\max}(\boldsymbol{H}f_0(\boldsymbol{\mu}_0)) < 0$.

In this case, (4.1) becomes an implication of Assumptions 2 and 4.1, with the constants τ and λ_0 depending on f_0. Suppose that f is endowed with the tensor-product B-splines prior as in (3.2), with normal basis coefficients. Define

$$\boldsymbol{\mu} = \underset{x\in(0,1)^d}{\arg\max}\, f(\boldsymbol{x}), \qquad \boldsymbol{\mu}_0 = \underset{x\in(0,1)^d}{\arg\max}\, f_0(\boldsymbol{x})$$

to be the locations of the maximum of f and f_0 respectively, with

$$M = f(\boldsymbol{\mu}), \qquad M_0 = f_0(\boldsymbol{\mu}_0)$$

their corresponding maximum values. To begin with, we show that $\boldsymbol{\mu}$ is almost always unique.

Lemma 4.1. *Assume that f is given the tensor-product B-splines prior as in* (3.2)*. Then $\boldsymbol{\mu}$ is unique for almost all sample paths of f under the posterior distribution given $\boldsymbol{Y}$ (empirical or hierarchical Bayes).*

Proof. For the empirical Bayes case, the reproducing kernel Hilbert space of the Gaussian posterior process $\{f(\boldsymbol{t}) : \boldsymbol{t} \in (0,1)^d\}$ given $\boldsymbol{Y}$ is the J-dimensional space of polynomial splines spanned by elements in $\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\cdot)$ (van der Vaart and van Zanten, 2008a). This implies that it has continuous sample path at every realization. Observe that $\widehat{\sigma}_n^2 > 0$ a.s. from (3.13). Now if

$$0 = \mathrm{Var}_{\widehat{\sigma}_n}(f(\boldsymbol{t}) - f(\boldsymbol{s})|\boldsymbol{Y}) = \widehat{\sigma}_n^2(\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{t}) - \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{s}))^T \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} (\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{t}) - \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{s})),$$

this implies that $B_{j_k,q_k}(t_k) = B_{j_k,q_k}(s_k)$ for $j_k = 1,\ldots,J_k$ and $k = 1,\ldots,d$. Since $q_k \geq \alpha > 2$ (B-splines are not step functions) for $k = 1,\ldots,d$, this further implies that $\boldsymbol{t} = \boldsymbol{s}$. Thus by Lemma 2.6 of Kim and Pollard (1990), $\boldsymbol{\mu}$ is unique for almost all sample paths of $\{f(\boldsymbol{t}) : \boldsymbol{t} \in (0,1)^d\}$ under the empirical posterior distribution given $\boldsymbol{Y}$.

For the hierarchical Bayes case, consider the conditional posterior process $\Pi(f|\boldsymbol{Y},\sigma^2)$ for arbitrary $\sigma^2 > 0$. The reproducing kernel Hilbert space of this Gaussian process is still the space of splines as before, and each realization of $\Pi(f|\boldsymbol{Y},\sigma^2)$ has continuous sample path. Now by substituting σ^2 for $\widehat{\sigma}_n^2$ in the preceding display, we see that $\mathrm{Var}(f(\boldsymbol{t})-f(\boldsymbol{s})|\boldsymbol{Y},\sigma^2) = 0$ implies $\boldsymbol{t} =$

$\boldsymbol{s}$. By Lemma 2.6 of Kim and Pollard (1990), almost every sample path of $\Pi(f|\boldsymbol{Y}, \sigma^2)$ has unique maximum for any $\sigma^2 > 0$. Note that $f \sim \Pi(f|\boldsymbol{Y})$ is generated from the following scheme: draw $\sigma^2 \sim \Pi(\sigma^2|\boldsymbol{Y}) = \text{IG}((\beta+n)/2, (\gamma + n\widehat{\sigma}_n^2)/2)$, and then $f \sim \Pi(f|\boldsymbol{Y}, \sigma^2) = \text{GP}(\boldsymbol{AY} + \boldsymbol{c\eta}, \sigma^2\boldsymbol{\Sigma})$. Hence, almost every draw of f will have unique maximum $\boldsymbol{\mu}$ under $\Pi(f|\boldsymbol{Y})$. □

Lemma 4.2 below describes the exact conditions needed in order for the induced posterior distribution of $\boldsymbol{\mu}$ to be consistent. Since the following consistency result holds in greater generality, we will temporarily leave the nonparametric regression setting.

Lemma 4.2. *Consider a sequence of statistical experiments* $\{\mathfrak{X}^{(n)}, \mathscr{X}^{(n)}, P_\theta^{(n)} : \theta \in \Theta\}$, *where* Θ *is an arbitrary parameter space with observations* $X^{(n)}$ *indexed by* n. *We assume that there is a true* θ_0 *that generates* $X^{(n)}$. *For any* $\theta \in \Theta$, *define transformation* $g_\theta : U \to \mathbb{R}$, *with* U *be some Euclidean space. Further, define* $\mu_0 = \arg\max_{x \in U} g_{\theta_0}(x)$ *and* $\mu = \arg\max_{x \in U} g_\theta(x)$.

Assume that we have posterior consistency for g_θ *under the supremum norm, i.e., for all* $\delta > 0$,

$$\Pi\left(\|g_\theta - g_{\theta_0}\|_\infty > \delta \middle| X^{(n)}\right) \xrightarrow{P_{\theta_0}^{(n)}} 0 \tag{4.2}$$

as $n \to \infty$. *In addition,* g_{θ_0} *has well-separated maximum, i.e., there exists* $\mu_0 \in U$ *such that*

$$\sup_{x \notin B(\mu_0, \epsilon)} g_{\theta_0}(x) < g_{\theta_0}(\mu_0) \quad \textit{for any } \epsilon > 0. \tag{4.3}$$

Then for any $\epsilon > 0$, *we have* $\Pi\left(\|\mu - \mu_0\| > \epsilon \middle| X^{(n)}\right) \xrightarrow{P_{\theta_0}^{(n)}} 0$ *as* $n \to \infty$, *i.e., the posterior distribution of* μ *is consistent at* μ_0.

Proof. For all $\epsilon > 0$ and $\delta > 0$,

$$\begin{aligned}
\Pi\left(\|\mu - \mu_0\| > \epsilon \middle| X^{(n)}\right) &\le \Pi\left(\sup_{x \notin B(\mu_0, \epsilon)} g_\theta(x) > g_\theta(\mu_0) \middle| X^{(n)}\right) \\
&\le \Pi\left(\sup_{x \notin B(\mu_0, \epsilon)} g_\theta(x) > g_\theta(\mu_0), \quad g_\theta(\mu_0) \ge g_{\theta_0}(\mu_0) - \delta/2 \middle| X^{(n)}\right) \\
&\quad + \Pi\left(g_\theta(\mu_0) < g_{\theta_0}(\mu_0) - \delta/2 \middle| X^{(n)}\right).
\end{aligned}$$

The second term above goes to 0 in $P_{\theta_0}^{(n)}$-probability by (4.2). In addition, (4.3) implies that there exists a $\delta > 0$, such that $g_{\theta_0}(x) < g_{\theta_0}(\mu_0) - \delta$ for $x \notin B(\mu_0, \epsilon)$. Hence, for this $\delta > 0$, the first term is bounded above by

$$\Pi\left(g_\theta(x) > g_{\theta_0}(x) + \delta/2 \quad \text{for some} \quad x \notin B(\mu_0, \epsilon) \middle| X^{(n)}\right) \le \Pi\left(\|g_\theta - g_{\theta_0}\|_\infty > \delta/2 \middle| X^{(n)}\right) \to 0$$

in $P_{\theta_0}^{(n)}$-probability as $n \to \infty$. Thus the posterior distribution of μ is consistent at μ_0. □

In the lemma above, θ can be a continuously differentiable function and g_θ is the derivative function. Note that if (4.2) holds uniformly over $\theta_0 \in D$ where D is a ball in Θ, then the posterior of μ is consistent uniformly in $\theta_0 \in D$. Returning to our present context, to show consistency for the posterior of $\boldsymbol{\mu}$, observe that the first condition (4.2) in Lemma 4.2 is satisfied through Theorem 3.5. By Assumption 4, f_0 has a well-separated maximum as defined in the second condition (4.3). Hence, Lemma 4.2 above concludes that the induced posterior of $\boldsymbol{\mu}$ given $\boldsymbol{Y}$ is consistent at $\boldsymbol{\mu}_0$ uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$ for $R > 0$.

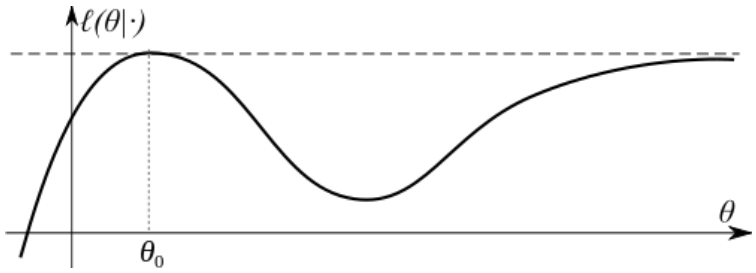


Figure 4.1: A function that has no well-separated maximum. This function $\ell(\theta|\cdot)$ has global maximum at θ_0, but the horizontal asymptote is approaching this maximum as well.

4.1 Posterior convergence rates for $\boldsymbol{\mu}$ and M

In the present case, the most elegant way of deriving posterior convergence rates of $\boldsymbol{\mu}$ and M is by inducing them from the corresponding posterior convergence rates of f and its mixed partial derivatives established in the previous chapter. To accomplish this, we need to find inequalities relating these two sets of quantities. Lemma 4.3 provides this for $\boldsymbol{\mu}$.

Lemma 4.3. *Under Assumptions 2 and 4, we have with posterior probability (empirical or hierarchical Bayes) tending to 1 that*

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| \leq \frac{\sqrt{d}}{\lambda_0} \max_{1 \leq k \leq d} \|D_k f - D_k f_0\|_\infty, \tag{4.4}$$

where λ_0 is the constant defined in (4.1).

Proof. By Taylor's theorem, we have

$$\nabla f_0(\boldsymbol{\mu}) = \nabla f_0(\boldsymbol{\mu}_0) + \boldsymbol{H} f_0(\boldsymbol{\mu}^*)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$$

for $\boldsymbol{\mu}^* = \lambda\boldsymbol{\mu} + (1-\lambda)\boldsymbol{\mu}_0, \lambda \in [0,1]$. Now since the posterior distribution of $\boldsymbol{\mu}$ is consistent at $\boldsymbol{\mu}_0$ by Lemma 4.2, we have that as $n \to \infty$ and $\forall \epsilon > 0$, $\Pi(\|\boldsymbol{\mu}^* - \boldsymbol{\mu}_0\| < \epsilon | \boldsymbol{Y}) \xrightarrow{P_0} 1$, where $\Pi(\cdot|\boldsymbol{Y})$ is a generic posterior that can be either empirical or hierarchical Bayes. Under both Assumptions 2 and 4, equation (4.1) implies that on the event $B_n = \{\lambda_{\max}(\boldsymbol{H} f_0(\boldsymbol{\mu}^*)) \leq -\lambda_0\}$, $\Pi(B_n|\boldsymbol{Y}) \xrightarrow{P_0} 1$. Hence, $\boldsymbol{H} f_0(\boldsymbol{\mu}^*)$ is invertible with probability tending to one under the posterior. Therefore, as $n \to \infty$ and on event B_n,

$$\boldsymbol{\mu} - \boldsymbol{\mu}_0 = \boldsymbol{H} f_0(\boldsymbol{\mu}^*)^{-1}(\nabla f_0(\boldsymbol{\mu}) - \nabla f_0(\boldsymbol{\mu}_0)).$$

Noting that $\nabla f_0(\boldsymbol{\mu}_0) = \nabla f(\boldsymbol{\mu}) = \mathbf{0}$, then

$$\begin{aligned}\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| &\leq \frac{1}{\lambda_0}\|\nabla f_0(\boldsymbol{\mu}) - \nabla f(\boldsymbol{\mu})\| \\ &\leq \frac{\sqrt{d}}{\lambda_0} \max_{1\leq k\leq d} \|D_k f - D_k f_0\|_\infty. \qquad \square\end{aligned}$$

Note that the preceding lemma holds uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$ if the posterior of $\boldsymbol{\mu}$ is consistent at $\boldsymbol{\mu}_0$ uniformly over the same set. For the anisotropic case, (4.4) holds assuming that f_0 satisfies Assumptions 3 and 4.

Theorem 4.4. *If $J_k \asymp (n/\log n)^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, we have for $R > 0$ and any $M_n \to \infty$,*

$$\begin{aligned}&\textit{Empirical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{W}^\alpha}\leq R} \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(\boldsymbol{\mu} : \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| > M_n(\log n/n)^{(\alpha-1)/(2\alpha+d)} | \boldsymbol{Y}) \to 0. \\ &\textit{Hierarchical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{W}^\alpha}\leq R} \mathrm{E}_0 \Pi(\boldsymbol{\mu} : \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| > M_n(\log n/n)^{(\alpha-1)/(2\alpha+d)} | \boldsymbol{Y}) \to 0.\end{aligned}$$

Proof. Apply Theorem 3.8 with $\boldsymbol{r} = \boldsymbol{e}_k$, the sup-norm convergence rate for $D_k f$ is $\epsilon_n = (\log n/n)^{(\alpha-1)/(2\alpha+d)}$ for all $k = 1, \ldots, d$. Using (4.4), for any $M_n \to \infty$ and uniformly in $\|f_0\|_{\mathcal{W}^\alpha} \leq R$,

$$\begin{aligned}\mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| > M_n\epsilon_n | \boldsymbol{Y}) &\leq \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}\left(\max_{1\leq k\leq d} \|D_k f - D_k f_0\|_\infty > \frac{\lambda_0}{\sqrt{d}} M_n\epsilon_n \middle| \boldsymbol{Y}\right) \\ &\leq \sum_{k=1}^d \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}\left(\|D_k f - D_k f_0\|_\infty > \frac{\lambda_0}{\sqrt{d}} M_n\epsilon_n \middle| \boldsymbol{Y}\right) \to 0.\end{aligned}$$

The same argument applies to the hierarchical Bayes version. $\square$

Theorem 4.5. *If $J_k \asymp (n/\log n)^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, we have for $R > 0$ and any $M_n \to \infty$,*

$$\textit{Empirical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{W}^{\boldsymbol{\alpha}}} \leq R} \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(M : |M - M_0| > M_n (\log n/n)^{\alpha/(2\alpha+d)} | \boldsymbol{Y}) \to 0.$$

$$\textit{Hierarchical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{W}^{\boldsymbol{\alpha}}} \leq R} \mathrm{E}_0 \Pi(M : |M - M_0| > M_n (\log n/n)^{\alpha/(2\alpha+d)} | \boldsymbol{Y}) \to 0.$$

Proof. Assume that $f(\boldsymbol{x}), f_0(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in (0,1)^d$. If not, we can always add a large enough constant $C > 0$ such that $f(\boldsymbol{x}) + C > 0$ and $f_0(\boldsymbol{x}) + C > 0$ for all $\boldsymbol{x} \in (0,1)^d$, which leaves $\|f - f_0\|_\infty$ and $|M - M_0|$ unchanged. Note that we can always find such a C not depending on f and f_0 since f_0 is uniformly bounded by Assumption 2, and the posterior support of f is a subspace of $\mathcal{H}^\alpha((0,1)^d)$ consisting of uniformly bounded functions. Then $\|f\|_\infty = \sup_{\boldsymbol{x} \in (0,1)^d} f(\boldsymbol{x}) = M$ and $\|f_0\|_\infty = \sup_{\boldsymbol{x} \in (0,1)^d} f_0(\boldsymbol{x}) = M_0$. Hence by the triangle inequality, we have

$$|M - M_0| = |\|f\|_\infty - \|f_0\|_\infty| \leq \|f - f_0\|_\infty. \tag{4.5}$$

Applying Theorem 3.5 and the inequality above will then induce the desired convergence rate on M. □

Apart from the logarithmic factors, these rates are the minimax rates for estimating $\boldsymbol{\mu}$ and M respectively (see Belitser et al., 2012). For the anisotropic case assuming that f_0 satisfies Assumptions 3 and 4, if $J_k \asymp (n/\log n)^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, then the posterior convergence rates for $\boldsymbol{\mu}$ and M are $(\log n/n)^{(\alpha^*-1)/(2\alpha^*+d)}$ and $(\log n/n)^{\alpha^*/(2\alpha^*+d)}$ uniformly in $\|f_0\|_{\mathcal{W}^\alpha,\infty} \leq R$ respectively.

4.2 Credible regions for $\boldsymbol{\mu}$ and M

We will discuss two methods to construct credible regions for $\boldsymbol{\mu}$ and M. The first method is by constructing Euclidean balls with some radius around the posterior means of $\boldsymbol{\mu}$ and M. Using (4.4), we can then establish that these balls have high posterior concentrations with frequentist coverage tending to 1 asymptotically. The second method is to first construct credible regions for f and its partial derivatives using techniques discussed in Section 3.6, and then we induce the corresponding credible regions for $\boldsymbol{\mu}$ and M by the argmax or max operators respectively.

We begin by discussing the first construction method. Let $\widehat{f}_n$ be some estimator of f_0. Define

$$\widehat{\boldsymbol{\mu}}_n = \operatorname*{arg\,max}_{\boldsymbol{x} \in (0,1)^d} \widehat{f}_n(\boldsymbol{x}). \tag{4.6}$$

Suppose $\widehat{f}_n$ converges to f_0 uniformly in $(0,1)^d$ under P_0-probability, i.e., as $n \to \infty$,

$$\sup_{\boldsymbol{x} \in (0,1)^d} |\widehat{f}_n(\boldsymbol{x}) - f_0(\boldsymbol{x})| \xrightarrow{P_0} 0. \tag{4.7}$$

By Assumption 4, f_0 has a well-separated maximum. Hence Theorem 5.7 of van der Vaart (1998) implies that $\widehat{\boldsymbol{\mu}}_n$ is consistent in estimating $\boldsymbol{\mu}_0$ under P_0-probability. We then have a frequentist analog of Lemma 4.3 as given below:

Lemma 4.6. *Suppose* (4.7) *holds for* $\widehat{f}_n$*. Under Assumptions 2 and 4, we have with probability* P_0 *tending to 1 that*

$$\|\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}_0\| \leq \frac{\sqrt{d}}{\lambda_0} \max_{1 \leq k \leq d} \|D_k \widehat{f}_n - D_k f_0\|_\infty, \tag{4.8}$$

where λ_0 *is the constant defined in* (4.1)*.*

Proof. The argument follows almost identically with the proof in Lemma 4.3, by substituting $\widehat{\boldsymbol{\mu}}_n$ for $\boldsymbol{\mu}$ and $\widehat{f}_n$ for f. In the present case, $P_0(B_n) \to 1$ follows from the implied consistency of $\widehat{\boldsymbol{\mu}}_n$ and (4.1). □

Recall that the empirical posterior mean process for f is $\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}\boldsymbol{\eta}$, where $\boldsymbol{A}$ and $\boldsymbol{c}$ were given by (3.10) and (3.11) respectively. Choose $\mathrm{E}_{\widehat{\sigma}_n}(f|\boldsymbol{Y})$ for $\widehat{f}_n$ such that

$$\widehat{\boldsymbol{\mu}}_n = \arg\max_{\boldsymbol{x} \in (0,1)^d} \mathrm{E}_{\widehat{\sigma}_n}[f(\boldsymbol{x})|\boldsymbol{Y}] = \arg\max_{\boldsymbol{x} \in (0,1)^d} [\boldsymbol{A}(\boldsymbol{x})\boldsymbol{Y} + \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\eta}] \tag{4.9}$$

is an estimator of the location of function maximum $\boldsymbol{\mu}_0$. By Theorem 3.5, it follows that $\boldsymbol{A}\boldsymbol{Y}+\boldsymbol{c}\boldsymbol{\eta}$ converges in P_0-probability to f_0 under L_∞-norm with rate $(\log n/n)^{\alpha/(2\alpha+d)}$. Hence $\widehat{\boldsymbol{\mu}}_n$ defined in (4.9) is consistent in estimating $\boldsymbol{\mu}_0$ under P_0-probability, and (4.8) holds in this case.

Recall that the empirical posterior of $D^{\boldsymbol{r}} f$ given $\boldsymbol{Y}$ is $\mathrm{GP}(\boldsymbol{A}_{\boldsymbol{r}}\boldsymbol{Y} + \boldsymbol{c}_{\boldsymbol{r}}\boldsymbol{\eta}, \widehat{\sigma}_n^2 \Sigma_{\boldsymbol{r}})$, where $\boldsymbol{A}_{\boldsymbol{r}}, \boldsymbol{c}_{\boldsymbol{r}}$ and $\Sigma_{\boldsymbol{r}}$ were given by (3.59), (3.60) and (3.61) respectively. For notational simplicity, if $\boldsymbol{r} = \boldsymbol{e}_k$ where $\boldsymbol{e}_k = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ with 1 in the kth position and the rest zero, we write $D^{\boldsymbol{e}_k}$ as D_k, $\boldsymbol{A}_{\boldsymbol{e}_k}$ as $\boldsymbol{A}_k$, $\boldsymbol{c}_{\boldsymbol{e}_k}$ as $\boldsymbol{c}_k$ and $\Sigma_{\boldsymbol{e}_k}$ as Σ_k respectively for any $k = 1, \ldots, d$. In view of Lemma 4.3 and utilizing the inflation factor approach (see the discussion preceding Theorem 3.19), we can construct a $1 - \omega$ credible ball for $\boldsymbol{\mu}$ centered at $\widehat{\boldsymbol{\mu}}_n$ with radius $\rho_n \widehat{\sigma}_n h_{\max}$. That is,

$$\widehat{\mathcal{C}}_{\boldsymbol{\mu}} = \{\boldsymbol{\mu} : \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n\| \leq \rho_n \widehat{\sigma}_n h_{\max}\}, \tag{4.10}$$

where $\rho_n \to \infty$ as $n \to \infty$ and $h_{\max} = \max_{1 \leq k \leq d} h_{\infty,k}$. For each $1 \leq k \leq d$, $h_{\infty,k}$ is determined

such that

$$\Pi_{\widehat{\sigma}_n}(f : \|D_k f - \boldsymbol{A}_k \boldsymbol{Y} - \boldsymbol{c}_k \boldsymbol{\eta}\|_\infty \leq \widehat{\sigma}_n h_{\infty,k} | \boldsymbol{Y}) = 1 - \omega, \tag{4.11}$$

where $\omega \in [0,1]$ does not depend on n and we assume that $\omega < 1/2$. Observe that $\boldsymbol{A}_k \boldsymbol{Y} + \boldsymbol{c}_k \boldsymbol{\eta} = \mathrm{E}(D_k f | \boldsymbol{Y}) = D_k[\mathrm{E}(f|\boldsymbol{Y})]$. By (4.4), the credibility of $\widehat{\mathcal{C}}_{\boldsymbol{\mu}}$ is at least $1-\omega$,

$$\begin{aligned}\Pi_{\widehat{\sigma}_n}(\widehat{\mathcal{C}}^c_{\boldsymbol{\mu}} | \boldsymbol{Y}) &\leq \Pi_{\widehat{\sigma}_n}\left(\max_{1\leq k\leq d} \|D_k f - D_k[\mathrm{E}(f|\boldsymbol{Y})]\|_\infty > \frac{\lambda_0}{\sqrt{d}} \rho_n \widehat{\sigma}_n h_{\max} \middle| \boldsymbol{Y}\right) \\ &\leq d \max_{1\leq k\leq d} \Pi_{\widehat{\sigma}_n}\left(\|D_k f - \boldsymbol{A}_k \boldsymbol{Y} - \boldsymbol{c}_k \boldsymbol{\eta}\|_\infty > \frac{\lambda_0}{\sqrt{d}} \rho_n \widehat{\sigma}_n h_{\max} \middle| \boldsymbol{Y}\right) \leq \omega. \end{aligned} \tag{4.12}$$

We shall study the frequentist coverage of $\widehat{\mathcal{C}}_{\boldsymbol{\mu}}$ as $n \to \infty$, given by

$$\inf_{\|f_0\|_{\mathcal{W}^\alpha} \leq R} P_0\left(\|\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n\| \leq \rho_n \widehat{\sigma}_n h_{\max}\right). \tag{4.13}$$

Theorem 4.7. *If $J_k \asymp (n/\log n)^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, the asymptotic coverage of (4.13) is 1 and the radius of $\widehat{\mathcal{C}}_{\boldsymbol{\mu}}$ is $O_{P_0}((\log n/n)^{(\alpha-1)/(2\alpha+d)} \rho_n)$ uniformly in $\|f_0\|_{\mathcal{W}^\alpha} \leq R$ for $R > 0$.*

Proof. Applying Theorem 3.19 for $\boldsymbol{r} = \boldsymbol{e}_k$ and in view of (4.8), we have

$$\begin{aligned} P_0\left(\|\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n\| > \rho_n \widehat{\sigma}_n h_{\max}\right) &\leq P_0\left(\max_{1\leq k\leq d} \|D_k f_0 - D_k[\mathrm{E}(f|\boldsymbol{Y})]\|_\infty > \frac{\lambda_0}{\sqrt{d}} \rho_n \widehat{\sigma}_n h_{\max}\right) \\ &\leq d \max_{1\leq k\leq d} P_0\left(\|D_k f_0 - \boldsymbol{A}_k \boldsymbol{Y} - \boldsymbol{c}_k \boldsymbol{\eta}\|_\infty > \frac{\lambda_0}{\sqrt{d}} \rho_n \widehat{\sigma}_n h_{\max}\right) \to 0, \end{aligned}$$

uniformly in $\|f_0\|_{\mathcal{W}^\alpha} \leq R$. □

Remark For the anisotropic case, assuming that f_0 satisfies both Assumptions 3 and 4, if $J_k \asymp (n/\log n)^{\alpha^*/\{\alpha_k(2\alpha+d)\}}$ for $k = 1, \ldots, d$, the asymptotic coverage of $\widehat{\mathcal{C}}_{\boldsymbol{\mu}}$ is 1, and its radius is $O_{P_0}((\log n/n)^{(\alpha^*-1)/(2\alpha^*+d)} \rho_n)$ uniformly in $\|f_0\|_{\mathcal{W}^\alpha,\infty} \leq R$.

Now let $\widehat{M}_n = \boldsymbol{A}(\widehat{\boldsymbol{\mu}}_n)\boldsymbol{Y} + \boldsymbol{c}(\widehat{\boldsymbol{\mu}}_n)\boldsymbol{\eta}$ be an estimator of the function maximum value M, where $\widehat{\boldsymbol{\mu}}_n$ was defined in (4.9). Using the inflation factor approach with $\omega < 1/2$, we construct a $1-\omega$ credible interval for M as

$$\widehat{\mathcal{C}}_M = \{M : |M - \widehat{M}_n| \leq \rho_n \widehat{\sigma}_n h_\infty\}, \tag{4.14}$$

where h_∞ is chosen such that $\Pi_{\widehat{\sigma}_n}(\|f - \boldsymbol{A}\boldsymbol{Y} - \boldsymbol{c}\boldsymbol{\eta}\|_\infty \leq \widehat{\sigma}_n h_\infty | \boldsymbol{Y}) = 1 - \omega$. Applying Theorem 3.19 with $\boldsymbol{r} = \boldsymbol{0}$ and identifying $h_{\infty,\boldsymbol{0}}$ with h_∞, (4.5) implies that $\widehat{\mathcal{C}}_M$ has credibility at least

$1-\omega$,

$$\Pi_{\widehat{\sigma}_n}(\widehat{\mathcal{C}}_M|\boldsymbol{Y}) \geq \Pi_{\widehat{\sigma}_n}(\|f - \boldsymbol{A}\boldsymbol{Y} - \boldsymbol{c}\boldsymbol{\eta}\|_\infty \leq \rho_n \widehat{\sigma}_n h_\infty|\boldsymbol{Y}) = 1-\omega.$$

We shall study the frequentist coverage of $\widehat{\mathcal{C}}_M$ as $n \to \infty$, given by

$$\inf_{\|f_0\|_{\mathcal{W}^\alpha} \leq R} P_0(|M_0 - \widehat{M}_n| \leq \rho_n \widehat{\sigma}_n h_\infty). \tag{4.15}$$

Theorem 4.8. *If $J_k \asymp (n/\log n)^{1/(2\alpha+d)}$ for $k = 1, \ldots, d$, the asymptotic coverage of (4.15) is 1, and the radius of $\widehat{\mathcal{C}}_M$ is $O_{P_0}((\log n/n)^{\alpha/(2\alpha+d)}\rho_n)$ uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$ for $R > 0$.*

Proof. Apply Theorem 3.19 with $\boldsymbol{r} = \boldsymbol{0}$ and use (4.5) to get uniformly in $\|f_0\|_{\mathcal{H}^\alpha} \leq R$,

$$1 \geq P_0(|M_0 - \widehat{M}_n| \leq \rho_n \widehat{\sigma}_n h_\infty) \geq P_0(\|f_0 - \boldsymbol{A}\boldsymbol{Y} - \boldsymbol{c}\boldsymbol{\eta}\|_\infty \leq \rho_n \widehat{\sigma}_n h_\infty) \to 1. \quad \square$$

Remark For the anisotropic case, assuming that f_0 satisfies both Assumptions 3 and 4, if $J_k \asymp (n/\log n)^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, the asymptotic coverage of $\widehat{\mathcal{C}}_M$ is 1, and its radius is $O_{P_0}((\log n/n)^{\alpha^*/(2\alpha^*+d)}\rho_n)$ uniformly in $\|f_0\|_{\mathcal{W}^{\boldsymbol{\alpha}},\infty} \leq R$.

For the second method, we begin by constructing a credible region for f using techniques discussed in Section 3.6. In particular, in view of the inequality in (4.4) and the inflation method used in Theorem 3.19, we construct a $1-\omega$ credible region for f as

$$\widehat{\mathcal{C}}_{\max} = \left\{ f : \max_{1 \leq k \leq d} \|D_k f - \boldsymbol{A}_k \boldsymbol{Y} - \boldsymbol{c}_k \boldsymbol{\eta}\|_\infty \leq \rho_n \widehat{\sigma}_n h \right\}, \tag{4.16}$$

where $h > 0$ is chosen such that $\Pi_{\widehat{\sigma}_n}(\widehat{\mathcal{C}}_{\max}|\boldsymbol{Y}) = 1-\omega$. Here, ω does not depend on n and we assume that $\omega < 1/2$. Also, h does not depend on $\boldsymbol{Y}$. Observe that the posterior mean $\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{c}\boldsymbol{\eta} \in \widehat{\mathcal{C}}_{\max}$. It also follows from Theorem 4.9 below that $f_0 \in \widehat{\mathcal{C}}_{\max}$ as $n \to \infty$, uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$ for any $R > 0$.

Theorem 4.9. *If $J_k \asymp (n/\log n)^{1/(2\alpha+1)}$ for $k = 1, \ldots, d$, then as $n \to \infty$,*

$$\inf_{\|f_0\|_{\mathcal{W}^\alpha} \leq R} P_0(f_0 \in \widehat{\mathcal{C}}_{\max}) \to 1,$$

and the radius of $\widehat{\mathcal{C}}_{\max}$ is $O_{P_0}((\log n/n)^{(\alpha-1)/(2\alpha+d)}\rho_n)$ uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$ for any $R > 0$.

Proof. The proof is similar to the proof of Theorem 3.19. For any $\xi_n = o(\sqrt{n})$, define $t =$

$\inf_{|\sigma^2-\sigma_0^2|\le\xi_n/\sqrt{n}}\sigma h$. By (3.43), to show $\widehat{\mathcal{C}}_{\max}$ has asymptotic coverage of 1, it suffices to show

$$\sup_{\|f_0\|_{\mathcal{W}^\alpha}\le R} P_0\left(\max_{1\le k\le d}\|D_k f_0 - \boldsymbol{A}_k\boldsymbol{Y} - \boldsymbol{c}_k\boldsymbol{\eta}\|_\infty > \rho_n t\right)\to 0. \tag{4.17}$$

Now, let $Z_{n,k}\sim \mathrm{GP}(0,\Sigma_k)$. Then $\Pi_{\widehat{\sigma}_n}(\widehat{\mathcal{C}}_{\max}|\boldsymbol{Y}) = 1-\omega$ implies that $P(\max_{1\le k\le d}\|Z_{n,k}\|_\infty > h) = \omega$ for $\omega < 1/2$. Since $P(\|Z_{n,i}\|_\infty > h)\le P(\max_{1\le k\le d}\|Z_{n,k}\|_\infty > h) = \omega$ for any $1\le i\le d$, this implies that $h\ge M(Z_{n,i})$ where $M(Z_{n,i})$ is the median of $\|Z_{n,i}\|_\infty$. Using the facts established in the proof of Theorem 3.19, we have $h\ge M(Z_{n,i})\asymp \mathrm{E}\|Z_{n,i}\|_\infty$, and it follows that $t^2\gtrsim\sigma_0^2h^2\gtrsim(\log n/n)J_i^3$. Now by (3.111), we have uniformly over $\|f_0\|_{\mathcal{W}^\alpha}\le R$ that

$$\begin{aligned}\mathrm{E}_0\left(\max_{1\le k\le d}\|D_k f_0 - \boldsymbol{A}_k\boldsymbol{Y} - \boldsymbol{c}_k\boldsymbol{\eta}\|_\infty\right) &\le d\max_{1\le k\le d}\mathrm{E}_0(\|D_k f_0 - \boldsymbol{A}_k\boldsymbol{Y} - \boldsymbol{c}_k\boldsymbol{\eta}\|_\infty)\\ &\lesssim \max_{1\le k\le d}\left(\frac{\log n}{n}J_k^3 + J_k^{-2(\alpha-1)}\right).\end{aligned} \tag{4.18}$$

By Markov's inequality, we have uniformly over $\|f_0\|_{\mathcal{W}^\alpha}\le R$ that

$$P_0\left(\max_{1\le k\le d}\|D_k f_0 - \boldsymbol{A}_k\boldsymbol{Y} - \boldsymbol{c}_k\boldsymbol{\eta}\|_\infty > \rho_n t\right)\lesssim \frac{\max_{1\le k\le d}\left((\log n/n)J_k^3 + J_k^{-2(\alpha-1)}\right)}{\rho_n^2 t^2}. \tag{4.19}$$

Now, since $J_k\asymp(n/\log n)^{1/(2\alpha+1)}$ for $k=1,\dots,d$, we have $\rho_n^2t^2\gg(\log n/n)J_k^3$ and $\rho_n^2t^2\gg J_k^{-2(\alpha-1)}$ for any $1\le k\le d$. Hence the fraction on the right hand side goes to 0 as $n\to\infty$. We conclude that $\inf_{\|f_0\|_{\mathcal{W}^\alpha}\le R}P_0(\max_{1\le k\le d}\|D_k f_0 - \boldsymbol{A}_k\boldsymbol{Y} - \boldsymbol{c}_k\boldsymbol{\eta}\|_\infty\le\rho_n t)\to 1$ if $J_k\asymp(n/\log n)^{1/(2\alpha+d)}$ for $k=1,\dots,d$. Applying (3.5) from Ledoux and Talagrand (1991) to $Z_{n,k}$ with $\|\cdot\|$ as the L_∞-norm, we have

$$\begin{aligned}\omega = P\left(\max_{1\le k\le d}\|Z_{n,k}\|_\infty > h\right) &\le d\max_{1\le k\le d}P(\|Z_{n,k}\|_\infty > h)\le 4d\max_{1\le k\le d}\exp\left(-\frac{h^2}{8\mathrm{E}\|Z_{n,k}\|_\infty^2}\right)\\ &= 4d\exp\left(-\frac{h^2}{8\max_{1\le k\le d}\mathrm{E}\|Z_{n,k}\|_\infty^2}\right).\end{aligned}$$

Hence, $h\lesssim\sqrt{\max_{1\le k\le d}\mathrm{E}\|Z_{n,k}\|_\infty^2\log(4d/\omega)}$. By (3.71), we have $\mathrm{E}\|Z_{n,k}\|_\infty^2\lesssim(\log n/n)J_k^3$. In view of (3.43), we have uniformly over $\|f_0\|_{\mathcal{W}^\alpha}\le R$ that the radius of $\widehat{\mathcal{C}}_{\max}$ is $\rho_n\widehat{\sigma}_n h = O_{P_0}((\log n/n)^{(\alpha-1)/(2\alpha+d)}\rho_n)$. □

Remark For the anisotropic version, we assume that f_0 satisfies Assumptions 3 and 4. If $J_k\asymp(n/\log n)^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k=1,\dots,d$, the asymptotic coverage of $\widehat{\mathcal{C}}_{\max}$ is 1 and its radius is $O_{P_0}((\log n/n)^{(\alpha^*-1)/(2\alpha^*+d)}\rho_n)$ uniformly over $\|f_0\|_{\mathcal{W}^{\boldsymbol{\alpha}},\infty}\le R$ for any $R>0$.

By applying the argmax operator to each element of $\widehat{\mathcal{C}}_{\max}$, we will induce a corresponding

credible region $\widehat{\mathcal{U}}_{\max}$ for $\boldsymbol{\mu}$. By Lemma 4.1, it follows that $\widehat{\mathcal{U}}_{\max}$ has credibility $1-\omega$ exactly,

$$\Pi_{\widehat{\sigma}_n}(\boldsymbol{\mu} \in \widehat{\mathcal{U}}_{\max}|\boldsymbol{Y}) = \Pi_{\widehat{\sigma}_n}(f \in \widehat{\mathcal{C}}_{\max}|\boldsymbol{Y}) = 1-\omega. \tag{4.20}$$

By Assumption 4, $\boldsymbol{\mu}_0$ is the unique point of maximum for f_0. It then follows from Theorem 4.9 that as $n \to \infty$,

$$P_0(\boldsymbol{\mu}_0 \in \widehat{\mathcal{U}}_{\max}) = P_0(f_0 \in \widehat{\mathcal{C}}_{\max}) \to 1, \tag{4.21}$$

uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$ for any $R > 0$. Using this particular $\widehat{\mathcal{C}}_{\max}$ to induce $\widehat{\mathcal{U}}_{\max}$ ensures that the diameter of $\widehat{\mathcal{U}}_{\max}$ is bounded above by some constant multiple of the minimax convergence rate, possibly up to a slowly varying factor.

Lemma 4.10. *The diameter of $\widehat{\mathcal{U}}_{\max}$ is $O_{P_0}((\log n/n)^{(\alpha-1)/(2\alpha+d)}\rho_n)$, uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$ for any $R > 0$.*

Proof. Let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ be two arbitrary elements of $\widehat{\mathcal{U}}_{\max}$. Then by Lemma 4.1, there are functions $f_1, f_2 \in \widehat{\mathcal{C}}_{\max}$ such that $\boldsymbol{\mu}_1 = \arg\max_{\boldsymbol{x}\in(0,1)^d} f_1(\boldsymbol{x})$ and $\boldsymbol{\mu}_2 = \arg\max_{\boldsymbol{x}\in(0,1)^d} f_2(\boldsymbol{x})$. Let $\widetilde{f}_n = \boldsymbol{AY} + \boldsymbol{c\eta}$. By (4.4) and the triangle inequality,

$$\begin{aligned}
&\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| \\
&\leq \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\| + \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_0\| \\
&\leq \frac{\sqrt{d}}{\lambda_0} \max_{1\leq k\leq d} \|D_k f_1 - D_k f_0\|_\infty + \frac{\sqrt{d}}{\lambda_0} \max_{1\leq k\leq d} \|D_k f_2 - D_k f_0\|_\infty \\
&\leq \frac{\sqrt{d}}{\lambda_0} \left(\max_{1\leq k\leq d} \|D_k f_1 - D_k \widetilde{f}_n\|_\infty + \max_{1\leq k\leq d} \|D_k f_2 - D_k \widetilde{f}_n\|_\infty + 2 \max_{1\leq k\leq d} \|D_k f_0 - D_k \widetilde{f}_n\|_\infty \right) \\
&= O_{P_0}(\rho_n \widehat{\sigma}_n h) = O_{P_0}((\log n/n)^{(\alpha-1)/(2\alpha+d)}\rho_n).
\end{aligned}$$

Both $\max_{1\leq k\leq d} \|D_k f_1 - D_k \widetilde{f}_n\|_\infty$ and $\max_{1\leq k\leq d} \|D_k f_2 - D_k \widetilde{f}_n\|_\infty$ are bounded by $\rho_n \widehat{\sigma}_n h$ follow from the definition of $\widehat{\mathcal{C}}_{\max}$. By Theorem 4.9, $\max_{1\leq k\leq d} \|D_k f_0 - D_k \widetilde{f}_n\|_\infty = O_{P_0}(\rho_n \widehat{\sigma}_n h)$. The last line then follows from (3.43) with the upper bound for h established in Theorem 4.9. □

We use the same technique to construct credible interval for M. That is, we construct a credible region for f and apply the max operator to each of its element. In view of (4.5) and Theorem 3.19 with $\omega < 1/2$, we construct a $1-\omega$ credible region for f as

$$\widehat{\mathcal{C}}_\infty = \{f : \|f - \boldsymbol{AY} - \boldsymbol{c\eta}\|_\infty \leq \rho_n \widehat{\sigma}_n h_\infty\}, \tag{4.22}$$

where $h_\infty > 0$ is chosen such that $\Pi_{\widehat{\sigma}_n}(\widehat{\mathcal{C}}_\infty|\boldsymbol{Y}) = 1-\omega$, and h_∞ does not depend on $\boldsymbol{Y}$. Note that the posterior mean $\boldsymbol{AY} + \boldsymbol{c\eta} \in \widehat{\mathcal{C}}_\infty$. It also follows from Theorem 3.19 that $P_0(f_0 \in \widehat{\mathcal{C}}_\infty) \to 1$

as $n \to \infty$ uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$ for any $R > 0$. By applying the max operator to each element of $\widehat{\mathcal{C}}_\infty$, we will induce a corresponding credible region $\widehat{\mathcal{M}}_\infty$ for M. By Lemma 4.1, $\widehat{\mathcal{M}}_\infty$ has credibility $1 - \omega$ exactly,

$$\Pi_{\widehat{\sigma}_n}(M \in \widehat{\mathcal{M}}_\infty | \boldsymbol{Y}) = \Pi_{\widehat{\sigma}_n}(f \in \widehat{\mathcal{C}}_\infty | \boldsymbol{Y}) = 1 - \omega. \tag{4.23}$$

By Assumption 4, $\boldsymbol{\mu}_0$ is the unique point of maximum for f_0 with maximum value M_0. It then follows from Theorem 3.19 that as $n \to \infty$,

$$P_0(M_0 \in \widehat{\mathcal{M}}_\infty) = P_0(f_0 \in \widehat{\mathcal{C}}_\infty) \to 1, \tag{4.24}$$

uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$ for any $R > 0$. Now, using this particular $\widehat{\mathcal{C}}_\infty$ to induce $\widehat{\mathcal{M}}_\infty$ ensures that the diameter of $\widehat{\mathcal{M}}_\infty$ is bounded above by some constant multiple of the minimax convergence rate, possibly up to some slowly varying factor.

Lemma 4.11. *The diameter of $\widehat{\mathcal{M}}_\infty$ is $O_{P_0}((\log n/n)^{\alpha/(2\alpha+d)}\rho_n)$, uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$ for any $R > 0$.*

Proof. Let M_1 and M_2 be two arbitrary elements of $\widehat{\mathcal{M}}_\infty$. Then by Lemma 4.1, there are functions $f_1, f_2 \in \widehat{\mathcal{C}}_\infty$ such that $M_1 = \max_{\boldsymbol{x} \in (0,1)^d} f_1(\boldsymbol{x})$ and $M_2 = \max_{\boldsymbol{x} \in (0,1)^d} f_2(\boldsymbol{x})$. By (4.5) and the triangle inequality, we have

$$|M_1 - M_2| \leq \|f_1 - f_2\|_\infty \leq \|f_1 - \boldsymbol{AY} - \boldsymbol{c\eta}\|_\infty + \|f_2 - \boldsymbol{AY} - \boldsymbol{c\eta}\|_\infty \leq 2\rho_n \widehat{\sigma}_n h_\infty. \tag{4.25}$$

By the definition of $\widehat{\mathcal{C}}_\infty$, both $\|f_1 - \boldsymbol{AY} - \boldsymbol{c\eta}\|_\infty$ and $\|f_2 - \boldsymbol{AY} - \boldsymbol{c\eta}\|_\infty$ are bounded by $\rho_n \widehat{\sigma}_n h_\infty$. The result then follows from (3.43) with the upper bound for h_∞ established in Theorem 3.19. □

Remark The anisotropic versions of the results established above follow by changing α to α^* and the norm $\|\cdot\|_{\mathcal{W}^\alpha}$ to $\|\cdot\|_{\mathcal{W}^{\boldsymbol{\alpha}},\infty}$.

Chapter 5

Two-stage Bayesian estimation of $\boldsymbol{\mu}$ and M

5.1 Introduction

The idea is to derive the posterior distributions of $\boldsymbol{\mu}$ and M in two stages. Given a sampling budget of n, we first allocate $n_1 < n$ samples to compute the posterior distribution of f, based on tensor-product B-splines prior with normal basis coefficients. Using this posterior, we then construct a credible region for $\boldsymbol{\mu}$ based on the methods discussed in the previous section. The rest $n_2 = n - n_1$ samples are then obtained from sampling points inside this region. Assuming that this region is small enough, we can approximate f around this vicinity with a multivariate polynomial. By further endowing the coefficients with a normal prior, we then use the second stage samples to construct an improved posterior distribution for f. By (4.4) and (4.5), this translates to improved posterior distributions for $\boldsymbol{\mu}$ and M, where improvement here refers to faster posterior convergence rates.

5.2 Two-stage Bayesian procedure

Let $\pi \in (0,1)$. In the first stage, we choose $n_1 \in \mathbb{N}$ design points $\{\widetilde{\boldsymbol{x}}_i, i = 1, \ldots, n_1\}$ such that $0 < n_1 < n$ and $n_1/n \to \pi$. If we have no information on these points, then they should be chosen as uniform as possible across the domain of interest. Since our domain is $(0,1)^d$ and assuming that $n_1 = m^d$ for some $m \in \mathbb{N}$, we can divide $(0,1)^d$ into n_1 hypercubes of length m, and sample from the uniform grid points formed. That is, $\widetilde{\boldsymbol{x}}_i \in \{(j-1)/(m-1), j = 1, \ldots, m\}^d$ for $i = 1, \ldots, n_1$. In general, for constants $c_1, c_2 > 0$, we can select these points such that the family of balls $\{B(\widetilde{\boldsymbol{x}}_i, c_1 n_1^{-1/d}), i = 1, \ldots, n_1\}$ covers $(0,1)^d$ and $\|\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j\| \geq c_2 n_1^{-1/d}$ for $i \neq j$ (see Belitser et al., 2012). Therefore, we have data $\mathcal{D}_1 = \{(\widetilde{\boldsymbol{x}}_i, Y_i), i = 1, \ldots, n_1\}$ for the model

in (3.1).

In view of $\mathcal{D}_1$, we endow f in (3.1) with tensor-product B-splines prior described in Section 3.2 to yield the empirical posterior $\Pi_{\widehat{\sigma}_n}(f|\mathcal{D}_1)$. Then, we construct a credible set for $\boldsymbol{\mu}$ as $C(\widetilde{\boldsymbol{\mu}}, v\delta_n) = \{\boldsymbol{\mu} : \|\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}\|_\infty \leq v\delta_n\}$, where $\widetilde{\boldsymbol{\mu}}$ is some preliminary estimator of $\boldsymbol{\mu}$. For example, $\widehat{\boldsymbol{\mu}}_n$ in (4.9). Here, v is the smallest integer such that $2v \geq m_\alpha$ for $m_\alpha = \lceil \alpha - 1 \rceil$, and $\delta_n > 0$ is a localization parameter such that $\delta_n \to 0$ as $n \to \infty$. Note that $C(\widetilde{\boldsymbol{\mu}}, v\delta_n)$ is a cube in $\mathbb{R}^d$ centered at $\widetilde{\boldsymbol{\mu}}$ with diameter $2v\delta_n$. Since $\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\| \leq \sqrt{d}\|\boldsymbol{x}\|_\infty$ for any $\boldsymbol{x} \in \mathbb{R}^d$, it follows from Theorem 4.7 that if $\rho_n \widehat{\sigma}_n h_{\max} = o_{p_0}(\delta_n)$, then $\Pi_{\widehat{\sigma}_n}(C(\widehat{\boldsymbol{\mu}}_n, v\delta_n)|\mathcal{D}_1) \geq 1 - \omega$ and $P_0(\boldsymbol{\mu}_0 \in C(\widehat{\boldsymbol{\mu}}_n, v\delta_n)) \to 1$ uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$ for any $R > 0$.

Therefore, we have $n_2 = n - n_1$ samples left for the second stage. Choose $n_3 \in \mathbb{N}$ such that $n_2 = n_3(2v+1)^d$. Divide $C(\widetilde{\boldsymbol{\mu}}, v\delta_n)$ into $(2v)^d$ smaller cubes of diameter δ_n. The stage two samples are then obtained by sampling n_3 times at each of the $(2v+1)^d$ vertices formed. To be more precise, define the set of $(2v+1)^d$ cube vertices as

$$\prod_{k=1}^{d} \{\widetilde{\mu}_k + t_k\delta_n, t_k = 0, \pm 1, \ldots, \pm v\} = \{\widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{(2v+1)^d}\}.$$

Denote the second stage design points as $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_2}\}$, and they are chosen such that for sets $I_j = \{1 \leq i \leq n_2 : \boldsymbol{x}_i = \widetilde{\boldsymbol{u}}_j\}$ with $j = 1, \ldots, (2v+1)^d$, we have $|I_j| = n_3$ where $|A|$ denotes the cardinality of set A. Hence, each vertices in $\{\widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_{(2v+1)^d}\}$ is replicated n_3 times to form $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_2}\}$. Note that $n_3 \geq cn$ for some constant $c > 0$. Using the model in (3.1), i.e., $Y_i = f(\boldsymbol{x}_i) + \varepsilon_i$ for $i = 1, \ldots, n_2$, we observe second stage samples $\mathcal{D}_2 = \{(\boldsymbol{x}_i, Y_i), i = 1, \ldots, n_2\}$.

At the second stage, we represent $f(\boldsymbol{x})$ at $\boldsymbol{x} = (x_1, \ldots, x_d)^T$ as a multivariate polynomial function of fixed order $m_\alpha = \lceil \alpha - 1 \rceil$, i.e.,

$$f_{\boldsymbol{\vartheta}}(\boldsymbol{x}) = \sum_{\boldsymbol{i} \in \mathbb{I}(m_\alpha)} \vartheta_{\boldsymbol{i}} \boldsymbol{x}^{\boldsymbol{i}} = \sum_{k=0}^{m_\alpha} \sum_{\boldsymbol{i} \in \mathbb{I}_k} \vartheta_{\boldsymbol{i}} \boldsymbol{x}^{\boldsymbol{i}} = \boldsymbol{p}(\boldsymbol{x})^T \boldsymbol{\vartheta}, \tag{5.1}$$

where $\boldsymbol{p}(\boldsymbol{x}) = (\boldsymbol{x}^{\boldsymbol{i}} : \boldsymbol{i} \in \mathbb{I}(m_\alpha))^T$ and $\boldsymbol{\vartheta} = (\vartheta_{\boldsymbol{i}} : \boldsymbol{i} \in \mathbb{I}(m_\alpha))^T$ are the corresponding basis coefficients. The elements of $\mathbb{I}(m_\alpha)$ can be enumerated as $\{\boldsymbol{i}_0, \boldsymbol{i}_1, \ldots, \boldsymbol{i}_W\}$ where $W + 1 = |\mathbb{I}(m_\alpha)| = \sum_{k=0}^{m_\alpha} \binom{d+k-1}{d-1}$ and note that $\boldsymbol{i}_0 = \boldsymbol{0}$. By observing second stage samples $\mathcal{D}_2$, define $\boldsymbol{Y} = (Y_1, \ldots, Y_{n_2})^T$, $\boldsymbol{X} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_{n_2}^T)^T$ and $\boldsymbol{P} = (\boldsymbol{p}(\boldsymbol{x}_1), \ldots, \boldsymbol{p}(\boldsymbol{x}_{n_2}))^T$. Moreover, we write $\boldsymbol{F}_0 = (f_0(\boldsymbol{x}_1), \ldots, f_0(\boldsymbol{x}_{n_2}))^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{n_2})^T$. Note that for $d = 1$, $\boldsymbol{P}$ is a Vandermonde

matrix. Hence, the model in (3.1) can be written compactly as

$$\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\vartheta}, \sigma^2 \sim \mathrm{N}_{n_2}(\boldsymbol{P}\boldsymbol{\vartheta}, \sigma^2 \boldsymbol{I}_{n_2}). \tag{5.2}$$

For the basis coefficients, we assign

$$\boldsymbol{\vartheta}|\sigma^2 \sim \mathrm{N}_{W+1}(\boldsymbol{\xi}, \sigma^2 \boldsymbol{V}). \tag{5.3}$$

We assume that the entries of $\boldsymbol{\xi}$ do not depend on n and $\boldsymbol{V} = \mathrm{diag}(C\delta_n^{-|\boldsymbol{i}_j|} : j = 0, \ldots, W)$ for some constant $C > 0$ not depending on n. Based on (5.2) and (5.3), it then follows that the induced conditional posterior for $f_{\boldsymbol{\vartheta}}$ is $\Pi(f_{\boldsymbol{\vartheta}}|\boldsymbol{Y}, \sigma^2) \sim \mathrm{GP}(\boldsymbol{T}\boldsymbol{Y} + \boldsymbol{d}\boldsymbol{\xi}, \sigma^2 \boldsymbol{H})$, where $\boldsymbol{T}$ and $\boldsymbol{d}$ are bounded linear operators mapping $\mathbb{R}^{n_2}$ and $\mathbb{R}^{W+1}$ respectively to $\mathcal{P}_{m_\alpha}$, and $\boldsymbol{H}$ is the covariance function defined on $C(\widetilde{\boldsymbol{\mu}}, v\delta_n) \times C(\widetilde{\boldsymbol{\mu}}, v\delta_n)$ such that for any $\boldsymbol{x}, \boldsymbol{y} \in C(\widetilde{\boldsymbol{\mu}}, v\delta_n)$,

$$\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{p}(\boldsymbol{x})^T \left(\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{V}^{-1}\right)^{-1} \boldsymbol{P}^T, \tag{5.4}$$

$$\boldsymbol{d}(\boldsymbol{x}) = \boldsymbol{p}(\boldsymbol{x})^T \left(\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{V}^{-1}\right)^{-1} \boldsymbol{V}^{-1}, \tag{5.5}$$

$$\boldsymbol{H}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{p}(\boldsymbol{x})^T \left(\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{V}^{-1}\right)^{-1} \boldsymbol{p}(\boldsymbol{y}). \tag{5.6}$$

Note that since the posterior mean can be written as $\sum_{i=1}^{n_2} T_i(\boldsymbol{x})Y_i + d(\boldsymbol{x})\boldsymbol{\eta}$ at a fixed $\boldsymbol{x}$, with $T_i(\boldsymbol{x})$ being the ith element of the row vector $\boldsymbol{T}(\boldsymbol{x})$, Assumption 2 with Properties 1 and 2 of Appendix C imply that $\boldsymbol{T}\boldsymbol{Y} + \boldsymbol{d}\boldsymbol{\xi}$ is a sub-Gaussian process under P_0.

As in Section 3.2, since σ^2 is unknown, we will consider the empirical and hierarchical Bayes approaches. In the present context, the empirical Bayes posterior is

$$\Pi_{\widehat{\sigma}_n}(f_{\boldsymbol{\vartheta}}|\boldsymbol{Y}) \sim \mathrm{GP}(\boldsymbol{T}\boldsymbol{Y} + \boldsymbol{d}\boldsymbol{\xi}, \widehat{\sigma}_n^2 \boldsymbol{H}), \tag{5.7}$$

where

$$\widehat{\sigma}_n^2 = \frac{(\boldsymbol{Y} - \boldsymbol{P}\boldsymbol{\xi})^T(\boldsymbol{P}\boldsymbol{V}\boldsymbol{P}^T + \boldsymbol{I}_{n_2})^{-1}(\boldsymbol{Y} - \boldsymbol{P}\boldsymbol{\xi})}{n_2}. \tag{5.8}$$

For the hierarchical Bayes approach, we further endow σ^2 with the inverse-gamma prior in (3.15). The corresponding posterior is given by

$$\sigma^2|\boldsymbol{Y} \sim \mathrm{IG}\left(\frac{\beta + n_2}{2}, \frac{\gamma + n_2\widehat{\sigma}_n^2}{2}\right), \tag{5.9}$$

where $\widehat{\sigma}_n^2$ is from (5.8) above with $\beta > 4$ and $\gamma > 0$. By Lemma 1 of Belitser et al. (2012), $\boldsymbol{P}$ has full column rank and it follows that $\boldsymbol{P}^T\boldsymbol{P}$ is invertible.

Let $\boldsymbol{r} = (r_1, \ldots, r_d)^T$ be such that $|\boldsymbol{r}| \leq m_\alpha$. Then the $\boldsymbol{r}$ mixed partial derivative of f with respect to $\boldsymbol{x}$ is

$$D^{\boldsymbol{r}} f_{\boldsymbol{\vartheta}}(\boldsymbol{x}) = \sum_{\boldsymbol{i} \in \mathbb{I}(m_\alpha)} \vartheta_{\boldsymbol{i}} \prod_{k=1}^{d} \frac{\partial^{r_k}}{\partial x_k^{r_k}} x_k^{i_k} = \sum_{\substack{\boldsymbol{i} \in \mathbb{I}(m_\alpha) \\ i_k \geq r_k, k=1,\ldots,d}} \vartheta_{\boldsymbol{i}} \frac{\boldsymbol{i}!}{(\boldsymbol{i}-\boldsymbol{r})!} \boldsymbol{x}^{\boldsymbol{i}-\boldsymbol{r}}, \tag{5.10}$$

which is a multivariate polynomial of order $m_\alpha - |\boldsymbol{r}|$. Now since $\boldsymbol{x}^{\boldsymbol{i}-\boldsymbol{r}} = 0$ if $\boldsymbol{i} < \boldsymbol{r}$, we can perform a change of variable and write $D^{\boldsymbol{r}} f_{\boldsymbol{\vartheta}}(\boldsymbol{x}) = \boldsymbol{p}_{\boldsymbol{r}}(\boldsymbol{x})^T \boldsymbol{N}_{\boldsymbol{r}} \boldsymbol{\vartheta}$, where $\boldsymbol{p}_{\boldsymbol{r}}(\boldsymbol{x}) = (\boldsymbol{x}^{\boldsymbol{i}} : \boldsymbol{i} \in \mathbb{I}(m_\alpha - |\boldsymbol{r}|))^T$. The $|\mathbb{I}(m_\alpha - |\boldsymbol{r}|)| \times |\mathbb{I}(m_\alpha)|$ matrix $\boldsymbol{N}_{\boldsymbol{r}}$ is of the following block matrix form: $\boldsymbol{N}_{\boldsymbol{r}} = [\boldsymbol{0} \quad \boldsymbol{R}]$, where $\boldsymbol{0}$ is a $|\mathbb{I}(m_\alpha - |\boldsymbol{r}|)| \times (|\mathbb{I}(m_\alpha)| - |\mathbb{I}(m_\alpha - |\boldsymbol{r}|)|)$ matrix of zeroes, and $\boldsymbol{R}$ is a $|\mathbb{I}(m_\alpha - |\boldsymbol{r}|)| \times |\mathbb{I}(m_\alpha - |\boldsymbol{r}|)|$ diagonal matrix with diagonal elements $\{(\boldsymbol{r}+\boldsymbol{i})!/\boldsymbol{i}! : \boldsymbol{i} \in \mathbb{I}(m_\alpha - |\boldsymbol{r}|)\}$. Note that in this framework, $\boldsymbol{N}_{\boldsymbol{r}}$ and $\boldsymbol{p}_{\boldsymbol{r}}(\boldsymbol{x})$ reduce to $\boldsymbol{I}_{|\mathbb{I}(m_\alpha)|}$ and $\boldsymbol{p}(\boldsymbol{x})$ respectively when $\boldsymbol{r} = \boldsymbol{0}$.

It then follows that $\Pi_{\widehat{\sigma}_n}(D^{\boldsymbol{r}} f_{\boldsymbol{\vartheta}} | \boldsymbol{Y}) \sim \mathrm{GP}(\boldsymbol{T}_{\boldsymbol{r}} \boldsymbol{Y} + \boldsymbol{d}_{\boldsymbol{r}} \boldsymbol{\xi}, \widehat{\sigma}_n^2 \boldsymbol{H}_{\boldsymbol{r}})$, such that for $\boldsymbol{x}, \boldsymbol{y} \in C(\widetilde{\boldsymbol{\mu}}, v\delta_n)$,

$$\boldsymbol{T}_{\boldsymbol{r}}(\boldsymbol{x}) = \boldsymbol{p}_{\boldsymbol{r}}(\boldsymbol{x})^T \boldsymbol{N}_{\boldsymbol{r}} \left(\boldsymbol{P}^T \boldsymbol{P} + \boldsymbol{V}^{-1} \right)^{-1} \boldsymbol{P}^T, \tag{5.11}$$

$$\boldsymbol{d}_{\boldsymbol{r}}(\boldsymbol{x}) = \boldsymbol{p}_{\boldsymbol{r}}(\boldsymbol{x})^T \boldsymbol{N}_{\boldsymbol{r}} \left(\boldsymbol{P}^T \boldsymbol{P} + \boldsymbol{V}^{-1} \right)^{-1} \boldsymbol{V}^{-1}, \tag{5.12}$$

$$\boldsymbol{H}_{\boldsymbol{r}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{p}_{\boldsymbol{r}}(\boldsymbol{x})^T \boldsymbol{N}_{\boldsymbol{r}} \left(\boldsymbol{P}^T \boldsymbol{P} + \boldsymbol{V}^{-1} \right)^{-1} \boldsymbol{N}_{\boldsymbol{r}}^T \boldsymbol{p}_{\boldsymbol{r}}(\boldsymbol{y}). \tag{5.13}$$

Note that if $\boldsymbol{r} = \boldsymbol{0}$, the three operators (5.11), (5.12) and (5.13) above reduce to that defined in (5.4), (5.5) and (5.6) respectively. Using the same reasoning as before, we see that under Assumption 2 with Properties 1 and 2 of Appendix C, $\boldsymbol{T}_{\boldsymbol{r}} \boldsymbol{Y} + \boldsymbol{d}_{\boldsymbol{r}} \boldsymbol{\xi}$ is a sub-Gaussian process under P_0. Define the location of maximum for $f_{\boldsymbol{\vartheta}}$ and its maximum value as

$$\boldsymbol{\mu} = \operatorname*{arg\,max}_{\boldsymbol{x} \in C(\widetilde{\boldsymbol{\mu}}, v\delta_n)} f_{\boldsymbol{\vartheta}}(\boldsymbol{x}), \qquad M = f_{\boldsymbol{\vartheta}}(\boldsymbol{\mu}). \tag{5.14}$$

Following the same reasoning as in Lemma 4.1, we see that $\boldsymbol{\mu}$ is unique for almost all sample paths of $f_{\boldsymbol{\vartheta}}$ under the empirical or hierarchical posterior. The following two theorems establish the second stage posterior convergence rates of $\boldsymbol{\mu}$ and M for any smoothness level $\alpha > 2$.

Theorem 5.1. *Under Assumptions 2 and 4, if $\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\| = o_{P_0}(\delta_n)$ uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$ with $R > 0$, then for any $M_n \to \infty$,*

$$\textit{Empirical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{W}^\alpha} \leq R} \mathrm{E}_0 \Pi_{\widehat{\sigma}_n}(\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| > M_n(n^{-1/2}\delta_n^{-1} + \delta_n^{\alpha-1}) | \boldsymbol{Y}) \to 0. \tag{5.15}$$

$$\textit{Hierarchical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{W}^\alpha} \leq R} \mathrm{E}_0 \Pi(\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| > M_n(n^{-1/2}\delta_n^{-1} + \delta_n^{\alpha-1}) | \boldsymbol{Y}) \to 0. \tag{5.16}$$

Theorem 5.2. *Under Assumptions 2 and 4, if* $\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}_0\| = o_{P_0}(\delta_n)$ *uniformly over* $\|f_0\|_{\mathcal{W}^\alpha} \le R$ *with* $R > 0$*, then for any* $M_n \to \infty$*,*

$$\textit{Empirical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{W}^\alpha}\le R} \mathrm{E}_0\Pi_{\widehat{\sigma}_n}(|M - M_0| > M_n(n^{-1/2} + \delta_n^\alpha)|\boldsymbol{Y}) \to 0. \tag{5.17}$$

$$\textit{Hierarchical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{W}^\alpha}\le R} \mathrm{E}_0\Pi(|M - M_0| > M_n(n^{-1/2} + \delta_n^\alpha)|\boldsymbol{Y}) \to 0. \tag{5.18}$$

Before we prove the two theorems above, we will first discuss some of their important consequences. The condition $\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\| = o_{P_0}(\delta_n)$ arises as a result of (1.10) in the definition of $\mathcal{H}^\alpha((0,1)^d)$. It says that we should construct credible region around $\widetilde{\boldsymbol{\mu}}$ at the second stage that is big enough to cover its estimation error. Note that by balancing the orders of the two terms in the posterior convergence rates for Theorems 5.1 and 5.2, we have $\delta_n \asymp n^{-1/(2\alpha)}$. This choice of δ_n is optimal as shown in the corollary below.

Corollary 5.3. *Suppose Assumptions 2 and 4 hold with* $\alpha > 1 + \sqrt{1 + d/2}$*. If the convergence rate of* $\widetilde{\boldsymbol{\mu}}$ *is faster than* $n^{-1/(2\alpha)}$ *under* P_0*, and the localization parameter is* $\delta_n = n^{-1/(2\alpha)}$*, then for any* $M_n \to \infty$*,*

Empirical Bayes:

$$\mathrm{E}_0\Pi_{\widehat{\sigma}_n}(\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| > M_n n^{-(\alpha-1)/(2\alpha)}|\boldsymbol{Y}) \to 0, \quad \mathrm{E}_0\Pi_{\widehat{\sigma}_n}(|M - M_0| > M_n n^{-1/2}|\boldsymbol{Y}) \to 0.$$

Hierarchical Bayes:

$$\mathrm{E}_0\Pi(\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| > M_n n^{-(\alpha-1)/(2\alpha)}|\boldsymbol{Y}) \to 0, \quad \mathrm{E}_0\Pi(|M - M_0| > M_n n^{-1/2}|\boldsymbol{Y}) \to 0,$$

uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \le R$ *for any* $R > 0$*.*

Proof. Let ϵ_n be the convergence rate for $\widetilde{\boldsymbol{\mu}}$, i.e., $\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}_0\| = O_{P_0}(\epsilon_n)$. By the discussion preceding Corollary 1 of Belitser et al. (2012), we choose $\delta_n = \max(\rho_n\epsilon_n, n^{-1/(2\alpha)})$ where ρ_n is a slowly varying sequence tending to infinity. Since $\epsilon_n \ll n^{-1/(2\alpha)}$ by assumption, we have $\delta_n = n^{-1/(2\alpha)}$ and $\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\| = o_{P_0}(\delta_n)$. Hence, Theorem 5.1 applies and the posterior convergence rate for $\boldsymbol{\mu}$ in this case follows from substituting $\delta_n = n^{-1/(2\alpha)}$. By Theorem 5.2, the posterior convergence rate for M is $\max(\rho_n^\alpha\epsilon_n^\alpha, n^{-1/2})$. Hence if $\epsilon_n \ll n^{-1/(2\alpha)}$, the rate becomes $n^{-1/2}$. By Belitser et al. (2012), the minimax rate for estimating $\boldsymbol{\mu}$ is $\epsilon_n = n^{-(\alpha-1)/(2\alpha+d)}$. Then $\epsilon_n \ll n^{-1/(2\alpha)}$ if $(\alpha-1)/(2\alpha+d) > 1/(2\alpha)$ or equivalently $\alpha > 1 + \sqrt{1 + d/2}$. The corresponding minimax rate for M is $n^{-\alpha/(2\alpha+d)}$ and it is $o(n^{-1/(2\alpha)})$ under the same lower bound for α. □

Remark If we used $\widehat{\boldsymbol{\mu}}_n$ defined in (4.9) as the preliminary estimator for $\boldsymbol{\mu}_0$, the first stage posterior convergence rate for $\boldsymbol{\mu}$ is $(\log n/n)^{(\alpha-1)/(2\alpha+d)}$ by Theorem 4.4. Then take $\delta_n = n^{-1/(2\alpha)}$ in Theorem 5.1 and we have second stage posterior convergence rate as $n^{-(\alpha-1)/(2\alpha)}$ for $\alpha > 1 + \sqrt{1 + d/2}$. Similarly, the first stage posterior convergence rate for M is $(\log n/n)^{\alpha/(2\alpha+d)}$ by

Theorem 4.5. Then the second stage posterior convergence rate in Theorem 5.2 becomes $n^{-1/2}$ for the same δ_n and same lower bound on α.

Observe that the dimension d, which is present in the first stage posterior convergence rates $(\log n/n)^{(\alpha-1)/(2\alpha+d)}$ and $(\log n/n)^{\alpha/(2\alpha+d)}$ for $\boldsymbol{\mu}$ and M (Theorems 4.4 and 4.5), does not effect the corresponding second stage rates $n^{-(\alpha-1)/(2\alpha)}$ and $n^{-1/2}$ respectively. Therefore, if the true function is smooth enough such that $\alpha > 1 + \sqrt{1 + d/2}$, then the curse of dimensionality is nearly avoided. Moreover, the Bayesian two-stage procedure removes the logarithmic factors from the first stage posterior convergence rates. The lower bound on α increases with dimension d. If $\alpha > 3$, then the posterior convergence rates established in Corollary 5.3 are optimal for all dimensions such that $3 \geq 1 + \sqrt{1 + d/2}$, which is $d = 1, \ldots, 6$, and includes the most commonly used dimensions $d = 1, 2, 3$.

We now prove Theorems 5.1 and 5.2. Before we can actually do that, we will first center the second stage design points and establish some preliminary results. Centering is a simplifying device that will make the analysis easier without changing the basic problem structure. Hence, we start by centering $\boldsymbol{z}_i = \boldsymbol{x}_i - \widetilde{\boldsymbol{\mu}}$ for $i = 1, \ldots, n_2$. Define $C(v\delta_n)$ to be a cube centered at $\mathbf{0}$ with radius $v\delta_n$, that is $C(v\delta_n) = [-v\delta_n, v\delta_n]^d \subseteq \mathbb{R}^d$. For notational simplicity, we write $\mathbb{I}(m_\alpha)$ as $\mathbb{I}$. Let $\boldsymbol{u}_j = \widetilde{\boldsymbol{u}}_j - \widetilde{\boldsymbol{\mu}}$ for $j = 1, \ldots, (2v+1)^d$. Then for all $i = 1, \ldots, n_2$, we have

$$\boldsymbol{z}_i \in \{0, \pm\delta_n, \ldots, \pm v\delta_n\}^d = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{(2v+1)^d}\} \subseteq C(v\delta_n). \tag{5.19}$$

Our shifted domain is now a cube of radius $v\delta_n$ centered at $\mathbf{0}$ instead of the original $\widetilde{\boldsymbol{\mu}}$. Hence, we are sampling $n_3 = n_2/(2v+1)^d$ times at each cube vertices to form $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{n_2}\}$. In view of (5.1), we represent $f(\boldsymbol{x})$ at $\boldsymbol{x} \in C(v\delta_n)$ as

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{\boldsymbol{i} \in \mathbb{I}} \theta_{\boldsymbol{i}} \boldsymbol{x}^{\boldsymbol{i}} = \boldsymbol{p}(\boldsymbol{x})^T \boldsymbol{\theta}, \tag{5.20}$$

and it follows that $f_{\boldsymbol{\vartheta}}(\boldsymbol{x}) = f_{\boldsymbol{\theta}}(\boldsymbol{x} - \widetilde{\boldsymbol{\mu}})$ for any $\boldsymbol{x} \in C(\widetilde{\boldsymbol{\mu}}, v\delta_n)$. We endow $\boldsymbol{\theta}$ with the prior $\boldsymbol{\theta}|\sigma^2 \sim \mathrm{N}_{W+1}(\boldsymbol{\xi}, \sigma^2 \boldsymbol{V})$, where the entries of $\boldsymbol{\xi}$ do not depend on n and $\boldsymbol{V} = \mathrm{diag}(C\delta_n^{-|\boldsymbol{i}_j|} : j = 0, 1 \ldots, W)$ for some constant $C > 0$ not depending on n. Define the design matrix as $\boldsymbol{Z} = (\boldsymbol{p}(\boldsymbol{z}_1), \ldots, \boldsymbol{p}(\boldsymbol{z}_{n_2}))^T$. Then it follows that $\Pi_{\widehat{\sigma}_n}(D^{\boldsymbol{r}} f_{\boldsymbol{\theta}} | \boldsymbol{Y}) \sim \mathrm{GP}(\boldsymbol{T}_{\boldsymbol{z},\boldsymbol{r}} \boldsymbol{Y} + \boldsymbol{d}_{\boldsymbol{z},\boldsymbol{r}} \boldsymbol{\xi}, \widehat{\sigma}_{\boldsymbol{z}}^2 \boldsymbol{H}_{\boldsymbol{z},\boldsymbol{r}})$, such that for $\boldsymbol{x}, \boldsymbol{y} \in C(v\delta_n)$,

$$\boldsymbol{T}_{\boldsymbol{z},\boldsymbol{r}}(\boldsymbol{x}) = \boldsymbol{p}_{\boldsymbol{r}}(\boldsymbol{x})^T \boldsymbol{N}_{\boldsymbol{r}} (\boldsymbol{Z}^T \boldsymbol{Z} + \boldsymbol{V}^{-1})^{-1} \boldsymbol{Z}^T, \tag{5.21}$$

$$\boldsymbol{d}_{\boldsymbol{z},\boldsymbol{r}}(\boldsymbol{x}) = \boldsymbol{p}_{\boldsymbol{r}}(\boldsymbol{x})^T \boldsymbol{N}_{\boldsymbol{r}} (\boldsymbol{Z}^T \boldsymbol{Z} + \boldsymbol{V}^{-1})^{-1} \boldsymbol{V}^{-1}, \tag{5.22}$$

$$\boldsymbol{H}_{\boldsymbol{z},\boldsymbol{r}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{p}_{\boldsymbol{r}}(\boldsymbol{x})^T \boldsymbol{N}_{\boldsymbol{r}} (\boldsymbol{Z}^T \boldsymbol{Z} + \boldsymbol{V}^{-1})^{-1} \boldsymbol{N}_{\boldsymbol{r}}^T \boldsymbol{p}_{\boldsymbol{r}}(\boldsymbol{y}). \tag{5.23}$$

Here, $\widehat{\sigma}_{\boldsymbol{z}}^2$ is

$$\widehat{\sigma}_{\boldsymbol{z}}^2 = \frac{(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\xi})^T(\boldsymbol{Z}\boldsymbol{V}\boldsymbol{Z}^T + \boldsymbol{I}_{n_2})^{-1}(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\xi})}{n_2}. \tag{5.24}$$

Note that we interpret $D^{\mathbf{0}} f_{\boldsymbol{\theta}} \equiv f_{\boldsymbol{\theta}}$. For the hierarchical Bayes approach, we endow σ^2 with the inverse-gamma prior in (3.15). The corresponding posterior is given by

$$\sigma^2 | \boldsymbol{Y} \sim \mathrm{IG}\left(\frac{\beta + n_2}{2}, \frac{\gamma + n_2 \widehat{\sigma}_{\boldsymbol{z}}^2}{2}\right), \tag{5.25}$$

with $\beta > 4$ and $\gamma > 0$. By shifting, the design points $\boldsymbol{z}_i$'s are symmetrically centered around zero at each dimension. Hence $\boldsymbol{Z}^T\boldsymbol{Z}$ will have a lot of zero entries because each entry contain products of elements in $\boldsymbol{z}_i$, and the entry is zero if some element is zero. Therefore, this simplifies calculations especially in deriving upper bounds for the posterior mean and variance. Define $f_{0,\boldsymbol{z}}(\boldsymbol{x} - \widetilde{\boldsymbol{\mu}}) = f_0(\boldsymbol{x})$ to be the shifted true function. Let $\boldsymbol{\theta}_0 = (\theta_{0,\boldsymbol{i}} : \boldsymbol{i} \in \mathbb{I})^T$ be a random vector such that $f_{\boldsymbol{\theta}_0}(\boldsymbol{x} - \widetilde{\boldsymbol{\mu}}) = T_{\boldsymbol{\mu}_0} f_0(\boldsymbol{x})$, where $T_{\boldsymbol{\mu}_0} f_0(\boldsymbol{x})$ is the Taylor polynomial of order m_α defined in (1.11) by expanding f_0 around $\boldsymbol{\mu}_0$, that is,

$$\sum_{\boldsymbol{i} \in \mathbb{I}} \theta_{0,\boldsymbol{i}} (\boldsymbol{x} - \widetilde{\boldsymbol{\mu}})^{\boldsymbol{i}} = f_0(\boldsymbol{\mu}_0) + \sum_{\boldsymbol{i} \in \mathbb{I}, |\boldsymbol{i}| \geq 2} \frac{D^{\boldsymbol{i}} f_0(\boldsymbol{\mu}_0)}{\boldsymbol{i}!} (\boldsymbol{x} - \boldsymbol{\mu}_0)^{\boldsymbol{i}}, \tag{5.26}$$

where $\nabla f_0(\boldsymbol{\mu}_0) = \mathbf{0}$ by Assumptions 2 and 4. Hence $\boldsymbol{\theta}_0$ can be thought of as the true $\boldsymbol{\theta}$ by projecting f_0 onto $\mathcal{P}_{m_\alpha}$. Note that $\boldsymbol{\theta}_0$ is random and depends on $\widetilde{\boldsymbol{\mu}}$, $\boldsymbol{\mu}_0$ and f_0. By applying $D^{\boldsymbol{i}}$ on both sides of (5.26) and evaluate $\boldsymbol{x}$ at $\widetilde{\boldsymbol{\mu}}$, we have $\boldsymbol{i}!\theta_{\boldsymbol{i}} = D^{\boldsymbol{i}} T_{\boldsymbol{\mu}_0} f_0(\widetilde{\boldsymbol{\mu}})$. Note that since $D^{\boldsymbol{i}} f_0(\boldsymbol{x}), \boldsymbol{i} \in \mathbb{I}$, are continuous by Assumption 2, they are bounded over $C(\widetilde{\boldsymbol{\mu}}, v\delta_n)$, and this implies that for any $\boldsymbol{i} \in \mathbb{I}$,

$$|\theta_{0,\boldsymbol{i}}| = O_{P_0}(1) \tag{5.27}$$

uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$. By Lemma 1 of Belitser et al. (2012), $\boldsymbol{Z}$ has full column rank, which is $W + 1$, and it follows that $\boldsymbol{Z}^T\boldsymbol{Z}$ is invertible. By equation (17), Lemma 2 of Belitser et al. (2012), it follows that for some constants $C_1, C_2 > 0$, we have uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$,

$$\|\boldsymbol{F}_0 - \boldsymbol{Z}\boldsymbol{\theta}_0\|_\infty \leq C_1 \delta_n^\alpha + C_2 \|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^\alpha. \tag{5.28}$$

Recall that we enumerate the elements of $\mathbb{I}$ as $\{\boldsymbol{i}_0, \ldots, \boldsymbol{i}_W\}$. For the rest of this section, we will follow this convention and index the entries of vectors $\boldsymbol{\xi}$, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ by elements of $\mathbb{I}$. For matrices $\boldsymbol{Z}^T\boldsymbol{Z}$ and $\boldsymbol{V}$, we will enumerate their rows and columns starting from 0 and ending at W. Recall that $n_3 \geq cn$ for some constant $c > 0$. Since $n_2 = n_3(2v+1)^d$, this implies that

$n \lesssim n_2 \leq n$. Hence, for the purpose of asymptotic analysis, n and n_2 are exchangeable. The next theorem shows that $\widehat{\sigma}^2_{\boldsymbol{z}}$ defined in (5.24) is a $\sqrt{n}$-consistent estimator of σ_0^2, uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$ for $R > 0$.

Theorem 5.4. *If $\delta_n \asymp n^{-1/(2\alpha)}$ and $\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\| = o_{P_0}(\delta_n)$ uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$, we have*

$$\sup_{\|f_0\|_{\mathcal{W}^\alpha} \leq R} \mathrm{E}_0(\widehat{\sigma}^2_{\boldsymbol{z}} - \sigma_0^2)^2 = O\left(\frac{1}{n}\right). \tag{5.29}$$

Proof. Let $\boldsymbol{U} = (\boldsymbol{Z}\boldsymbol{V}\boldsymbol{Z}^T + \boldsymbol{I}_n)^{-1}$. By equation (33) of page 355 in Searle (1982), the absolute conditional bias $|\mathrm{E}(\widehat{\sigma}^2_{\boldsymbol{z}}|\boldsymbol{\theta}_0) - \sigma_0^2|$ is

$$\begin{aligned}
&\left|\frac{\sigma_0^2}{n}\mathrm{tr}(\boldsymbol{U}) - \sigma_0^2\right| + \frac{1}{n}(\boldsymbol{F}_0 - \boldsymbol{Z}\boldsymbol{\xi})^T\boldsymbol{U}(\boldsymbol{F}_0 - \boldsymbol{Z}\boldsymbol{\xi}) \\
&\lesssim \frac{1}{n}[\mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{U}) + (\boldsymbol{F}_0 - \boldsymbol{Z}\boldsymbol{\theta}_0)^T\boldsymbol{U}(\boldsymbol{F}_0 - \boldsymbol{Z}\boldsymbol{\theta}_0) + (\boldsymbol{Z}\boldsymbol{\theta}_0 - \boldsymbol{Z}\boldsymbol{\xi})^T\boldsymbol{U}(\boldsymbol{Z}\boldsymbol{\theta}_0 - \boldsymbol{Z}\boldsymbol{\xi})],
\end{aligned} \tag{5.30}$$

where we used $(\boldsymbol{x} + \boldsymbol{y})^T\boldsymbol{D}(\boldsymbol{x} + \boldsymbol{y}) \leq 2\boldsymbol{x}^T\boldsymbol{D}\boldsymbol{x} + 2\boldsymbol{y}^T\boldsymbol{D}\boldsymbol{y}$ for any $\boldsymbol{D} \geq \boldsymbol{0}$ in the last line. Let $\boldsymbol{P}_{\boldsymbol{Z}} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$. Suppose that $\boldsymbol{A}$ is an $m \times m$ matrix, $\boldsymbol{G}$ an $m \times r$ matrix, $\boldsymbol{T}$ an $r \times r$ matrix, and $\boldsymbol{W}$ an $r \times m$ matrix. Assume that $\boldsymbol{A}$ and $\boldsymbol{T}$ are invertible. Then, the binomial inverse theorem (see Theorem 18.2.8 of Harville, 1997 for a proof) says

$$(\boldsymbol{A} + \boldsymbol{G}\boldsymbol{T}\boldsymbol{W})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{G}(\boldsymbol{T}^{-1} + \boldsymbol{W}\boldsymbol{A}^{-1}\boldsymbol{G})^{-1}\boldsymbol{W}\boldsymbol{A}^{-1}. \tag{5.31}$$

Therefore, two applications of (5.31) to $\boldsymbol{U}$ yield

$$\boldsymbol{U} = (\boldsymbol{Z}\boldsymbol{V}\boldsymbol{Z}^T + \boldsymbol{I}_n)^{-1} = \boldsymbol{I}_n - \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + \boldsymbol{V}^{-1})^{-1}\boldsymbol{Z}^T = \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Z}} + \boldsymbol{C}, \tag{5.32}$$

where $\boldsymbol{C} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}[\boldsymbol{V} + (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}]^{-1}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T \geq \boldsymbol{0}$. Hence the first term in (5.30) is

$$\frac{1}{n}\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{C}) \leq \frac{1}{n}\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{Z}}) = \frac{W+1}{n}. \tag{5.33}$$

Note that $\boldsymbol{U} \leq \boldsymbol{I}_n$ since $\boldsymbol{Z}\boldsymbol{V}\boldsymbol{Z}^T \geq \boldsymbol{0}$, and the second term in (5.30) is bounded by

$$\frac{1}{n}\|\boldsymbol{U}\|_{(2,2)}\|\boldsymbol{F}_0 - \boldsymbol{Z}\boldsymbol{\theta}_0\|^2 \leq \|\boldsymbol{F}_0 - \boldsymbol{Z}\boldsymbol{\theta}_0\|^2_\infty \lesssim \delta_n^{2\alpha} + \|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^{2\alpha}, \tag{5.34}$$

in view of (5.28). By (5.32) and the fact that $(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{Z}=\boldsymbol{0}$, the last term in (5.30) is

$$\begin{aligned}\frac{1}{n}(\boldsymbol{\theta}_0-\boldsymbol{\xi})^T[\boldsymbol{V}+(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}]^{-1}(\boldsymbol{\theta}_0-\boldsymbol{\xi}) &\leq \frac{1}{n}(\boldsymbol{\theta}_0-\boldsymbol{\xi})^T\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0-\boldsymbol{\xi}) \\ &= \frac{1}{n}\sum_{j=0}^{W}\delta_n^{|\boldsymbol{i}_j|}(\theta_{0,\boldsymbol{i}_j}-\xi_{\boldsymbol{i}_j})^2 = O_{P_0}\left(\frac{1}{n}\right), \qquad (5.35)\end{aligned}$$

where we have used the fact that $\delta_n = o(1)$, $\theta_{0,\boldsymbol{i}_j} = O_{P_0}(1)$ by (5.27) and $\xi_{\boldsymbol{i}_j} = O(1)$ by assumption on the prior for any $0 \leq j \leq W$. Combining the bounds in (5.33), (5.34) and (5.35) into (5.30), we obtain $|\mathrm{E}(\widehat{\sigma}^2_{\boldsymbol{z}}|\boldsymbol{\theta}_0)-\sigma_0^2| = O_{P_0}(1/n)+\delta_n^{2\alpha}+\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}_0\|^{2\alpha}$.

Let $\boldsymbol{Y}=\boldsymbol{F}_0+\boldsymbol{\varepsilon}$ and write $n\widehat{\sigma}^2_{\boldsymbol{z}} = (\boldsymbol{F}_0-\boldsymbol{Z}\boldsymbol{\xi})^T\boldsymbol{U}(\boldsymbol{F}_0-\boldsymbol{Z}\boldsymbol{\xi})+2(\boldsymbol{F}_0-\boldsymbol{Z}\boldsymbol{\xi})^T\boldsymbol{U}\boldsymbol{\varepsilon}+\boldsymbol{\varepsilon}^T\boldsymbol{U}\boldsymbol{\varepsilon}$. Observe that $\boldsymbol{\varepsilon}$ and $\boldsymbol{\theta}_0$ are independent by definition. Using the fact $\mathrm{Var}(T_1+T_2) \leq 2\mathrm{Var}(T_1)+2\mathrm{Var}(T_2)$, it follows that $\mathrm{Var}(\widehat{\sigma}^2_{\boldsymbol{z}}|\boldsymbol{\theta}_0)$ is bounded up to a constant multiple by

$$\frac{1}{n^2}[(\boldsymbol{F}_0-\boldsymbol{Z}\boldsymbol{\theta}_0)^T\boldsymbol{U}^2(\boldsymbol{F}_0-\boldsymbol{Z}\boldsymbol{\theta}_0)+(\boldsymbol{Z}\boldsymbol{\theta}_0-\boldsymbol{Z}\boldsymbol{\xi})^T\boldsymbol{U}^2(\boldsymbol{Z}\boldsymbol{\theta}_0-\boldsymbol{Z}\boldsymbol{\xi})+\mathrm{Var}(\boldsymbol{\varepsilon}^T\boldsymbol{U}\boldsymbol{\varepsilon}). \qquad (5.36)$$

In view of (5.28) and $\boldsymbol{U} \leq \boldsymbol{I}_n$, the first term above is bounded by

$$\frac{1}{n^2}\|\boldsymbol{U}\|^2_{(2,2)}\|\boldsymbol{F}_0-\boldsymbol{Z}\boldsymbol{\theta}_0\|^2 \leq \frac{1}{n}\|\boldsymbol{F}_0-\boldsymbol{Z}\boldsymbol{\theta}_0\|^2_\infty \lesssim \frac{1}{n}(\delta_n^{2\alpha}+\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}_0\|^{2\alpha}). \qquad (5.37)$$

Observe that since $\boldsymbol{V}>\boldsymbol{0}$, we have

$$\begin{aligned}\boldsymbol{Z}^T\boldsymbol{C}^2\boldsymbol{Z} &= [\boldsymbol{V}+(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}]^{-1}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}[\boldsymbol{V}+(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}]^{-1} \\ &< [\boldsymbol{V}+(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}]^{-1} < \boldsymbol{Z}^T\boldsymbol{Z}. \qquad (5.38)\end{aligned}$$

Using (5.32), idempotency of $\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{Z}}$ and $(\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{Z}=\boldsymbol{0}$, the second term in (5.36) can be written as

$$\begin{aligned}\frac{1}{n^2}(\boldsymbol{\theta}_0-\boldsymbol{\xi})^T\boldsymbol{Z}^T(\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{Z}}+\boldsymbol{C})^2\boldsymbol{Z}(\boldsymbol{\theta}_0-\boldsymbol{\xi}) &= \frac{1}{n^2}(\boldsymbol{\theta}_0-\boldsymbol{\xi})^T\boldsymbol{Z}^T\boldsymbol{C}^2\boldsymbol{Z}(\boldsymbol{\theta}_0-\boldsymbol{\xi}) \\ &\leq \frac{1}{n^2}(\boldsymbol{\theta}_0-\boldsymbol{\xi})^T\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{\theta}_0-\boldsymbol{\xi}), \qquad (5.39)\end{aligned}$$

where the last line follows from (5.38). In the proof of Lemma 3 in Belitser et al. (2012), $\boldsymbol{Z}^T\boldsymbol{Z} = n_3\boldsymbol{\Delta}\boldsymbol{A}\boldsymbol{\Delta}$ where $\boldsymbol{\Delta} = \mathrm{diag}(\delta_n^{|\boldsymbol{i}_j|} : j=0,\ldots,W)$ and $\boldsymbol{A}$ is an $(W+1)\times(W+1)$ invertible matrix of constants. Note that $\|\boldsymbol{\Delta}\|_{(2,2)}=1$. Using the fact $n_3 \lesssim n$, this further implies that

the right hand side of (5.39) is bounded by

$$\frac{n_3}{n^2}\|\boldsymbol{A}\|_{(2,2)}\|\boldsymbol{\Delta}\|_{(2,2)}^2\|\boldsymbol{\theta}_0-\boldsymbol{\xi}\|^2 = O_{P_0}\left(\frac{1}{n}\right), \tag{5.40}$$

where $\|\boldsymbol{\theta}_0-\boldsymbol{\xi}\|^2=\sum_{j=0}^{W}(\theta_{0,\boldsymbol{i}_j}-\xi_{\boldsymbol{i}_j})^2=O_{P_0}(1)$ is bounded using (5.27) and the assumption on $\boldsymbol{\xi}$. By Lemma D.3, the last term in (5.36) is $O(1/n)$. Combining this with the bounds established in (5.37) and (5.40) into (5.36), we obtain $\mathrm{Var}(\widehat{\sigma}_{\boldsymbol{z}}^2|\boldsymbol{\theta}_0)=O_{P_0}(1/n)$. By taking $\delta_n \asymp n^{-1/(2\alpha)}$, the mean square error is

$$\mathrm{E}_0(\widehat{\sigma}_{\boldsymbol{z}}^2-\sigma_0^2)^2=\mathrm{E}\{\mathrm{E}[(\widehat{\sigma}_{\boldsymbol{z}}^2-\sigma_0^2)^2|\boldsymbol{\theta}_0]\}\lesssim\frac{1}{n}+\delta_n^{4\alpha}\lesssim\frac{1}{n}. \quad \square$$

A Bayesian analog of Theorem 5.4 above is given below, which shows that the posterior of σ^2 in (5.25) converges to σ_0^2 at the parametric rate $1/\sqrt{n}$.

Theorem 5.5. *If $\delta_n \asymp n^{-1/(2\alpha)}$ and $\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}_0\|=o_{P_0}(\delta_n)$ uniformly over $\|f_0\|_{\mathcal{W}^\alpha}\le R$, we have for $R>0$ and any $M_n\to\infty$,*

$$\sup_{\|f_0\|_{\mathcal{W}^\alpha}\le R}\mathrm{E}_0\Pi\left(|\sigma^2-\sigma_0^2|>\frac{M_n}{\sqrt{n}}\middle|\boldsymbol{Y}\right)\to 0. \tag{5.41}$$

The proof of Theorem 5.5 is identical to the proof of Theorem 3.2. For notational simplicity, we write $\sigma^2\in\mathcal{U}_n$ if $|\sigma^2-\sigma_0^2|\le\zeta_n/\sqrt{n}$ for $\zeta_n=o(\sqrt{n})$.

Lemma 5.6. *Under Assumption 2, we have for any $\boldsymbol{i}\in\mathbb{I}$ that*

$$\sup_{\sigma^2\in\mathcal{U}_n}\mathrm{E}[(\theta_{\boldsymbol{i}}-\theta_{0,\boldsymbol{i}})^2|\boldsymbol{Y},\sigma^2]=O_{P_0}(n^{-1}\delta_n^{-2|\boldsymbol{i}|})+O(\delta_n^{2\alpha-2|\boldsymbol{i}|})+O(\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}_0\|^{2\alpha}\delta_n^{-2|\boldsymbol{i}|}).$$

Proof. Let $0\le h\le W$. Since $\boldsymbol{V}>\boldsymbol{0}$ by assumption, we have by Lemma 3 of Belitser et al. (2012) that the conditional posterior variance is bounded by

$$\sup_{\sigma^2\in\mathcal{U}_n}\mathrm{Var}(\theta_{\boldsymbol{i}_h}|\boldsymbol{Y},\sigma^2)=(\sigma_0^2+\zeta_n/\sqrt{n})[(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}]_{h,h}\lesssim[(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}]_{h,h}\lesssim\frac{1}{n}\delta_n^{-2|\boldsymbol{i}_h|}. \tag{5.42}$$

Now the bias for the conditional posterior mean in vector form is

$$\begin{aligned}\mathrm{E}(\boldsymbol{\theta}|\boldsymbol{Y},\sigma^2)-\boldsymbol{\theta}_0&=(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}(\boldsymbol{Z}^T\boldsymbol{Y}+\boldsymbol{V}^{-1}\boldsymbol{\xi})-\boldsymbol{\theta}_0\\&=(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}[\boldsymbol{Z}^T\boldsymbol{\varepsilon}+\boldsymbol{Z}^T(\boldsymbol{F}_0-\boldsymbol{Z}\boldsymbol{\theta}_0)+\boldsymbol{V}^{-1}(\boldsymbol{\xi}-\boldsymbol{\theta}_0)].\end{aligned} \tag{5.43}$$

Since $\mathrm{E}_0(\boldsymbol{\varepsilon})=\boldsymbol{0}$ and $\mathrm{Var}_0[(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\varepsilon}]=\sigma_0^2(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}$, it follows from Markov's inequality that the order of each entry in $(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\varepsilon}$ is bounded

by the diagonal entries of its corresponding covariance matrix. Following the same reasoning as in (5.42),

$$\begin{aligned}\sigma_0^2[(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}]_{h,h} &< \sigma_0^2[\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}]_{h,h} \\ &< \sigma_0^2[(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}]_{h,h} \lesssim \frac{1}{n}\delta_n^{-2|\boldsymbol{i}_h|}. \end{aligned} \tag{5.44}$$

Therefore, this implies that

$$(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\varepsilon} = ((O_{P_0}(n^{-1/2}), O_{P_0}(n^{-1/2}\delta_n^{-|\boldsymbol{i}_1|}, \ldots, O_{P_0}(n^{-1/2}\delta_n^{-|\boldsymbol{i}_W|}))^T. \tag{5.45}$$

Let β_{ij} be the (i,j)th element of $(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}$, v^{ij} be the (i,j)th element of $\boldsymbol{V}^{-1}$ and γ_i be the ith entry of $\boldsymbol{F}_0 - \boldsymbol{Z}\boldsymbol{\theta}_0$. By (5.28), we have uniformly over $1 \le i \le n$ that $|\gamma_i| \lesssim \delta_n^\alpha + \|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^\alpha$. Now using the fact that for $\boldsymbol{A} > \boldsymbol{0}$, $a_{ij} \le \sqrt{a_{ii}a_{jj}}$ by the Cauchy-Schwarz inequality, we have for $0 \le h, j \le W$,

$$\begin{aligned}[(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}]_{h,j} &\le \sqrt{[(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}]_{h,h}[(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}]_{j,j}} \\ &< \sqrt{[(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}]_{h,h}[(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}]_{j,j}} \\ &\lesssim \frac{1}{n}\delta_n^{-(|\boldsymbol{i}_h|+|\boldsymbol{i}_j|)}. \end{aligned} \tag{5.46}$$

Observe that since $\boldsymbol{z}_j \in C(v\delta_n)$, $j = 1, \ldots, n_2$, we have $|\boldsymbol{z}_j^{\boldsymbol{i}}| \le v^{m_\alpha}\delta_n^{|\boldsymbol{i}|}$ for $\boldsymbol{i} \in \mathbb{I}$. Therefore, using the fact that $n_2 \le n$, $(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}\boldsymbol{Z}^T(\boldsymbol{F}_0 - \boldsymbol{Z}\boldsymbol{\theta}_0)$ is

$$\begin{pmatrix} \beta_{00}\sum_{j=1}^{n_2}\boldsymbol{z}_j^{\boldsymbol{i}_0}\gamma_j + \ldots + \beta_{0W}\sum_{j=1}^{n_2}\boldsymbol{z}_j^{\boldsymbol{i}_W}\gamma_j \\ \beta_{10}\sum_{j=1}^{n_2}\boldsymbol{z}_j^{\boldsymbol{i}_0}\gamma_j + \ldots + \beta_{1W}\sum_{j=1}^{n_2}\boldsymbol{z}_j^{\boldsymbol{i}_W}\gamma_j \\ \vdots \\ \beta_{W0}\sum_{j=1}^{n_2}\boldsymbol{z}_j^{\boldsymbol{i}_0}\gamma_j + \ldots + \beta_{WW}\sum_{j=1}^{n_2}\boldsymbol{z}_j^{\boldsymbol{i}_W}\gamma_j \end{pmatrix} = \begin{pmatrix} O(\delta_n^\alpha) + O(\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^\alpha) \\ O(\delta_n^{\alpha-|\boldsymbol{i}_1|}) + O(\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^\alpha\delta_n^{-|\boldsymbol{i}_1|}) \\ \vdots \\ O(\delta_n^{\alpha-|\boldsymbol{i}_W|}) + O(\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^\alpha\delta_n^{-|\boldsymbol{i}_W|}) \end{pmatrix}. \tag{5.47}$$

It then remains to bound each entry of the last term in (5.43). Note that $|\theta_{0,\boldsymbol{i}_j} - \xi_{0,\boldsymbol{i}_j}| = O_{P_0}(1)$ for $j = 0, \ldots, W$ by (5.27) and the assumption on $\boldsymbol{\xi}$. By Lemma 3 of Belitser et al. (2012) and

the assumption on $\boldsymbol{V}$, $(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}\boldsymbol{V}^{-1}(\boldsymbol{\xi}-\boldsymbol{\theta}_0)$ is

$$\begin{pmatrix} \beta_{00}\sum_{j=0}^{W} v^{0j}(\xi_{\boldsymbol{i}_j}-\theta_{0,\boldsymbol{i}_j})+\ldots+\beta_{0W}\sum_{j=0}^{W} v^{Wj}(\xi_{\boldsymbol{i}_j}-\theta_{0,\boldsymbol{i}_j}) \\ \beta_{10}\sum_{j=0}^{W} v^{0j}(\xi_{\boldsymbol{i}_j}-\theta_{0,\boldsymbol{i}_j})+\ldots+\beta_{1W}\sum_{j=0}^{W} v^{Wj}(\xi_{\boldsymbol{i}_j}-\theta_{0,\boldsymbol{i}_j}) \\ \vdots \\ \beta_{W0}\sum_{j=0}^{W} v^{0j}(\xi_{\boldsymbol{i}_j}-\theta_{0,\boldsymbol{i}_j})+\ldots+\beta_{WW}\sum_{j=0}^{W} v^{Wj}(\xi_{\boldsymbol{i}_j}-\theta_{0,\boldsymbol{i}_j}) \end{pmatrix} = \begin{pmatrix} O_{P_0}(n^{-1}) \\ O_{P_0}(n^{-1}\delta_n^{-|\boldsymbol{i}_1|}) \\ \vdots \\ O_{P_0}(n^{-1}\delta_n^{-|\boldsymbol{i}_W|}) \end{pmatrix}. \tag{5.48}$$

Hence, combining the bounds established into (5.43), we have that the squared bias is

$$(\mathrm{E}(\theta_{\boldsymbol{i}_h}|\boldsymbol{Y},\sigma^2)-\theta_{0,\boldsymbol{i}_h})^2 = O_{P_0}(n^{-1}\delta_n^{-2|\boldsymbol{i}_h|})+O(\delta_n^{2\alpha-2|\boldsymbol{i}_h|})+O(\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}_0\|^{2\alpha}\delta_n^{-2|\boldsymbol{i}_h|}). \tag{5.49}$$

Combining the above with (5.42), we conclude that for any $\boldsymbol{i}\in\mathbb{I}$,

$$\sup_{\sigma^2\in\mathcal{U}_n}\mathrm{E}((\theta_{\boldsymbol{i}}-\theta_{0,\boldsymbol{i}})^2|\boldsymbol{Y},\sigma^2) = O_{P_0}(n^{-1}\delta_n^{-2|\boldsymbol{i}|})+O(\delta_n^{2\alpha-2|\boldsymbol{i}|})+O(\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}_0\|^{2\alpha}\delta_n^{-2|\boldsymbol{i}|}). \quad \square$$

Remark From (5.48), we see the reason for choosing the prior variance as $\boldsymbol{V}=\mathrm{diag}(C\delta_n^{-|\boldsymbol{i}_j|}: j=0,\ldots,W)$ with $C>0$. The diagonal entries of $\boldsymbol{V}$ are used to cancel out one of the δ_n factors in each the entries of $(\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{V}^{-1})^{-1}$. Another possible prior on $\boldsymbol{\theta}$ is Zellner's prior, i.e., $\boldsymbol{\theta}\sim \mathrm{N}_{W+1}(\boldsymbol{0},\lambda(\boldsymbol{Z}^T\boldsymbol{Z})^{-1})$ or $\boldsymbol{\vartheta}\sim\mathrm{N}_{W+1}(\boldsymbol{0},\lambda(\boldsymbol{P}^T\boldsymbol{P})^{-1})$ in the original design points for $\lambda>0$. We see that Theorems 5.4, 5.5 and Lemma 5.6 still hold under this prior with some modifications. This is the prior that we will use for our simulation study (see Section 5.3).

Theorem 5.7. *Under Assumptions 2 and 4, if $\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}_0\|=o_{P_0}(\delta_n)$ uniformly over $\|f_0\|_{\mathcal{W}^\alpha}\le R$, we have for any $|\boldsymbol{r}|\le m_\alpha$, $\boldsymbol{x}\in C(v\delta_n)$ and $M_n\to\infty$, uniformly over $\|f_0\|_{\mathcal{W}^\alpha}\le R$ that*

$$\textit{Empirical Bayes: } \mathrm{E}_0\Pi_{\widehat{\sigma}_n}(|D^{\boldsymbol{r}}f_{\boldsymbol{\theta}}(\boldsymbol{x})-D^{\boldsymbol{r}}f_{0,\boldsymbol{z}}(\boldsymbol{x})|>M_n(n^{-1/2}\delta_n^{-|\boldsymbol{r}|}+\delta_n^{\alpha-|\boldsymbol{r}|})|\boldsymbol{Y})\to 0.$$
$$\textit{Hierarchical Bayes: } \mathrm{E}_0\Pi(|D^{\boldsymbol{r}}f_{\boldsymbol{\theta}}(\boldsymbol{x})-D^{\boldsymbol{r}}f_{0,\boldsymbol{z}}(\boldsymbol{x})|>M_n(n^{-1/2}\delta_n^{-|\boldsymbol{r}|}+\delta_n^{\alpha-|\boldsymbol{r}|})|\boldsymbol{Y})\to 0.$$

Proof. In view of (5.10) with $\theta_{\boldsymbol{i}}$ instead of $\vartheta_{\boldsymbol{i}}$, we have

$$D^{\boldsymbol{r}}f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{\boldsymbol{i}\in\mathbb{I}:\boldsymbol{i}\ge\boldsymbol{r}}\frac{\boldsymbol{i}!}{(\boldsymbol{i}-\boldsymbol{r})!}\theta_{\boldsymbol{i}}\boldsymbol{x}^{\boldsymbol{i}-\boldsymbol{r}} = \boldsymbol{r}!\theta_{\boldsymbol{r}}+\sum_{\boldsymbol{i}\in\mathbb{I}:\boldsymbol{i}\ge\boldsymbol{r},\boldsymbol{i}\ne\boldsymbol{r}}\frac{\boldsymbol{i}!}{(\boldsymbol{i}-\boldsymbol{r})!}\theta_{\boldsymbol{i}}\boldsymbol{x}^{\boldsymbol{i}-\boldsymbol{r}}. \tag{5.50}$$

Observe that for any $\boldsymbol{x}\in C(v\delta_n)$, $|\boldsymbol{x}^{\boldsymbol{i}-\boldsymbol{r}}|\le v^{m_\alpha}\delta_n^{|\boldsymbol{i}|-|\boldsymbol{r}|}$. Also, by noting that $r_k\le i_k\le m_\alpha$ for

$k = 1, \ldots, d$, we have $\boldsymbol{r}! \le m_\alpha^{dm_\alpha}$ and $\boldsymbol{i}! \le m_\alpha^{dm_\alpha}$. Therefore,

$$\begin{aligned}|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_{\boldsymbol{\theta}_0}(\boldsymbol{x})| &\le \boldsymbol{r}!|\theta_{\boldsymbol{r}} - \theta_{0,\boldsymbol{r}}| + \sum_{\boldsymbol{i}\in\mathbb{I}:\boldsymbol{i}\ge\boldsymbol{r},\boldsymbol{i}\ne\boldsymbol{r}} \frac{\boldsymbol{i}!}{(\boldsymbol{i}-\boldsymbol{r})!}|(\theta_{\boldsymbol{i}} - \theta_{0,\boldsymbol{i}})\boldsymbol{x}^{\boldsymbol{i}-\boldsymbol{r}}| \\ &\le m_\alpha^{dm_\alpha}|\theta_{\boldsymbol{r}} - \theta_{0,\boldsymbol{r}}| + m_\alpha^{dm_\alpha} v^{m_\alpha} \sum_{\boldsymbol{i}\in\mathbb{I}:\boldsymbol{i}\ge\boldsymbol{r},\boldsymbol{i}\ne\boldsymbol{r}} |\theta_{\boldsymbol{i}} - \theta_{0,\boldsymbol{i}}|\delta_n^{|\boldsymbol{i}|-|\boldsymbol{r}|}. \end{aligned} \tag{5.51}$$

Using the fact $(\sum_{i=1}^n |a_i|)^p \le n^{p-1}\sum_{i=1}^n |a_i|^p$ for $p \ge 1$ and squaring both sides above yield

$$|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_{\boldsymbol{\theta}_0}(\boldsymbol{x})|^2 \le 2m_\alpha^{2dm_\alpha}|\theta_{\boldsymbol{r}} - \theta_{0,\boldsymbol{r}}|^2 + 2|\mathbb{I}|m_\alpha^{2dm_\alpha} v^{2m_\alpha} \sum_{\boldsymbol{i}\in\mathbb{I}:\boldsymbol{i}\ge\boldsymbol{r},\boldsymbol{i}\ne\boldsymbol{r}} |\theta_{\boldsymbol{i}} - \theta_{0,\boldsymbol{i}}|^2\delta_n^{2|\boldsymbol{i}|-2|\boldsymbol{r}|}.$$

Therefore, for any $|\boldsymbol{r}| \le m_\alpha$ and any $\boldsymbol{x} \in C(v\delta_n)$, we have uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \le R$ that

$$\begin{aligned}&\mathrm{E}_0 \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}(|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_{\boldsymbol{\theta}_0}(\boldsymbol{x})|^2|\boldsymbol{Y}, \sigma^2) \\ &\lesssim \mathrm{E}_0 \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}[(\theta_{\boldsymbol{r}} - \theta_{0,\boldsymbol{r}})^2|\boldsymbol{Y}, \sigma^2] + \sum_{\boldsymbol{i}\in\mathbb{I}:\boldsymbol{i}\ge\boldsymbol{r},\boldsymbol{i}\ne\boldsymbol{r}} \delta_n^{2|\boldsymbol{i}|-2|\boldsymbol{r}|}\mathrm{E}_0 \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}[(\theta_{\boldsymbol{i}} - \theta_{0,\boldsymbol{i}})^2|\boldsymbol{Y}, \sigma^2]. \end{aligned} \tag{5.52}$$

Observe that $\sup_{\|f_0\|_{\mathcal{W}^\alpha}\le R} \mathrm{E}_0\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\| = o(\delta_n)$ by assumption. In view of Lemma 5.6, the first term above is bounded by $n^{-1}\delta_n^{-2|\boldsymbol{r}|} + \delta_n^{2\alpha-2|\boldsymbol{r}|}$. By noting that the sum over $\mathbb{I}$ such that $\boldsymbol{i} \ge \boldsymbol{r}$ and $\boldsymbol{i} \ne \boldsymbol{r}$ has at most $|\mathbb{I}|$ terms, where $|\mathbb{I}|$ depends only on α and d, another application of Lemma 5.6 implies that the second term in (5.52) is bounded above by

$$\sum_{\boldsymbol{i}\in\mathbb{I}:\boldsymbol{i}\ge\boldsymbol{r},\boldsymbol{i}\ne\boldsymbol{r}} \delta_n^{2|\boldsymbol{i}|-2}(n^{-1}\delta_n^{-2|\boldsymbol{i}|} + \delta_n^{2\alpha-2|\boldsymbol{i}|} + \mathrm{E}_0\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^{2\alpha}\delta_n^{-2|\boldsymbol{i}|}) \lesssim n^{-1}\delta_n^{-2|\boldsymbol{r}|} + \delta_n^{2\alpha-2|\boldsymbol{r}|}.$$

Observe that for $\boldsymbol{x} \in C(v\delta_n)$, we have $\|\boldsymbol{x}\| \le \sqrt{d}v\delta_n$. Using the c_r-inequality, $|a+b|^r \le \max(1, 2^{r-1})(|a|^r + |b|^r), r > 0$, and (1.11) in the definition of $\mathcal{H}^\alpha((0,1)^d)$, we have

$$\begin{aligned}|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}_0}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}(\boldsymbol{x})| &= |D^{\boldsymbol{r}} T_{\boldsymbol{\mu}_0}(\boldsymbol{x} + \widetilde{\boldsymbol{\mu}}) - D^{\boldsymbol{r}} f_0(\boldsymbol{x} + \widetilde{\boldsymbol{\mu}})| \\ &\le C\|\boldsymbol{x} + \widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^{\alpha-|\boldsymbol{r}|} \\ &\lesssim \|\boldsymbol{x}\|^{\alpha-|\boldsymbol{r}|} + \|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^{\alpha-|\boldsymbol{r}|} \\ &\lesssim \delta_n^{\alpha-|\boldsymbol{r}|} + \|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^{\alpha-|\boldsymbol{r}|}. \end{aligned} \tag{5.53}$$

Therefore, uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \le R$,

$$\mathrm{E}_0|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}_0}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}(\boldsymbol{x})|^2 \lesssim \delta_n^{2\alpha-2|\boldsymbol{r}|} + \mathrm{E}_0\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^{2\alpha-2|\boldsymbol{r}|} \lesssim \delta_n^{2\alpha-2|\boldsymbol{r}|}. \tag{5.54}$$

Combining the bounds established for (5.52) and (5.54), we conclude that uniformly over

$\|f_0\|_{\mathcal{W}^\alpha} \le R$,

$$\begin{aligned}\mathrm{E}_0 \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}[(D^{\boldsymbol{r}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}(\boldsymbol{x}))^2|\boldsymbol{Y},\sigma^2] &\le 2\mathrm{E}_0 \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}(|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_{\boldsymbol{\theta}_0}(\boldsymbol{x})|^2|\boldsymbol{Y},\sigma^2) \\ &\quad + 2\mathrm{E}_0|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}_0}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}(\boldsymbol{x})|^2 \\ &\lesssim n^{-1}\delta_n^{-2|\boldsymbol{r}|} + \delta_n^{2\alpha-2|\boldsymbol{r}|}. \end{aligned} \tag{5.55}$$

Observe that by Markov's inequality, Theorem 5.4 implies for $R > 0$ and any $M_n \to \infty$,

$$\inf_{\|f_0\|_{\mathcal{W}^\alpha}\le R} P_0\left(\sigma_0^2 - \frac{M_n}{\sqrt{n}} \le \widehat{\sigma}_{\boldsymbol{z}}^2 \le \sigma_0^2 + \frac{M_n}{\sqrt{n}}\right) \to 1. \tag{5.56}$$

For the empirical Bayes version, by Markov's inequality, (5.55) and (5.56) above, we have for $\epsilon_n = n^{-1/2}\delta_n^{-|\boldsymbol{r}|} + \delta_n^{\alpha-|\boldsymbol{r}|}$ and any $M_n \to \infty$,

$$\begin{aligned}&\mathrm{E}_0\Pi_{\widehat{\sigma}_n}(|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}(\boldsymbol{x})| > M_n\epsilon_n|\boldsymbol{Y}) \\ &\quad\le \frac{1}{M_n^2\epsilon_n^2}\mathrm{E}_0 \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}([D^{\boldsymbol{r}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}(\boldsymbol{x})]^2|\boldsymbol{Y},\sigma^2) + o(1) \to 0, \end{aligned} \tag{5.57}$$

uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \le R$. For the hierarchical Bayes procedure, we have for $\epsilon_n = n^{-1/2}\delta_n^{-|\boldsymbol{r}|} + \delta_n^{\alpha-|\boldsymbol{r}|}$ and any $M_n \to \infty$,

$$\begin{aligned}\mathrm{E}_0\Pi(|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}(\boldsymbol{x})| > M_n\epsilon_n|\boldsymbol{Y}) &\le \mathrm{E}_0 \sup_{\sigma^2\in\mathcal{U}_n} \Pi(|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}(\boldsymbol{x})| > M_n\epsilon_n|\boldsymbol{Y},\sigma^2) \\ &\quad + \mathrm{E}_0\Pi\left(|\sigma^2 - \sigma_0^2| > \frac{\zeta_n}{\sqrt{n}}\middle|\boldsymbol{Y}\right), \end{aligned} \tag{5.58}$$

uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \le R$. The first term on the right hand side is $o(1)$ by (5.55), while the second term goes to zero by Theorem 5.5. □

Corollary 5.8. *Under Assumptions 2 and 4, if $\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\| = o_{P_0}(\delta_n)$ uniformly in $\|f_0\|_{\mathcal{W}^\alpha} \le R$, we have for any $|\boldsymbol{r}| \le m_\alpha$ and $M_n \to \infty$ that*

$$\textit{Empirical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{W}^\alpha}\le R} \mathrm{E}_0\Pi_{\widehat{\sigma}_n}(\|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}} - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}\|_\infty > M_n(n^{-1/2}\delta_n^{-|\boldsymbol{r}|} + \delta_n^{\alpha-|\boldsymbol{r}|})|\boldsymbol{Y}) \to 0.$$

$$\textit{Hierarchical Bayes:} \quad \sup_{\|f_0\|_{\mathcal{W}^\alpha}\le R} \mathrm{E}_0\Pi(\|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}} - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}\|_\infty > M_n(n^{-1/2}\delta_n^{-|\boldsymbol{r}|} + \delta_n^{\alpha-|\boldsymbol{r}|})|\boldsymbol{Y}) \to 0.$$

Proof. By (5.51), we have

$$\begin{aligned}\|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}} - D^{\boldsymbol{r}} f_{\boldsymbol{\theta}_0}\|_\infty &= \sup_{\boldsymbol{x}\in C(v\delta_n)} |D^{\boldsymbol{r}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_{\boldsymbol{\theta}_0}(\boldsymbol{x})| \\ &\lesssim |\theta_{\boldsymbol{r}} - \theta_{0,\boldsymbol{r}}| + \sum_{\boldsymbol{i}\in\mathbb{I}:\boldsymbol{i}\geq\boldsymbol{r},\boldsymbol{i}\neq\boldsymbol{r}} |\theta_{\boldsymbol{i}} - \theta_{0,\boldsymbol{i}}|\delta_n^{|\boldsymbol{i}|-|\boldsymbol{r}|}. \end{aligned} \tag{5.59}$$

This implies that the same upper bounds established in (5.52) of Theorem 5.7 apply in the present case, therefore uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$,

$$\mathrm{E}_0 \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}[\|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}} - D^{\boldsymbol{r}} f_{\boldsymbol{\theta}_0}\|_\infty^2 | \boldsymbol{Y}, \sigma^2] \lesssim n^{-1}\delta_n^{-2|\boldsymbol{r}|} + \delta_n^{2\alpha-2|\boldsymbol{r}|}. \tag{5.60}$$

Moreover, since the bound in (5.53) is uniform for all $\boldsymbol{x} \in C(v\delta_n)$, it follows from (5.54) that

$$\sup_{\|f_0\|_{\mathcal{W}^\alpha}\leq R} \mathrm{E}_0\|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}_0} - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}\|_\infty^2 \lesssim \delta_n^{2\alpha-2|\boldsymbol{r}|}. \tag{5.61}$$

Therefore, we conclude that uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$,

$$\begin{aligned}\mathrm{E}_0 \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}[\|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}} - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}\|_\infty^2 | \boldsymbol{Y}] &\lesssim \mathrm{E}_0 \sup_{\sigma^2\in\mathcal{U}_n} \mathrm{E}[\|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}} - D^{\boldsymbol{r}} f_{\boldsymbol{\theta}_0}\|_\infty^2 | \boldsymbol{Y}] + \mathrm{E}_0\|D^{\boldsymbol{r}} f_{\boldsymbol{\theta}_0} - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}\|_\infty^2 \\ &\lesssim n^{-1}\delta_n^{-2|\boldsymbol{r}|} + \delta_n^{2\alpha-2|\boldsymbol{r}|}. \end{aligned} \tag{5.62}$$

The empirical and hierarchical posterior convergence rates then follow from (5.57) and (5.58) respectively with absolute values replaced by sup-norms. □

We are now ready to prove Theorems 5.1 and 5.2.

Proof of Theorem 5.1. We will prove the empirical Bayes case, the corresponding proof for hierarchical Bayes is identical, i.e., by substituting $\Pi_{\widehat{\sigma}_n}(\cdot|\boldsymbol{Y})$ with $\Pi(\cdot|\boldsymbol{Y})$. Observe that for any $\boldsymbol{r}$ including $\boldsymbol{r} = \boldsymbol{0}$,

$$\begin{aligned}\sup_{\boldsymbol{x}\in C(\widetilde{\boldsymbol{\mu}},v\delta_n)} |D^{\boldsymbol{r}} f_{\boldsymbol{\vartheta}}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_0(\boldsymbol{x})| &= \sup_{\boldsymbol{x}\in C(\widetilde{\boldsymbol{\mu}},v\delta_n)} |D^{\boldsymbol{r}} f_{\boldsymbol{\theta}}(\boldsymbol{x} - \widetilde{\boldsymbol{\mu}}) - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}(\boldsymbol{x} - \widetilde{\boldsymbol{\mu}})| \\ &= \sup_{\boldsymbol{x}\in C(v\delta_n)} |D^{\boldsymbol{r}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) - D^{\boldsymbol{r}} f_{0,\boldsymbol{z}}(\boldsymbol{x})|. \end{aligned} \tag{5.63}$$

Define $\epsilon_n = n^{-1/2}\delta_n^{-1} + \delta_n^{\alpha-1}$. By (4.4) and Corollary 5.8 with $\boldsymbol{r} = \boldsymbol{e}_k$, for any $M_n \to \infty$,

$$\begin{aligned}
\mathrm{E}_0\Pi_{\widehat{\sigma}_n}(\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| > M_n\epsilon_n|\boldsymbol{Y}) &\leq \mathrm{E}_0\Pi_{\widehat{\sigma}_n}\left(\max_{1\leq k\leq d}\|D_k f_{\boldsymbol{\vartheta}} - D_k f_0\|_\infty > \frac{\lambda_0}{\sqrt{d}}M_n\epsilon_n\middle|\boldsymbol{Y}\right) \\
&\leq \sum_{k=1}^{d}\mathrm{E}_0\Pi_{\widehat{\sigma}_n}\left(\|D_k f_{\boldsymbol{\vartheta}} - D_k f_0\|_\infty > \frac{\lambda_0}{\sqrt{d}}M_n\epsilon_n\middle|\boldsymbol{Y}\right) \\
&= \sum_{k=1}^{d}\mathrm{E}_0\Pi_{\widehat{\sigma}_n}\left(\|D_k f_{\boldsymbol{\theta}} - D_k f_{0,\boldsymbol{z}}\|_\infty > \frac{\lambda_0}{\sqrt{d}}M_n\epsilon_n\middle|\boldsymbol{Y}\right) \to 0,
\end{aligned}$$

uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$. □

Proof of Theorem 5.2. By (4.5), Corollary 5.8 with $\boldsymbol{r} = \boldsymbol{0}$, and (5.63), for any $M_n \to \infty$,

$$\mathrm{E}_0\Pi_{\widehat{\sigma}_n}(|M - M_0| > M_n(n^{-1/2} + \delta_n^\alpha)|\boldsymbol{Y}) \leq \mathrm{E}_0\Pi_{\widehat{\sigma}_n}(\|f_{\boldsymbol{\theta}} - f_{0,\boldsymbol{z}}\|_\infty > M_n(n^{-1/2} + \delta_n^\alpha)|\boldsymbol{Y}) \to 0,$$

uniformly over $\|f_0\|_{\mathcal{W}^\alpha} \leq R$. The hierarchical Bayes case follows similarly by substituting $\Pi_{\widehat{\sigma}_n}(\cdot|\boldsymbol{Y})$ with $\Pi(\cdot|\boldsymbol{Y})$. □

5.3 Simulation

In this section, we will compare the performance of our proposed two-stage Bayesian procedure with three other estimation methods. Namely, the corresponding single-stage Bayesian, single-stage frequentist, and two-stage frequentist as proposed by Belitser et al. (2012). We consider a bivariate true function $f_0 : [0,1]^2 \to \mathbb{R}$ given by

$$f_0(x,y) = 5x(x-1)y(y-1)\sin(11x)\sin(11y). \tag{5.64}$$

A 3-D plot of f_0 is in Figure 5.1. In the first stage, we observed f_0 on a uniform 25×25 grid with i.i.d. errors distributed as $\mathrm{N}(0, 0.01)$ (see Figure 5.2 with gray points as observations). We used bivariate tensor B-splines with normal coefficients as our prior (see Section 3.2). Note that f_0 is a product of two identical functions, hence it has the same smoothness at each dimension. This implies that the isotropic assumption applies with the same set of knots and hence identical J for each B-spline component. At each dimension, the B-spline is of order 4 (cubic) with uniform knot sequence, and we added 4 duplicate knots at each of the boundary points $\{0,1\}$. For the prior parameters, we set $\boldsymbol{\eta} = \boldsymbol{0}$ and $\boldsymbol{\Omega} = \boldsymbol{I}_{J^2}$ for $J = 11$ (we set it the same as the J for single-stage Bayesian as will be discussed below). Figure 5.3 shows the surface of the posterior mean, where the red point is the point of maximum with corresponding location $\widehat{\boldsymbol{\mu}}_n$. We zoom-in around this vicinity by constructing a credible region $C(\widehat{\boldsymbol{\mu}}_n, \delta) = \{\boldsymbol{\mu} : \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n\|_\infty \leq \delta\}$.

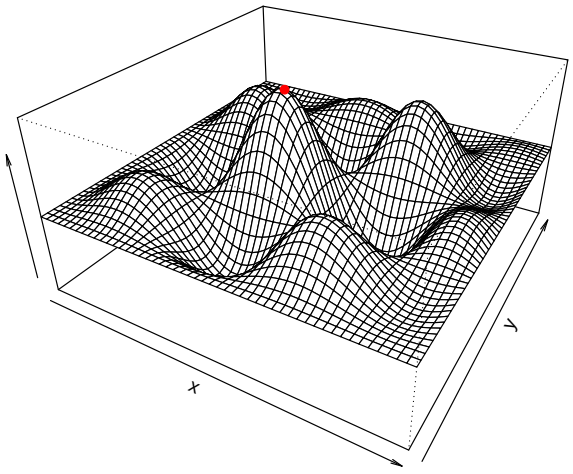


Figure 5.1: A plot of f_0, with the true maximum indicated by the red dot.

We take $\delta = 0.08$ (see Figure 5.5) and generate 70 observations at each of the 9 cube vertices $(\widehat{\mu}_1 + t_1\delta, \widehat{\mu}_2 + t_2\delta), t_1, t_2 = 0, \pm 1$ (see gray points in Figure 5.3). With this second stage samples, we use a bivariate (quadratic) polynomial with normal coefficients as prior (see (5.1) and (5.3)). That is for $x, y \in [0, 1]$,

$$f_{\boldsymbol{\vartheta}}(x, y) = \vartheta_0 + \vartheta_1 x + \vartheta_2 y + \vartheta_3 x^2 + \vartheta_4 y^2 + \vartheta_5 xy, \tag{5.65}$$

and we used Zellner's prior $\boldsymbol{\vartheta} \sim \mathrm{N}_5(\mathbf{0}, (\boldsymbol{P}^T\boldsymbol{P})^{-1})$. The empirical posterior for $\boldsymbol{\vartheta}$ is

$$\Pi_{\widehat{\sigma}_n}(\boldsymbol{\vartheta}|\boldsymbol{Y}) \sim \mathrm{N}_5((\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}^T\boldsymbol{Y}/2, \widehat{\sigma}_n^2(\boldsymbol{P}^T\boldsymbol{P})^{-1}/2), \tag{5.66}$$

where $\boldsymbol{Y}$ are the 630 second stage samples and $\widehat{\sigma}_n^2$ is computed using (5.8). Now, to compute $\boldsymbol{\mu} = \arg\max_{\boldsymbol{x} \in C(\widehat{\boldsymbol{\mu}}_n, \delta)} f_{\boldsymbol{\vartheta}}(\boldsymbol{x})$ for a fixed $\boldsymbol{\vartheta}$, we solve the following system of equation $\nabla f_{\boldsymbol{\vartheta}}(\boldsymbol{\mu}) = \mathbf{0}$, which is equivalent to solving

$$\begin{pmatrix} 2\vartheta_3 & \vartheta_5 \\ \vartheta_5 & 2\vartheta_4 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} -\vartheta_1 \\ -\vartheta_2 \end{pmatrix} \tag{5.67}$$

for $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$. Therefore, to induce the posterior distribution of $\boldsymbol{\mu}$, we draw samples from $\Pi_{\widehat{\sigma}_n}(\boldsymbol{\vartheta}|\boldsymbol{Y})$ and solve for $\boldsymbol{\mu}$ using (5.67) for each sample.

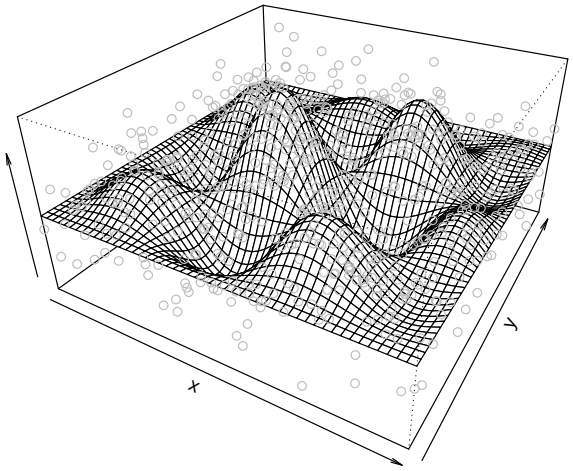


Figure 5.2: A plot of f_0, with gray points as the first stage 625 observations.

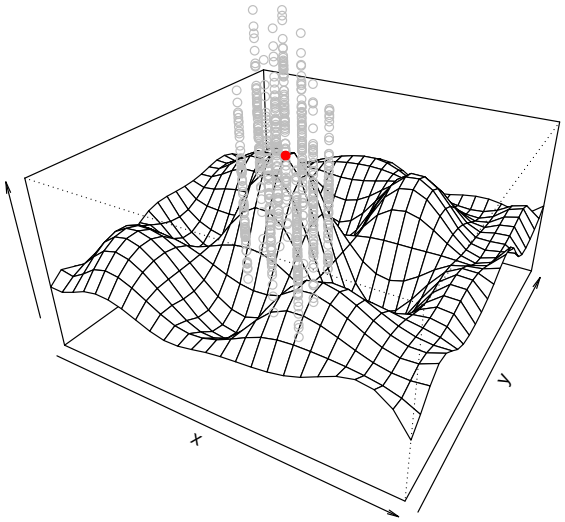


Figure 5.3: Posterior mean based on bivariate B-spline prior, the location corresponding to red dot is $\widehat{\boldsymbol{\mu}}_n$, and grey points are the 670 second stage samples.

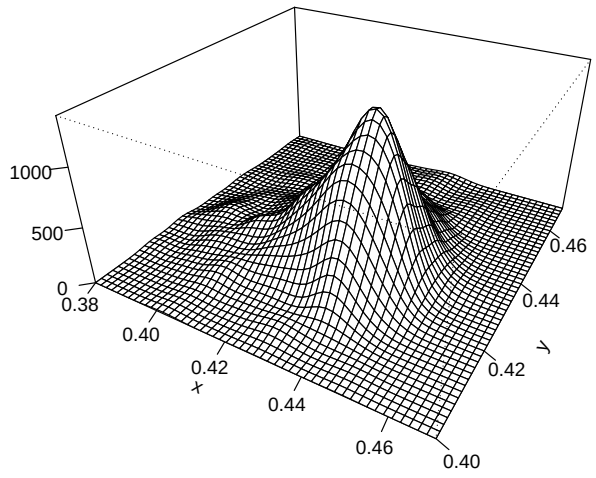
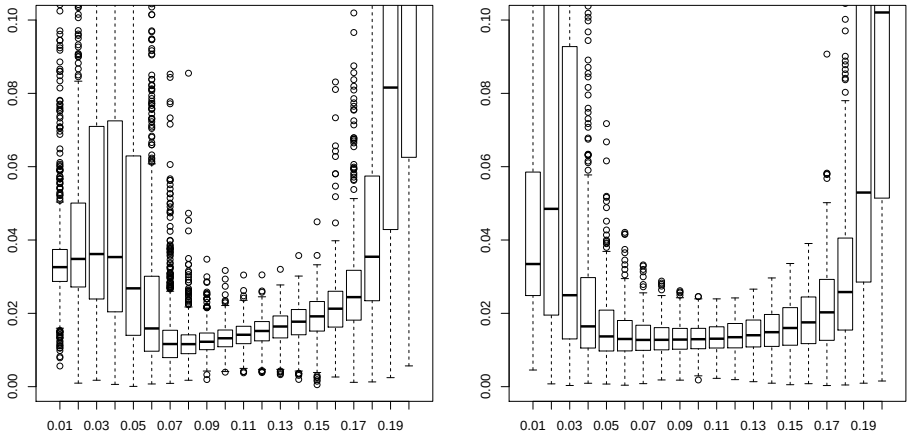


Figure 5.4: The second stage posterior density for $\boldsymbol{\mu}$.

Figure 5.4 shows the posterior density of $\boldsymbol{\mu}$ by sampling 1000 times from $\Pi_{\widehat{\sigma}_n}(\boldsymbol{\vartheta}|\boldsymbol{Y})$. The mean of this induced posterior $\widetilde{\boldsymbol{\mu}}_{2\text{nd}}$ will be an improved point estimator of $\boldsymbol{\mu}_0$ over the first stage estimator $\widehat{\boldsymbol{\mu}}_n$. The choice of $\delta = 0.08$ comes from repeating the above experiment 1000 times for each different δ. For each replicated experiment, we computed $\|\widetilde{\boldsymbol{\mu}}_{2\text{nd}} - \boldsymbol{\mu}_0\|$. The left panel of Figure 5.5 shows the corresponding box-plots of the 1000 computed $\|\widetilde{\boldsymbol{\mu}}_{2\text{nd}} - \boldsymbol{\mu}_0\|$ for a range of possible δ's. We see that the optimal δ is 0.08. The procedure described above is implemented in the statistical software package `R`. Univariate B-splines are constructed using the `bs` function from the `splines` package. We then use `tensor.prod.model.matrix` from the `crs` package to form their corresponding tensor-product.

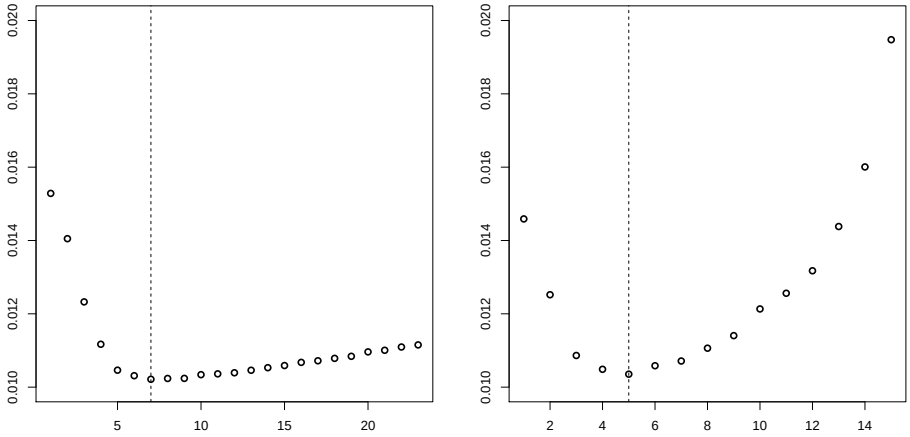
Observe that we have used a total of $25 \times 25 + 9 \times 70 = 1255$ observations. We then compare our two-stage procedure with a Bayesian single-stage method that uses roughly the same amount of samples. The closest is a uniform 36×36 grid of 1296 points. As in the previous case, we observed f_0 at these points with i.i.d errors $\mathrm{N}(0, 0.01)$. We then use bivariate tensor B-splines with normal coefficients as prior with the same setting as before. However, the number of interior knots N is determined by leave-one-out cross validation (CV), with prediction mean square error as our criterion function. From the left panel of Figure 5.6, we see that $N = 7$ or $J = N + q = 11$ at each dimension is optimal. To be consistent, we set the same J for both single and two-stage Bayesian procedures.



(a) Two-stage Bayes: $\delta = 0.08$ (b) Two-stage frequentist: $\delta = 0.1$

Figure 5.5: Box plots of 1000 computed $\|\widetilde{\boldsymbol{\mu}}_{2\text{nd}} - \boldsymbol{\mu}_0\|$, where $\widetilde{\boldsymbol{\mu}}_{2\text{nd}}$ is the second stage Bayesian (left) or frequentist (right) estimator of $\boldsymbol{\mu}_0$ for each $\delta = (0.01, \ldots, 0.2)$.

For comparison, we conducted the same experiment described using single and two-stage frequentist procedures. For the single-stage frequentist method, we used the least squares estimator $\widehat{f}(x, y) = \boldsymbol{b}_{J,q}(x, y)^T (\boldsymbol{B}^T \boldsymbol{B})^{-1} \boldsymbol{B}^T \boldsymbol{Y}$ to estimate $f_0(x, y)$. As in the Bayesian case, we used bivariate B-splines to construct $\boldsymbol{b}_{J,q}(x, y)$ and $\boldsymbol{B}$, such that $q = 4$ and $J = N + q = 9$ at each dimension. The optimal N is determined by leave-one-out cross validation (see right panel of Figure 5.6), and we used the same $J = 9$ for the single and two-stage frequentist procedures.



(a) Bayes single-stage: $N = 7$ (b) Frequentist single-stage: $N = 5$

Figure 5.6: CV mean square error against N number of interior knots at each dimension. Vertical dash lines indicate optimal N.

The frequentist two-stage procedure follows closely with its Bayesian counterpart, except that we use the location of maximum for $\widehat{f}$ as our preliminary estimator. Also, the corresponding optimal δ is 0.1 as shown in the right panel of Figure 5.5. The coefficients for the bivariate polynomial is estimated using least squares, i.e., $\widehat{\boldsymbol{\vartheta}} = (\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}^T\boldsymbol{Y}$, and the corresponding second stage estimator of $\boldsymbol{\mu}_0$ is computed using (5.67).

To compare the performance of the Bayesian and frequentist methods with single and two-stage procedures, we replicated the experiment 1000 times for each of the four methods by setting their corresponding optimal values for the tuning parameters J and δ. For each replicated experiment, we computed the L_2-norm of the difference between the Bayesian or frequentist estimator with $\boldsymbol{\mu}_0$. Note here that we sampled 1000 times from the posterior of $\boldsymbol{\mu}$ to compute the posterior mean as our Bayesian estimator for each replicated experiment and for each single or two-stage procedures. Figure 5.7 shows the box-plots of these 1000 computed L_2-norms for Bayesian and frequentist methods with single or two-stage procedures.

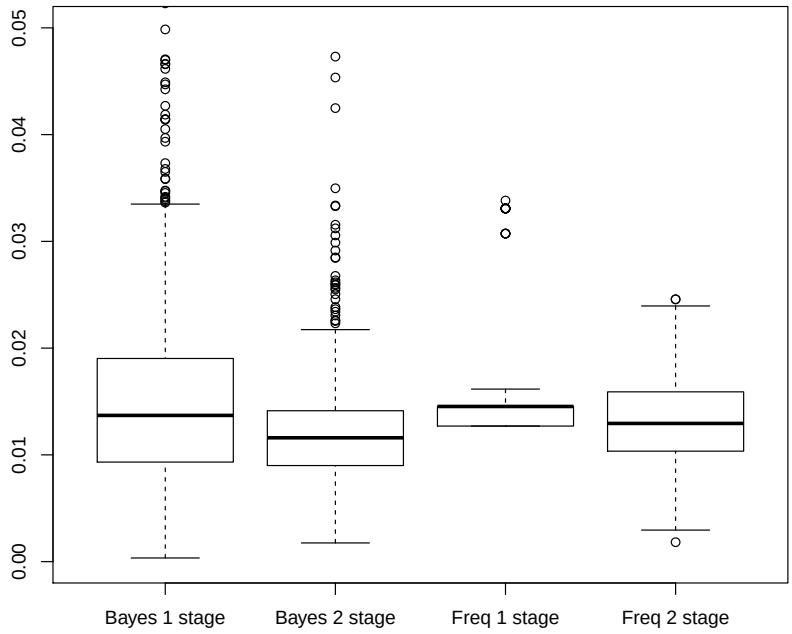


Figure 5.7: L_2-norms of the difference between Bayesian or frequentist estimator with $\boldsymbol{\mu}_0$ for 1000 Monte Carlo replications.

We see that for both Bayesian and frequentist methods, the two-stage procedures perform slightly better than single-stage estimators, in terms of lower mean square errors and smaller

spread, especially for the Bayesian case. This is true despite the fact that the second stage estimators have used less observations, namely 1255 as compared to 1296 used by single-stage estimators. We also note that our proposed Bayesian two-stage method seems to perform slightly better than its two-stage frequentist counterpart. However, our Bayesian two-stage procedure has relatively much wider spread when compared to the two-stage frequentist procedure. This is due to the fact that we are sampling from the empirical posterior of $\boldsymbol{\mu}$ to calculate its posterior mean, and it takes into account the error in estimating $\widehat{\sigma}_n^2$ and also the posterior spread. On the other hand, the frequentist estimators do not use σ^2 in their computation and this results in smaller variation.

REFERENCES

Barron, A., Schervish, M., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561.

Belitser, E. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *The Annals of Statistics*, 31(2):536–559.

Belitser, E., Ghosal, S., and van Zanten, J. H. (2012). Optimal two-stage procedures for estimating location and size of the maximum of a multivariate regression function. *The Annals of Statistics*, 40(6):2850–2876.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144.

Blum, J. R. (1954). Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, 25(4):737–744.

Bontemps, D. (2011). Bernstein-von Mises theorems for Gaussian regression with increasing number of regressors. *The Annals of Statistics*, 39(5):2557–2584.

Cai, T. and Low, M. (2006). Adaptive confidence balls. *The Annals of Statistics*, 34(1):202–228.

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

Castillo, I. and Nickl, R. (2013). Nonparametric Bernstein-von Mises theorems in Gaussian white noise. *The Annals of Statistics*, 41(4):1999–2028.

Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *The Annals of Statistics*, 21(2):903–923.

de Blasi, P., Lijoi, A., and Prünster, I. (2013). An asymptotic analysis of a class of discrete nonparametric priors. *Statistics Sinica*, 23:1299–1322.

de Boor, C. (2001). *A Practical Guide to Splines*. Springer-Verlag New York, Inc., revised edition.

de Jonge, R. and van Zanten, J. H. (2012). Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electronic Journal of Statistics*, 6:1984–2001.

Demko, S., Moss, W. F., and Smith, P. W. (1984). Decay rates for inverses of band matrices. *Mathematics of Computation*, 43(168):491–499.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1):1–26.

Diaconis, P. and Freedman, D. (1998). Consistency of Bayes estimates for nonparametric regression: normal theory. *Bernoulli*, 4(4):411–444.

Doss, H. (1984). Bayesian estimation in the symmetric location problem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 68(2):127–147.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent Advances in Statistics*, 24:287–302.

Freedman, D. (1963). On the asymptotic behavior of Bayes estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386–1403.

Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite dimensional parameters. *Annals of Statisitcs*, 27(4):1119–1140.

Ghosal, S., , Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158.

Ghosal, S., Lember, J., and van der Vaart, A. W. (2003). On Bayesian adaptation. *Acta Applicandae Mathematica*, 79(1-2):165–175.

Ghosal, S., Lember, J., and van der Vaart, A. W. (2008). Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89.

Ghosal, S. and van der Vaart, A. W. (2007). Convergence rate of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223.

Ghosal, S. and van der Vaart, A. W. (2014). Fundamentals of Nonparametric Bayesian Inference. To be published.

Giné, E. and Nickl, R. (2011). Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *The Annals of Statistics*, 39(6):2795–3443.

Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag New York, Inc.

Hoffman, M., Rousseau, J., and Schmidt-Hieber, J. (2013). On adaptive posterior concentration rates. Archive preprint arXiv:1305.5270v1 [math.ST].

Huang, T.-M. (2004). Convergence rates for posterior distributions and adaptive estimation. *The Annals of Statistics*, 32(4):1556–1593.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Juditsky, A. and Lambert-Lacroix, S. (2003). Nonparametric confidence set estimation. *Math. Methods Statist.*, 12(4):410–428.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, 18(1):191–219.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41(2):495–502.

Klemelä, J. (2005). Adaptive estimation of the mode of a multivariate density. *Journal of Nonparametric Statistics*, 17(1):83–105.

Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. (2011). Bayesian inverse problems with Gaussian priors. *The Annals of Statistics*, 39(5):2626–2657.

Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. (2013). Bayesian recovery of the initial condition for the heat equation. *Communications in Statistics - Theory and Methods*, 42(7):1294–1313.

Lan, Y., Banerjee, M., and Michailidis, G. (2009). Change-point estimation under adaptive sampling. *The Annals of Statistics*, 37(4):1752–1791.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer-Verlag New York, Inc.

Leahu, H. (2011). On the Bernstein-von Mises phenomenon in the Gaussian white noise model. *Electronic Journal of Statistics*, 5:373–404.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*. Springer-Verlag Berlin Heidelberg.

Lenk, P. J. (1988). The Logistic normal distribution for Bayeisan, nonparametric, predictive densities. *Journal of the American Statistical Association*, 83(402):509–516.

Leonard, T. (1978). Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(2):113–146.

Lo, A. Y. (1984). On a class of Bayesian nonparametrics estimates: I. density estimates. *The Annals of Statistics*, 12(1):351–357.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation*, 23(3):727–741.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238.

Meinardus, G. (1967). *Approximation of Functions, Theory and Numerical Methods*. Springer-Verlag New York, Inc. Translated by Larry L. Schumaker.

Mokkadem, A. and Pelletier, M. (2007). A companion for the KieferWolfowitzBlum stochastic approximation algorithm. *The Annals of Statistics*, 35(4):1749–1772.

Müller, H.-G. (1985). Kernel estimators of zeros and of location and size of extrema of regression functions. *Scandinavian Journal of Statistics*, 12(3):221–232.

Müller, H.-G. (1989). Adaptive nonparametric peak estimation. *The Annals of Statistics*, 17(3):1053–1069.

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Racine, J. S. and Nie, Z. (2014). *crs: Categorical Regression Splines*. R package version 0.15-23.

Ray, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electronic Journal of Statistics*, 7:2516–2549.

Rivoirard, V. and Rousseau, J. (2012). Bernstein-von Mises theorem for linear functionals of the density. *The Annals of Statistics*, 40(3):1489–1523.

Robins, J. and van der Vaart, A. W. (2006). Adaptive nonparametric confidence sets. *The Annals of Statistics*, 34(1):229–253.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press, New York, third edition.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26.

Scricciolo, C. (2014). Adaptive Bayesian density estimation in L^p-metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Analysis*, 9(2):475–520.

Searle, S. R. (1982). *Matrix Algebra useful for Statistics*. John Wiley and Sons, Inc.

Shen, W. and Ghosal, S. (2014). Adaptive Bayesian procedures using random series prior. Archive preprint arXiv:1403.0625v1 [math.ST].

Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical process with applications to statistics*. John Wiley and Sons, Inc.

Shoung, J.-M. and Zhang, C.-H. (2001). Least squares estimators of the mode of a unimodal regression function. *The Annals of Statistics*, 29(3):648–665.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053.

Szabo, B., van der Vaart, A. W., and van Zanten, J. H. (2014). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics*. To appear.

Tang, R., Banerjee, M., and Michailidis, G. (2011). A two-stage hybrid procedure for estimating an inverse regression function. *The Annals of Statistics*, 39(2):956–989.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

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

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

van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics*, 37(5B):2655–2675.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Process With Applications to Statistics*. Springer-Verlag New York, Inc.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer-Verlag New York, Inc., fourth edition.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(3):364–372.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation*, 36(1):45–54.

Zhao, L. H. (2000). Bayesian aspects of some nonparametric problems. *The Annals of Statistics*, 28(2):532–552.

Zhou, S., Shen, X., and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, 26(5):1760–1782.

APPENDICES

Appendix A

New Results on Univariate B-splines

We introduce new results on univariate B-splines that we developed during the course of writing this dissertation.

Lemma A.1. *Under assumption* (2.5)*, we have* $\Delta \asymp N^{-1} \sim J^{-1}$.

Proof. By (2.5), we have $\min_{1\le l\le N}\delta_l \le \Delta \le C\min_{1\le l\le N}\delta_l$. This implies that $\Delta \asymp \min_{1\le l\le N}\delta_l$. Then for all $1 \le l^{'} \le N$, we have $\min_{1\le l\le N}\delta_l \le \delta_{l'} \le \Delta$, and hence $N\min_{1\le l\le N}\delta_l \le \sum_{l'=1}^{N}\delta_{l'} \le N\Delta$. Using the fact that $\sum_{l'=1}^{N}\delta_{l'} \le 1$, we have $N^{-1} \lesssim \Delta$ and $\min_{1\le l\le N}\delta_l \lesssim N^{-1}$. Since $\Delta \lesssim \min_{1\le l\le N}\delta_l$, we obtain $\Delta \asymp N^{-1}$. By definition, $J = q+N$, where q does not depend on n and both J and N are assumed to increase with n, hence $J/N = q/N + 1$ and $N/J = 1 - q/J$ are $1 + o(1)$. □

Lemma A.2. *For* $1 \le j \le J$*, we have* $\sum_{i=1}^{n} B_{j,q}(X_i)^p = O(n/J)$ *for any fixed* $p \in \mathbb{N}$.

Proof. As $B_{j,q}(\cdot)^p$ is bounded by 1 for any $p \in \mathbb{N}$, and is positive only inside (t_{j-q}, t_j),

$$\sum_{i=1}^{n} B_{j,q}(X_i)^p \le |\{i : X_i \in (t_{j-q}, t_j)\}| = nG_n(t_{j-q}, t_j). \tag{A.1}$$

By assumption (2.5), we have $t_j - t_{j-q} \le q\Delta$ and $t_j - t_{j-q} \ge q\min_{1\le l\le N}\delta_l \ge q\Delta/C$. Moreover, assumption (2.6) and Lemma A.1 imply that

$$\begin{aligned} nG_n(t_{j-q}, t_j) &= nG(t_{j-q}, t_j) + o(n/N) \\ &\lesssim n\Delta + o(n\Delta) \\ &\lesssim n/J. \end{aligned} \tag{A.2}$$

Hence, from (A.1) above, we have $\sum_{i=1}^{n} B_{j,q}(X_i)^p = O(n/J)$. □

Lemma A.3. *Let* $\boldsymbol{b}_{J,q}(x) = (B_{1,q}(x), \ldots, B_{J,q}(x))^T$. *Then* $\|\boldsymbol{b}_{J,q}(t) - \boldsymbol{b}_{J,q}(s)\|^2 \lesssim J^2|t-s|^2$ *for any* $t, s \in (0,1)$.

Proof. By equation (8) of Chapter X in de Boor (2001) and the triangle inequality, we have

$$\begin{aligned}
|B'_{j,q}(x)| &= (q-1)\left(\frac{|B_{j,q-1}(x)|}{t_{j+q-1}-t_j} + \frac{|B_{j+1,q-1}(x)|}{t_{j+q}-t_{j+1}}\right) \\
&\leq (q-1)\frac{2}{\min_{1\leq l\leq N}\delta_l} \\
&\leq 2(q-1)\frac{C}{\Delta} \lesssim J,
\end{aligned}$$

where we have used the quasi-uniform knot assumption (2.5) and Lemma A.1 in the last line. Since at most $2q$ elements in $\boldsymbol{b}_{J,q}(t)$ and $\boldsymbol{b}_{J,q}(s)$ will be nonzero for any $t, s \in (0,1)$, the bound above together with the mean value theorem give

$$\begin{aligned}
\|\boldsymbol{b}_{J,q}(t) - \boldsymbol{b}_{J,q}(s)\|^2 &= \sum_{j=1}^{J} |B_{j,q}(t) - B_{j,q}(s)|^2 \\
&\leq 2q \max_{1\leq j\leq J} |B_{j,q}(t) - B_{j,q}(s)|^2 \\
&\lesssim J^2|t-s|^2. \qquad \square
\end{aligned}$$

Lemma A.4. *Let* $m \in \mathbb{N}_0$ *be such that* $m < \alpha$. *Let* $Z_{n,m} \sim \mathrm{GP}(0, \Sigma_m)$ *and* $Q_{n,m}$ *be a sub-Gaussian process with mean function* 0 *and covariance function* $\sigma_0^2\Psi_m$, *where* Σ_m *and* Ψ_m *were defined in* (2.66) *and* (2.71) *respectively. If* $m = 0$, *we write* $Z_{n,m} = Z_n \sim \mathrm{GP}(0, \Sigma)$ *and* $Q_{n,m} = Q_n$ *is a mean zero sub-Gaussian process with covariance function* $\sigma_0^2\Psi$, *where* Σ *and* Ψ *were defined in* (2.13) *and* (2.37) *respectively. Let* $J \asymp n^{1/(2\alpha+1)}$, *then for any* $t, s \in (0,1)$, *we have* $\mathrm{Var}[Z_{n,m}(t) - Z_{n,m}(s)] \leq Cn|t-s|^2$ *and* $\mathrm{Var}[Q_{n,m}(t) - Q_{n,m}(s)] \leq C'n|t-s|^2$ *for constants* $C, C' > 0$.

Proof. Let $J \asymp n^{1/(2\alpha+1)}$, then we have

$$\begin{aligned}
&\mathrm{Var}[Z_{n,m}(t) - Z_{n,m}(s)] \\
&= (\boldsymbol{b}_{J,q,m}(t) - \boldsymbol{b}_{J,q,m}(s))^T \boldsymbol{W}_m \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \boldsymbol{W}_m^T (\boldsymbol{b}_{J,q,m}(t) - \boldsymbol{b}_{J,q,m}(s)) \\
&\leq \|\boldsymbol{b}_{J,q,m}(t) - \boldsymbol{b}_{J,q,m}(s)\|^2 \left\|\left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1}\right\|_{(2,2)} \|\boldsymbol{W}_m^T\boldsymbol{W}_m\|_{(2,2)} \\
&\lesssim \frac{J^{2m+3}}{n}|t-s|^2 \lesssim n|t-s|^2,
\end{aligned}$$

where we have used Lemma A.3, equations (2.19) and (2.68) to bound the three norms in line

3 respectively. Similarly,

$$
\begin{aligned}
&\mathrm{Var}[Q_{n,m}(t) - Q_{n,m}(s)] \\
&= \sigma_0^2 (\boldsymbol{b}_{J,q,m}(t) - \boldsymbol{b}_{J,q,m}(s))^T \boldsymbol{W}_m \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \\
&\qquad \times \boldsymbol{B}^T\boldsymbol{B} \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \boldsymbol{W}_m^T (\boldsymbol{b}_{J,q,m}(t) - \boldsymbol{b}_{J,q,m}(s)) \\
&\lesssim \|\boldsymbol{b}_{J,q,m}(t) - \boldsymbol{b}_{J,q,m}(s)\|^2 \left\| \left(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{\Omega}^{-1}\right)^{-1} \right\|_{(2,2)}^2 \|\boldsymbol{B}^T\boldsymbol{B}\|_{(2,2)} \|\boldsymbol{W}_m^T\boldsymbol{W}_m\|_{(2,2)} \\
&\lesssim \frac{J^{2m+3}}{n} |t-s|^2 \lesssim n|t-s|^2,
\end{aligned}
$$

where we have used Lemma A.3, equations (2.19), (2.18) and (2.68) to bound the four norms in line 4 respectively. □

Appendix B

New Results on Tensor-product B-splines

We introduce new results on tensor-product B-splines that we developed during the course of writing this dissertation.

Lemma B.1. *Under assumption* (3.4)*, we have* $\Delta_k \asymp N_k^{-1} \sim J_k^{-1}$ *for* $k = 1, \ldots, d$.

Proof. By (3.4), $\min_{1\le l\le N_k} \delta_{k,l} \le \Delta_k \le C_k \min_{1\le l\le N_k} \delta_{k,l}$. Therefore, $\Delta_k \asymp \min_{1\le l\le N_k} \delta_{k,l}$. Then for all $1 \le l^{'} \le N_k$, we have $\min_{1\le l\le N_k} \delta_{k,l} \le \delta_{k,l'} \le \Delta_k$, and hence $N_k \min_{1\le l\le N_k} \delta_{k,l} \le \sum_{l'=1}^{N_k} \delta_{k,l'} \le N_k\Delta_k$. Using the fact that $\sum_{l'=1}^{N_k} \delta_{k,l'} \le 1$, we have $N_k^{-1} \lesssim \Delta_k$ and $\min_{1\le l\le N_k} \delta_{k,l} \lesssim N_k^{-1}$. Since $\Delta_k \lesssim \min_{1\le l\le N_k} \delta_{k,l}$, we obtain $\Delta_k \asymp N_k^{-1}$. By definition, $J_k = q_k + N_k$, where q_k does not depend on n and both N_k and J_k are assumed to increase with n, hence $J_k/N_k = 1 + q_k/N_k$ and $N_k/J_k = 1 - q_k/J_k$ are $1 + o(1)$ for $k = 1, \ldots, d$. □

Lemma B.2. *For* $1 \le j_k \le J_k$ *and* $p_k \in \mathbb{N}$ *where* $k = 1, \ldots, d$, $\sum_{i=1}^{n} \prod_{k=1}^{d} B_{j_k,q_k}(X_{ik})^{p_k} = O\left(n \prod_{k=1}^{d} J_k^{-1}\right)$.

Proof. For $k = 1, \ldots, d$, as $B_{j_k,q_k}(\cdot)^{p_k}$ is bounded by 1 and is positive only inside $(t_{k,j_k-q_k}, t_{k,j_k})$,

$$\begin{aligned}\sum_{i=1}^{n} \prod_{k=1}^{d} B_{j_k,q_k}(X_{ik})^{p_k} &\le n \int_{[0,1]^d} \prod_{k=1}^{d} \mathbb{1}_{(t_{k,j_k-q_k}, t_{k,j_k}]}(\boldsymbol{x}) dG_n(\boldsymbol{x}) \\ &= nG_n[\prod_{k=1}^{d} (t_{k,j_k-q_k}, t_{k,j_k})]. \end{aligned} \tag{B.1}$$

By assumption (3.4), we have $t_{k,j_k} - t_{k,j_k-q_k} \le q_k\Delta_k$ and $t_{k,j_k} - t_{k,j_k-q_k} \ge q_k \min_{1\le l\le N_k} \delta_{k,l} \ge q_k\Delta_k/C_k$. This implies that $t_{k,j_k} - t_{k,j_k-q_k} \asymp \Delta_k$ for $k = 1, \ldots, d$. Moreover, assumption (3.5)

and Lemma B.1 imply that

$$
\begin{aligned}
nG_n\left[\prod_{k=1}^{d}(t_{k,j_k-q_k},t_{k,j_k})\right] &= nG\left[\prod_{k=1}^{d}(t_{k,j_k-q_k},t_{k,j_k})\right]+o\left(n\prod_{k=1}^{d}N_k^{-1}\right)\\
&\lesssim n\prod_{k=1}^{d}\Delta_k+o\left(n\prod_{k=1}^{d}\Delta_k\right)\\
&\lesssim n\prod_{k=1}^{d}J_k^{-1}.
\end{aligned}
$$

Hence, from (B.1) above, we have $\sum_{i=1}^{n}\prod_{k=1}^{d}B_{j_k,q_k}(X_{ik})^{p_k}=O\left(n\prod_{k=1}^{d}J_k^{-1}\right)$. □

Lemma B.3. *Let* $\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})=\{\prod_{k=1}^{d}B_{j_k,q_k}(x_k)\}_{j_1=1,\ldots,j_d=1}^{J_1,\ldots,J_d}$ *for* $\boldsymbol{J}=(J_1,\ldots,J_d)^T$ *and* $\boldsymbol{q}=(q_1,\ldots,q_d)^T$. *Then* $\|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{t})-\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{s})\|^2\lesssim\sum_{k=1}^{d}J_k^2\|\boldsymbol{t}-\boldsymbol{s}\|^2$ *for any* $\boldsymbol{t},\boldsymbol{s}\in(0,1)^d$.

Proof. By equation (8) of Chapter X in de Boor (2001) and the triangle inequality, we have

$$
\begin{aligned}
|B'_{j_k,q_k}(x_k)| &= (q_k-1)\left(\frac{|B_{j_k,q_k-1}(x_k)|}{t_{j_k+q_k-1}-t_{j_k}}+\frac{|B_{j_k+1,q_k-1}(x_k)|}{t_{j_k+q_k}-t_{j_k+1}}\right)\\
&\le (q_k-1)\frac{2}{\min_{1\le l\le N_k}\delta_{k,l}}\\
&\le 2(q_k-1)\frac{C_k}{\Delta_k}\lesssim J_k, \qquad \text{(B.2)}
\end{aligned}
$$

where we have used the quasi-uniform knot assumption (3.4) and Lemma B.1 in the last line. Using the fact that $|\prod_{i=1}^{d}a_i-\prod_{i=1}^{d}b_i|\le\sum_{i=1}^{d}|a_i-b_i|$ for $|a_i|\le1$, $|b_i|\le1$, $i=1,\ldots,d$, and applying the mean value theorem in view of (B.2), we have

$$
\begin{aligned}
\left|\prod_{k=1}^{d}B_{j_k,q_k}(t_k)-\prod_{k=1}^{d}B_{j_k,q_k}(s_k)\right| &\le \sum_{k=1}^{d}|B_{j_k,q_k}(t_k)-B_{j_k,q_k}(s_k)|\\
&\lesssim \sum_{k=1}^{d}J_k|t_k-s_k|\le\left(\sum_{k=1}^{d}J_k^2\right)^{1/2}\|\boldsymbol{t}-\boldsymbol{s}\|,
\end{aligned}
$$

where we have used Cauchy-Schwarz inequality in the last line. Since at most $2\prod_{k=1}^{d}q_k$ elements

in both $\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{t})$ and $\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{s})$ will be nonzero for any $\boldsymbol{t},\boldsymbol{s}\in(0,1)^d$,

$$\begin{aligned}\|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{t})-\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{s})\|^2 &= \sum_{j_1=1}^{J_1}\cdots\sum_{j_d=1}^{J_d}\left|\prod_{k=1}^{d}B_{j_k,q_k}(t_k)-\prod_{k=1}^{d}B_{j_k,q_k}(s_k)\right|^2\\ &\lesssim \left(2\prod_{k=1}^{d}q_k\right)\sum_{k=1}^{d}J_k^2\|\boldsymbol{t}-\boldsymbol{s}\|^2\lesssim\sum_{k=1}^{d}J_k^2\|\boldsymbol{t}-\boldsymbol{s}\|^2.\end{aligned}$$

□

Lemma B.4. *Let $\boldsymbol{r}\in\mathbb{N}_0^d$ be such that $\sum_{k=1}^d r_k<\alpha$. Let $Z_{n,\boldsymbol{r}}\sim\mathrm{GP}(0,\Sigma_{\boldsymbol{r}})$ and $Q_{n,\boldsymbol{r}}$ be a sub-Gaussian process with mean function 0 and covariance function $\sigma_0^2\Psi_{\boldsymbol{r}}$, where $\Sigma_{\boldsymbol{r}}$ and $\Psi_{\boldsymbol{r}}$ were defined in* (3.61) *and* (3.66) *respectively. If $\boldsymbol{r}=\boldsymbol{0}$, we write $Z_{n,\boldsymbol{r}}=Z_n\sim\mathrm{GP}(0,\Sigma)$ and $Q_{n,\boldsymbol{r}}=Q_n$ is a zero mean sub-Gaussian process with covariance function $\sigma_0^2\Psi$, where Σ and Ψ were defined in* (3.12) *and* (3.34) *respectively. Let $J_k\asymp n^{1/(2\alpha+d)}$ for $k=1,\ldots,d$, then for any $\boldsymbol{t},\boldsymbol{s}\in(0,1)^d$, we have $\mathrm{Var}[Z_{n,\boldsymbol{r}}(\boldsymbol{t})-Z_{n,\boldsymbol{r}}(\boldsymbol{s})]\le Cn\|\boldsymbol{t}-\boldsymbol{s}\|^2$ and $\mathrm{Var}[Q_{n,\boldsymbol{r}}(\boldsymbol{t})-Q_{n,\boldsymbol{r}}(\boldsymbol{s})]\le C'n\|\boldsymbol{t}-\boldsymbol{s}\|^2$ for some constants $C,C'>0$.*

Proof. Let $J_k\asymp n^{1/(2\alpha+d)}$ for $k=1,\ldots,d$, we then have

$$\begin{aligned}&\mathrm{Var}[Z_{n,\boldsymbol{r}}(\boldsymbol{t})-Z_{n,\boldsymbol{r}}(\boldsymbol{s})]\\ &=(\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{t})-\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{s}))^T\boldsymbol{W}_{\boldsymbol{r}}\left(\boldsymbol{B}^T\boldsymbol{B}+\boldsymbol{\Omega}^{-1}\right)^{-1}\boldsymbol{W}_{\boldsymbol{r}}^T(\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{t})-\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{s}))\\ &\le\|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{t})-\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{s})\|^2\left\|\left(\boldsymbol{B}^T\boldsymbol{B}+\boldsymbol{\Omega}^{-1}\right)^{-1}\right\|_{(2,2)}\|\boldsymbol{W}_{\boldsymbol{r}}^T\boldsymbol{W}_{\boldsymbol{r}}\|_{(2,2)}\\ &\lesssim\frac{1}{n}\left(\prod_{k=1}^{d}J_k^{2r_k+1}\right)\left(\sum_{k=1}^{d}J_k^2\right)\|\boldsymbol{t}-\boldsymbol{s}\|^2\lesssim n\|\boldsymbol{t}-\boldsymbol{s}\|^2,\end{aligned}$$

where we have used Lemma B.3, equations (3.18) and (3.62) to bound the three norms in line 3 respectively. Similarly,

$$\begin{aligned}&\mathrm{Var}[Q_{n,\boldsymbol{r}}(\boldsymbol{t})-Q_{n,\boldsymbol{r}}(\boldsymbol{s})]\\ &=\sigma_0^2(\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{t})-\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{s}))^T\boldsymbol{W}_{\boldsymbol{r}}\left(\boldsymbol{B}^T\boldsymbol{B}+\boldsymbol{\Omega}^{-1}\right)^{-1}\\ &\quad\times\boldsymbol{B}^T\boldsymbol{B}\left(\boldsymbol{B}^T\boldsymbol{B}+\boldsymbol{\Omega}^{-1}\right)^{-1}\boldsymbol{W}_{\boldsymbol{r}}^T(\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{t})-\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{s}))\\ &\lesssim\|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{t})-\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\boldsymbol{s})\|^2\left\|\left(\boldsymbol{B}^T\boldsymbol{B}+\boldsymbol{\Omega}^{-1}\right)^{-1}\right\|_{(2,2)}^2\|\boldsymbol{B}^T\boldsymbol{B}\|_{(2,2)}\|\boldsymbol{W}_{\boldsymbol{r}}^T\boldsymbol{W}_{\boldsymbol{r}}\|_{(2,2)}\\ &\lesssim\frac{1}{n}\left(\prod_{k=1}^{d}J_k^{2r_k+1}\right)\left(\sum_{k=1}^{d}J_k^2\right)\|\boldsymbol{t}-\boldsymbol{s}\|^2\lesssim n\|\boldsymbol{t}-\boldsymbol{s}\|^2,\end{aligned}$$

where we have used Lemma B.3, equations (3.18), (3.17) and (3.62) to bound the four norms in line 4 respectively. □

Remark For the anisotropic case where $\sum_{k=1}^{d}(r_k/\alpha_k) < 1, \boldsymbol{r} \in \mathbb{N}_0^d$, we let $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \ldots, d$, and both $\mathrm{Var}[Z_{n,\boldsymbol{r}}(\boldsymbol{t}) - Z_{n,\boldsymbol{r}}(\boldsymbol{s})]$ and $\mathrm{Var}[Q_{n,\boldsymbol{r}}(\boldsymbol{t}) - Q_{n,\boldsymbol{r}}(\boldsymbol{s})]$ will be of the order $n\|\boldsymbol{t} - \boldsymbol{s}\|^2$.

Lemma B.5. *Let $f(\boldsymbol{x}) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})^T\boldsymbol{\theta}$ and define $I_{j_1,\ldots,j_d} = \prod_{k=1}^{d}[t_{k,j_k-q_k}, t_{k,j_k}]$. Furthermore, let $f|_{I_{j_1,\ldots,j_d}}$ be the restriction of f onto $I_{j_1,\ldots,j_d}$. Then there exists constant $C > 0$ depending on $\boldsymbol{q} = (q_1 \ldots, q_d)^T$ such that*

$$\|f|_{I_{j_1,\ldots,j_d}}\|_\infty \le C\prod_{k=1}^{d}(t_{k,j_k} - t_{k,j_k-q_k})^{-1/2}\|f|_{I_{j_1,\ldots,j_d}}\|_2.$$

Proof. By equation (12.8) of Theorem 12.2 from Schumaker (2007), we have that

$$\begin{aligned} f(\boldsymbol{x})|_{I_{j_1,\ldots,j_d}} &= \sum_{m_1=1}^{J_1}\cdots\sum_{m_d=1}^{J_d}\theta_{m_1,\ldots,m_d}\prod_{k=1}^{d}B_{m_k,q_k}(x_k)|_{I_{j_1,\ldots,j_d}} \\ &= \sum_{l_1=0}^{q_1-1}\cdots\sum_{l_d=0}^{q_d-1}\alpha_{l_1,\ldots,l_d}\prod_{k=1}^{d}x_k^{l_k} \quad \text{for } t_{k,j_k-q_k} \le x_k \le t_{k,j_k}, k = 1, \ldots, d. \end{aligned}$$

If $\boldsymbol{x} \in I_{j_1,\ldots,j_d}$, then $\boldsymbol{x} \in \prod_{k=1}^{d}[t_{k,j_k-h_k-1}, t_{k,j_k-h_k}]$ for some $h_k = 0, 1, \ldots, q_k - 1$. This implies that only terms associated with coefficients $\boldsymbol{\gamma} = \{\theta_{m_1,\ldots,m_d}\}_{m_1=j_1-h_1,\ldots,m_d=j_d-h_d}^{j_1-h_1+q_1-1,\ldots,j_d-h_d+q_d-1}$ will be nonzero. Also, we define $\boldsymbol{\alpha} = \{\alpha_{l_1,\ldots,l_d}\}_{l_1=0,\ldots,l_d=0}^{q_1-1,\ldots,q_d-1}$. From the two equivalent representation of f on $I_{j_i,\ldots,j_d}$ above, we observe that there is a one-to-one mapping between $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, i.e., each element of $\boldsymbol{\alpha}$ is a linear combination of elements in $\boldsymbol{\gamma}$ and vice-versa. Hence, there are matrices $\boldsymbol{T}$ and $\boldsymbol{V}$ of dimension $\prod_{k=1}^{d}q_k \times \prod_{k=1}^{d}q_k$ respectively, such that $\boldsymbol{T}\boldsymbol{\gamma} = \boldsymbol{\alpha}$ and $\boldsymbol{V}\boldsymbol{\alpha} = \boldsymbol{\gamma}$. Since these two linear transformations have entries and dimensions not depending on n, we have $\|\boldsymbol{T}\|_{(\infty,\infty)} = O(1)$ and $\|\boldsymbol{V}\|_{(\infty,\infty)} = O(1)$, with constants in $O(1)$ depending only on $\boldsymbol{q}$. Let $\boldsymbol{U}_{k,q_k} = (1, U_k, U_k^2, \ldots, U_k^{q_k-1})^T$ where $U_k \sim \mathrm{Uniform}(t_{k,j_k-q_k}, t_{k,j_k}), k = 1, \ldots, d$. Since $\mathrm{E}(\boldsymbol{U}_{k,q_k}\boldsymbol{U}_{k,q_k}^T)$ is nonsingular, its minimum eigenvalue is bounded below by a positive constant.

Therefore,

$$
\begin{aligned}
\|f|_{I_{j_1,\ldots,j_d}}\|_2^2 &= \int_{I_{j_1,\ldots,j_d}} \left(\sum_{l_1=0}^{q_1-1} \cdots \sum_{l_d=0}^{q_d-1} \alpha_{l_1,\ldots,l_d} \prod_{k=1}^{d} x_k^{l_k} \right)^2 d\boldsymbol{x} \\
&= \sum_{l_1=0}^{q_1-1} \cdots \sum_{l_d=0}^{q_d-1} \sum_{l_1'=0}^{q_1-1} \cdots \sum_{l_d'=0}^{q_d-1} \alpha_{l_1,\ldots,l_d} \alpha_{l_1',\ldots,l_d'} \prod_{k=1}^{d} \int_{[t_{k,j_k-q_k}, t_{k,j_k}]} x_k^{l_k+l_k'} dx_1 \ldots dx_d \\
&= \prod_{k=1}^{d} \left[(t_{k,j_k} - t_{k,j_k-q_k}) \boldsymbol{\alpha}^T \mathrm{E}(\boldsymbol{U}_{k,q_k} \boldsymbol{U}_{k,q_k}^T) \boldsymbol{\alpha} \right] \\
&\geq \prod_{k=1}^{d} (t_{k,j_k} - t_{k,j_k-q_k}) \lambda_{\min} \left(\mathrm{E}(\boldsymbol{U}_{k,q_k} \boldsymbol{U}_{k,q_k}^T) \right) \|\boldsymbol{\alpha}\|^2 \\
&\gtrsim \prod_{k=1}^{d} (t_{k,j_k} - t_{k,j_k-q_k}) \|\boldsymbol{\alpha}\|_\infty^2 \\
&\geq \prod_{k=1}^{d} (t_{k,j_k} - t_{k,j_k-q_k}) \|\boldsymbol{V}\|_{(\infty,\infty)}^{-2} \|\boldsymbol{\gamma}\|_\infty^2 \\
&\geq C \prod_{k=1}^{d} (t_{k,j_k} - t_{k,j_k-q_k}) \|f|_{I_{j_1,\ldots,j_d}}\|_\infty^2,
\end{aligned}
$$

where $\|f|_{I_{j_1,\ldots,j_d}}\|_\infty^2 \leq \|\boldsymbol{\gamma}\|_\infty^2 \| \sum_{m_1=1}^{J_1} \cdots \sum_{m_d=1}^{J_d} \prod_{k=1}^{d} B_{m_k,q_k}(\cdot)|_{I_{j_1,\ldots,j_d}}\|_\infty^2 \leq \|\boldsymbol{\gamma}\|_\infty^2$ for the last line. □

Lemma B.6. *There exists constant $c > 0$ depending only on $\boldsymbol{q} = (q_1 \ldots, q_d)^T$ such that for $f(\boldsymbol{x}) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})^T \boldsymbol{\theta}$, we have*

$$
c \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1,\ldots,j_d}^2 \prod_{k=1}^{d} (t_{k,j_k} - t_{k,j_k-q_k}) \leq \|f\|_2^2 \leq \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1,\ldots,j_d}^2 \prod_{k=1}^{d} (t_{k,j_k} - t_{k,j_k-q_k}).
$$

Proof. Since $\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})$ is a probability vector at a given $\boldsymbol{x}$, we use Jensen's inequality to write

$$
\begin{aligned}
\int_{[0,1]^d} f(\boldsymbol{x})^2 d\boldsymbol{x} &\leq \int_{[0,1]^d} \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1,\ldots,j_d}^2 \prod_{k=1}^{d} B_{j_k,q_k}(x_k) d\boldsymbol{x} \\
&\leq \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1,\ldots,j_d}^2 \int_{[0,1]^d} \prod_{k=1}^{d} \mathbb{1}_{(t_{k,j_k-q_k}, t_{k,j_k})}(x_k) d\boldsymbol{x} \\
&= \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1,\ldots,j_d}^2 \prod_{k=1}^{d} (t_{k,j_k} - t_{k,j_k-q_k}).
\end{aligned}
$$

Using Lemma B.5 and equation (5) of Chapter XI from de Boor (2001), we have

$$\begin{aligned}\|f\|_2^2 &= \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \|f|_{I_{j_1,\ldots,j_d}}\|_2^2 \\ &\gtrsim \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \prod_{k=1}^{d} (t_{k,j_k} - t_{k,j_k-q_k}) \|f|_{I_{j_1,\ldots,j_d}}\|_\infty^2 \\ &\geq c \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1,\ldots,j_d}^2 \prod_{k=1}^{d} (t_{k,j_k} - t_{k,j_k-q_k}). \end{aligned}$$

□

The following result is a multivariate generalization from Lemma 6.1 of Zhou et al. (1998).

Lemma B.7. *Let $f(\boldsymbol{x}) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x})^T\boldsymbol{\theta}$. Under assumptions* (3.4) *and* (3.5)*, we have that for any $\boldsymbol{\theta} \in \mathbb{R}^J$,*

$$n\|f\|_{2,G_n}^2 = \boldsymbol{\theta}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{\theta} \asymp n\left(\prod_{k=1}^{d} J_k^{-1}\right)\|\boldsymbol{\theta}\|^2.$$

Proof. To obtain upper bound for $\boldsymbol{\theta}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{\theta}$, we first use assumption (3.4), Lemma B.6 and write $g_{\max} = \sup_{\boldsymbol{x}\in[0,1]^d} g(\boldsymbol{x}) < \infty$ to get

$$\begin{aligned}\int_{[0,1]^d} f(\boldsymbol{x})^2 dG(\boldsymbol{x}) &\leq g_{\max}\|f\|_2^2 \\ &\leq g_{\max} \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1,\ldots,j_d}^2 \prod_{k=1}^{d} (t_{k,j_k} - t_{k,j_k-q_k}) \\ &\leq g_{\max}\|\boldsymbol{\theta}\|^2 \prod_{k=1}^{d} q_k\Delta_k. \end{aligned} \tag{B.3}$$

Noting that $(G_n - G)(\mathbf{1}_d) = (G_n - G)(\mathbf{0}) = 0$, we use multivariate integration by parts and

assumption (3.5) to write

$$\begin{aligned}
&\left|\int_{[0,1]^d} f(\boldsymbol{x})^2 d(G_n - G)(\boldsymbol{x})\right| \\
&= \left| f(\mathbf{1}_d)^2[(G_n - G)(\mathbf{1}_d)] - f(\mathbf{0})^2[(G_n - G)(\mathbf{0})] - 2\int_{[0,1]^d} (G_n - G)(\boldsymbol{x}) f(\boldsymbol{x}) \frac{\partial^d f(\boldsymbol{x})}{\partial x_1 \dots \partial x_d} d\boldsymbol{x} \right| \\
&\leq 2 \sup_{\boldsymbol{x}\in[0,1]^d} |G_n(\boldsymbol{x}) - G(\boldsymbol{x})| \int_{[0,1]^d} \left| f(\boldsymbol{x}) \frac{\partial^d f(\boldsymbol{x})}{\partial x_1 \dots \partial x_d} \right| d\boldsymbol{x} \\
&= o\left(\prod_{k=1}^d N_k^{-1}\right) \|f\|_2 \left\| \frac{\partial^d f}{\partial x_1 \dots \partial x_d} \right\|_2,
\end{aligned}$$

where we have used the Cauchy-Schwarz inequality in the last line. From (3.53), we have that $\mathfrak{D}^1_{j_k}\theta_{j_1,\dots,j_k} = (q_k - 1)\mathfrak{d}_{j_k}\theta_{j_1,\dots,j_d}(t_{k,j_k} - t_{k,j_k-q_k+1})^{-1}$, where $\mathfrak{d}_{j_k}\theta_{j_1,\dots,j_d} = \theta_{j_1,\dots,j_{k-1},j_k+1,j_{k+1},\dots,j_d} - \theta_{j_1,\dots,j_{k-1},j_k,j_{k+1},\dots,j_d}$. By setting $\boldsymbol{r} = \mathbf{1}_d$ in (3.54), we have

$$\begin{aligned}
\frac{\partial^d f(\boldsymbol{x})}{\partial x_1 \dots \partial x_d} &= \sum_{j_1=1}^{J_1-1} \dots \sum_{j_d=1}^{J_d-1} \prod_{k=1}^d \mathfrak{D}^1_{j_k}\theta_{j_1,\dots,j_d} B_{j_k,q_k-1}(x_k) \\
&= \sum_{j_1=1}^{J_1-1} \dots \sum_{j_d=1}^{J_d-1} \prod_{k=1}^d \mathfrak{d}_{j_k}\theta_{j_1,\dots,j_d} \prod_{k=1}^d \frac{q_k - 1}{t_{j_k} - t_{j_k-q_k+1}} B_{j_k,q_k-1}(x_k).
\end{aligned}$$

Hence, applying Lemma B.6 to f and its derivatives,

$$\begin{aligned}
&\left|\int_{[0,1]^d} f(\boldsymbol{x})^2 d(G_n - G)(\boldsymbol{x})\right| \\
&= o\left(\prod_{k=1}^d N_k^{-1}\right) \left(\sum_{j_1=1}^{J_1} \dots \sum_{j_d=1}^{J_d} \theta^2_{j_1,\dots,j_d} \prod_{i=1}^d (t_{k,j_k} - t_{k,j_k-q_k})\right)^{1/2} \\
&\quad \times \left[\sum_{j_1=1}^{J_1-1} \dots \sum_{j_d=1}^{J_d-1} \left(\prod_{k=1}^d \mathfrak{d}_{j_k}\theta_{j_1,\dots,j_d}\right)^2 \prod_{k=1}^d \frac{(q_k-1)^2}{t_{j_k} - t_{j_k-q_k+1}}\right]^{1/2} \\
&= o\left(\prod_{k=1}^d N_k^{-1}\right) \prod_{k=1}^d \left(\frac{\Delta_k}{\min_{1\leq l\leq N_k} \delta_{k,l}}\right) \|\boldsymbol{\theta}\| \left[\sum_{j_1=1}^{J_1-1} \dots \sum_{j_d=1}^{J_d-1} \left(\prod_{k=1}^d \mathfrak{d}_{j_k}\theta_{j_1,\dots,j_d}\right)^2\right]^{1/2} \\
&= o\left(\prod_{k=1}^d N_k^{-1}\right) \|\boldsymbol{\theta}\|^2, \qquad \text{(B.4)}
\end{aligned}$$

where the last line follows from assumption (3.4) and $\sum_{j_1=1}^{J_1-1} \dots \sum_{j_d=1}^{J_d-1} \left(\prod_{k=1}^d \mathfrak{d}_{j_k}\theta_{j_1,\dots,j_d}\right)^2 \leq$

$2^{2d}\|\boldsymbol{\theta}\|^2$. Combining this result with (B.3) and using Lemma B.1, we will obtain

$$\begin{aligned}\int_{[0,1]^d} f^2(\boldsymbol{x}) dG_n(\boldsymbol{x}) &= o\left(\prod_{k=1}^d N_k^{-1}\right)\|\boldsymbol{\theta}\|^2 + g_{\max}\|\boldsymbol{\theta}\|^2 \prod_{k=1}^d q_k\Delta_k \\ &\lesssim \left(\prod_{k=1}^d J_k^{-1}\right)\|\boldsymbol{\theta}\|^2.\end{aligned}$$

To obtain the lower bound, we use the same technique by applying assumption (3.4), Lemma B.6 and write $g_{\min} = \inf_{\boldsymbol{x}\in[0,1]^d} g(\boldsymbol{x}) > 0$ to get

$$\begin{aligned}\int_{[0,1]^d} f(\boldsymbol{x})^2 dG(\boldsymbol{x}) &\geq g_{\min}\|f\|_2^2 \\ &\geq c^2 g_{\min} \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1,\ldots,j_d}^2 \prod_{k=1}^d (t_{k,j_k} - t_{k,j_k-q_k}) \\ &\gtrsim g_{\min}\|\boldsymbol{\theta}\|^2 \prod_{k=1}^d C_k^{-1}\Delta_k.\end{aligned}$$

Therefore, combining the above with (B.4) and appealing to Lemma B.1, we get

$$\begin{aligned}\int_{[0,1]^d} f^2(\boldsymbol{x}) dG_n(\boldsymbol{x}) &= g_{\min}\|\boldsymbol{\theta}\|^2 \prod_{k=1}^d C_k^{-1}\Delta_k - o\left(\prod_{k=1}^d N_k^{-1}\right)\|\boldsymbol{\theta}\|^2 \\ &\gtrsim \left(\prod_{k=1}^d J_k^{-1}\right)\|\boldsymbol{\theta}\|^2.\end{aligned}$$

The proof is completed by observing that $\|f\|_{2,G_n}^2 = \int_{[0,1]^d} f^2(\boldsymbol{x}) dG_n(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{B}^T \boldsymbol{B}\boldsymbol{\theta}/n$. □

Lemma B.8. *Let $f_0 \in \mathcal{H}^{\boldsymbol{\alpha}}((0,1)^d)$, then for constant $C > 0$ depending on $\boldsymbol{q}$ and d, we have*

$$\|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\cdot)^T \boldsymbol{W}_{\boldsymbol{r}}\boldsymbol{\theta}_\infty - D^{\boldsymbol{r}} f_0\|_\infty \leq C\left(\sum_{k=1}^d J_k^{-(\alpha_k - r_k)} \left\|D^{(\alpha_k-r_k)\boldsymbol{e}_k} D^{\boldsymbol{r}} f_0\right\|_\infty\right),$$

where $\boldsymbol{\theta}_\infty$ is the same as the one in (3.19).

Proof. Let $I_{j_1,\ldots,j_d} = \prod_{k=1}^d [t_{k,j_k-q_k}, t_k]$ and define a bounded linear operator Q such that for any $f \in \mathcal{H}^{\boldsymbol{\alpha}}((0,1)^d)$,

$$Qf(\boldsymbol{x}) = \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} (\lambda_{j_1,\ldots,j_d} f) \prod_{k=1}^d B_{j_k,q_k}(x_k),$$

where $\lambda_{j_1,\ldots,j_d} = \prod_{k=1}^{d} \lambda_{k,j_k}$ and λ_{k,j_k} is the dual basis of $B_{j_k,q_k}(\cdot)$, i.e., λ_{k,j_k} is a linear functional such that $\lambda_{k,i_k} B_{j_k,q_k}(\cdot) = \mathbb{1}_{\{i_k=j_k\}}(\cdot)$ for $k = 1,\ldots,d$ (see Section 4.6 of Schumaker, 2007). Using Theorem 13.20 of Schumaker (2007), there exists a tensor-product Taylor's polynomial $p_{j_1,\ldots,j_d}(\boldsymbol{x}) = \sum_{j_1=0}^{\alpha_1-1} \cdots \sum_{j_d=0}^{\alpha_d-1} \zeta_{j_1,\ldots,j_d} \prod_{k=1}^{d} x_k^{j_k}$ such that

$$\|D^{\boldsymbol{r}}(f_0 - p_{j_1,\ldots,j_d})|_{I_{j_1,\ldots,j_d}}\|_\infty \leq C \sum_{k=1}^{d} J_k^{-(\alpha_k - r_k)} \left\| D^{(\alpha_k - r_k)\boldsymbol{e}_k} D^{\boldsymbol{r}} f_0 \Big|_{I_{j_1,\ldots,j_d}} \right\|_\infty,$$

with constant $C > 0$ depending on $\boldsymbol{q}$ and d. By equations (12.30) and (12.31) of Theorem 12.6 in Schumaker (2007), we have

$$\begin{aligned}
\|(D^{\boldsymbol{r}} f_0 - Q D^{\boldsymbol{r}} f_0)|_{I_{j_1,\ldots,j_d}}\|_\infty &\leq \|D^{\boldsymbol{r}}(f_0 - p_{j_1,\ldots,j_d})|_{I_{j_1,\ldots,j_d}}\|_\infty + \|Q(D^{\boldsymbol{r}} f_0 - D^{\boldsymbol{r}} p_{j_1,\ldots,j_d})|_{I_{j_1,\ldots,j_d}}\|_\infty \\
&\leq C \|D^{\boldsymbol{r}}(f_0 - p_{j_1,\ldots,j_d})|_{I_{j_1,\ldots,j_d}}\|_\infty \\
&\leq C \sum_{k=1}^{d} J_k^{-(\alpha_k - r_k)} \left\| D^{(\alpha_k - r_k)\boldsymbol{e}_k} D^{\boldsymbol{r}} f_0 \Big|_{I_{j_1,\ldots,j_d}} \right\|_\infty.
\end{aligned}$$

Since $Q D^{\boldsymbol{r}} f_0 = D^{\boldsymbol{r}} Q f_0$, identifying $(\boldsymbol{\theta}_\infty)_{j_1,\ldots,j_d}$ defined in (3.19) with $\lambda_{j_1,\ldots,j_d} f_0$ and applying equations (15) and (16) of Chapter X in de Boor, 2001, we see that $Q D^{\boldsymbol{r}} f_0$ is equal to $\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q},\boldsymbol{r}}(\cdot)^T \boldsymbol{W}_{\boldsymbol{r}} \boldsymbol{\theta}_\infty$. The proof is completed by summing both sides over $1 \leq j_k \leq J_k, k = 1,\ldots,d$. □

Appendix C

Sub-Gaussian Inequalities

Here we list important properties of sub-Gaussian random variables used in this dissertation.

1. If X is sub-Gaussian, then $X + a$ is also sub-Gaussian for any $a \in \mathbb{R}$.

2. For independent sub-Gaussian random variables $Y_1, \ldots, Y_n$, then $\sum_{i=1}^n a_i Y_i$ is also sub-Gaussian for any $a_i \in \mathbb{R}, i = 1, \ldots, n$. Moreover, for any $x > 0$,

$$P\left(\left|\sum_{i=1}^n a_i Y_i\right| > x\right) \leq 2\exp\left(-\frac{x^2}{2\|\boldsymbol{a}\|^2}\right),$$

   where $\boldsymbol{a} = (a_1, \ldots, a_n)^T$.

Lemma C.1 below bounds the L_p-moments of the sup-norm of a sub-Gaussian process with finite pointwise maximums of its mean and variance, by incurring a logarithmic factor.

Lemma C.1. *Let $\{X(\boldsymbol{t}) : \boldsymbol{t} \in [0,1]^d\}$ be a sub-Gaussian process with respect to the semi-metric $d(\boldsymbol{t}, \boldsymbol{s}) = \sqrt{\mathrm{Var}[X(\boldsymbol{t}) - X(\boldsymbol{s})]}$ such that $\mathrm{Var}[X(\boldsymbol{t}) - X(\boldsymbol{s})] \lesssim C(n)\|\boldsymbol{t} - \boldsymbol{s}\|^2$ for any $\boldsymbol{t}, \boldsymbol{s} \in [0,1]^d$, where $C(n)$ is a polynomial in n. Choose points $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{T_n})^T$ in $[0,1]^d$ such that $\bigcup_{i=1}^{T_n} B(\boldsymbol{u}_i, \delta_n) \supseteq [0,1]^d$, for any sequence $\delta_n \to 0$ as $n \to \infty$ with $\delta_n < 1$ and $T_n = N(\delta_n, [0,1]^d, \|\cdot\|)$. Then for all $p \geq 1$,*

$$\mathrm{E}\|X\|_\infty^p \lesssim \{\log(1/\delta_n)\}^{p/2}\left\{\left(\delta_n\sqrt{C(n)}\right)^p + \max_{1\leq i\leq T_n}|\mathrm{E}[X(\boldsymbol{u}_i)]|^p + \max_{1\leq i\leq T_n}\{\mathrm{Var}[X(\boldsymbol{u}_i)]\}^{p/2}\right\}. \tag{C.1}$$

Proof. Using the c_p-inequality, we have

$$\begin{aligned}
\mathrm{E}(\|X\|_\infty^p) &\le \mathrm{E}\left(\max_{1\le i\le T_n}\sup_{\boldsymbol{t}\in B(\boldsymbol{u}_i,\delta_n)}|X(\boldsymbol{t})|^p\right)\\
&\le 2^{p-1}\mathrm{E}\left(\max_{1\le i\le T_n}\sup_{\boldsymbol{t}\in B(\boldsymbol{u}_i,\delta_n)}|X(\boldsymbol{t})-X(\boldsymbol{u}_i)|^p+\max_{1\le i\le T_n}|X(\boldsymbol{u}_i)|^p\right)\\
&\lesssim \mathrm{E}\left(\sup_{\|\boldsymbol{t}-\boldsymbol{s}\|\le\delta_n}|X(\boldsymbol{t})-X(\boldsymbol{s})|^p\right)+\mathrm{E}\left(\max_{1\le i\le T_n}|X(\boldsymbol{u}_i)|^p\right). \qquad \text{(C.2)}
\end{aligned}$$

Since $\{X(\boldsymbol{t}):\boldsymbol{t}\in[0,1]^d\}$ is a sub-Gaussian process by assumption, we have from Section 2.2.1 of van der Vaart and Wellner (1996) that

$$\begin{aligned}
\|X(\boldsymbol{t})-X(\boldsymbol{s})\|_{\psi_2} &\le \sqrt{6}[\mathrm{Var}\{X(\boldsymbol{t})-X(\boldsymbol{s})\}]^{1/2}\\
&\lesssim \sqrt{C(n)}\|\boldsymbol{t}-\boldsymbol{s}\|,
\end{aligned}$$

where the last line follows by assumption. Applying Corollary 2.2.8 of van der Vaart and Wellner (1996) with $d(\boldsymbol{t},\boldsymbol{s})=\|\boldsymbol{t}-\boldsymbol{s}\|$,

$$\begin{aligned}
\left\|\sup_{\|\boldsymbol{t}-\boldsymbol{s}\|\le\delta_n}|X(\boldsymbol{t})-X(\boldsymbol{s})|\right\|_{\psi_2} &\lesssim \sqrt{C(n)}\int_0^{\delta_n}\sqrt{\log N(\varepsilon,[0,1]^d,\|\cdot\|)}d\varepsilon\\
&\lesssim \sqrt{C(n)}\int_0^{\delta_n}\sqrt{\log\left(1/\varepsilon^d\right)}d\varepsilon\\
&\lesssim \delta_n\sqrt{C(n)\log\left(1/\delta_n\right)}.
\end{aligned}$$

Since $\|Y\|_p\le p!\sqrt{\log 2}\|Y\|_{\psi_2}$ for any random variable Y with finite ψ_2-Orlicz norm, the first term in (C.2) can be bounded by

$$\mathrm{E}\left(\sup_{\|\boldsymbol{t}-\boldsymbol{s}\|\le\delta_n}|X(\boldsymbol{t})-X(\boldsymbol{s})|^p\right)\lesssim\left\|\sup_{\|\boldsymbol{t}-\boldsymbol{s}\|\le\delta_n}|X(\boldsymbol{t})-X(\boldsymbol{s})|\right\|_{\psi_2}^p\lesssim\delta_n^p\{C(n)\log\left(1/\delta_n\right)\}^{p/2}. \qquad \text{(C.3)}$$

In order to bound the second term in (C.2), first observe that for $Y\sim\mathrm{N}(\mu,\sigma^2)$ and $Z\sim\mathrm{N}(0,1)$,

$$\|Y\|_{\psi_2}=\|\mu+\sigma Z\|_{\psi_2}\le\|\mu\|_{\psi_2}+\sigma\|Z\|_{\psi_2}=|\mu|/\sqrt{\log 2}+\sigma\sqrt{8/3}. \qquad \text{(C.4)}$$

Therefore, applying Lemma 2.2.2 of van der Vaart and Wellner (1996) and using (C.4) above,

$$\begin{aligned}
\mathrm{E}\left[\max_{1\le i\le T_n}|X(\boldsymbol{u}_i)|^p\right] &\lesssim \left\|\max_{1\le i\le T_n}|X(\boldsymbol{u}_i)|\right\|_{\psi_2}^p \\
&\lesssim \{\log(1+T_n)\}^{p/2}\max_{1\le i\le T_n}\|X(\boldsymbol{u}_i)\|_{\psi_2}^p \\
&\lesssim \{\log(1+T_n)\}^{p/2}\left(\max_{1\le i\le T_n}|\mathrm{E}[X(\boldsymbol{u}_i)]|^p+\max_{1\le i\le T_n}\{\mathrm{Var}[X(\boldsymbol{u}_i)]\}^{p/2}\right).
\end{aligned} \tag{C.5}$$

Since $T_n \ge 2$ if $\delta_n < 1$, this implies that $\log(1+T_n) \le 2\log T_n \lesssim \log(\mathrm{diam}\{[0,1]^d\}/\delta_n^d) \le \log(1/\delta_n^d)$. Combining the bounds derived from (C.3) and (C.5) into (C.2), we will obtain

$$\mathrm{E}(\|X\|_\infty^p) \lesssim \{\log(1/\delta_n)\}^{p/2}\left\{\left(\delta_n\sqrt{C(n)}\right)^p+\max_{1\le i\le T_n}|\mathrm{E}[X(\boldsymbol{u}_i)]|^p+\max_{1\le i\le T_n}\{\mathrm{Var}[X(\boldsymbol{u}_i)]\}^{p/2}\right\}. \quad \square$$

Appendix D

Miscellaneous Results

Lemma D.1. *Let $\boldsymbol{A}$ be a $J \times J$ symmetric, positive definite and q-banded matrix. Assume that the eigenvalues of $\boldsymbol{A}$ are contained in the interval $[a\tau_m, b\tau_m]$ for fixed $0 < a < b < \infty$ and some sequence τ_m. Then $\|\boldsymbol{A}^{-1}\|_{(\infty,\infty)} = O(\tau_m^{-1})$.*

Proof. Since we can scale $\boldsymbol{A}$ by τ_m such that its eigenvalues are in $[a, b]$, we set $\tau_m = 1$ without loss of generality. Using Proposition 2.2 from Demko et al. (1984) and writing h_{ij} to be the (i, j)th element of $\boldsymbol{A}^{-1}$,

$$\|\boldsymbol{A}^{-1}\|_{(\infty,\infty)} = \max_{1\le i\le J} \sum_{j=1}^{J} |h_{i,j}| \le \max_{1\le i\le J} \sum_{j=1}^{J} C\lambda^{|i-j|},$$

where $C = \max\{1/a, (1 + \sqrt{b/a})^2/(2b)\}$ and $\lambda = [(\sqrt{b/a} - 1)/(\sqrt{b/a} + 1)]^{1/q}$. Since $b > a > 0$, we have that $0 < \lambda < 1$. Therefore,

$$\|\boldsymbol{A}^{-1}\|_{(\infty,\infty)} \le C \max_{1\le i\le J} \left(1 + 2\sum_{j=1}^{J} \lambda^j\right) \le C \left(1 + 2\sum_{j=1}^{\infty} \lambda^j\right) < \infty. \qquad \square$$

Lemma D.2. *Let $\boldsymbol{A}$ be a $J \times J$ symmetric and positive definite matrix with its rows and columns indexed by d-dimensional multi-indices, i.e., for $\boldsymbol{i} = (i_1, \ldots, i_d)$ and $\boldsymbol{j} = (j_1, \ldots, j_d)$, such that $1 \le i_k, j_k \le J_k, k = 1, \ldots, d, J = \prod_{k=1}^{d} J_k$, the $(\boldsymbol{i}, \boldsymbol{j})$th element of $\boldsymbol{A}$ is $a_{\boldsymbol{i},\boldsymbol{j}} = \boldsymbol{A}\{(i_1, \ldots, i_d), (j_1, \ldots, j_d)\}$. Given $\boldsymbol{q} = (q_1, \ldots, q_d)^T$, we assume $\boldsymbol{A}$ is $\boldsymbol{q}$-banded as in Definition 3.1. Furthermore, we assume that the eigenvalues of $\boldsymbol{A}$ are contained in $[a\tau_m, b\tau_m]$ for fixed $0 < a < b < \infty$ and some sequence τ_m. Then $\|\boldsymbol{A}^{-1}\|_{(\infty,\infty)} = O(\tau_m^{-1})$.*

Proof. We adapt the proof given in Proposition 2.2 of Demko et al. (1984) to the case of multi-dimensional banded matrices. Note that all matrices involved in this proof are indexed by d-dimensional multi-indices of total size J. We first establish by induction that if $\boldsymbol{A}$ is $\boldsymbol{q}$-banded

as in Definition 3.1, then $\boldsymbol{A}^n$ is $n\boldsymbol{q}$-banded for $n \in \mathbb{N}$. That is, $\boldsymbol{A}^n\{(i_1, \ldots, i_d), (j_1, \ldots, j_d)\} = 0$ if $|i_k - j_k| \geq nq_k$ for some $k = 1, \ldots, d$.

Suppose $\boldsymbol{A}$ is $\boldsymbol{q} = (q_1, \ldots, q_d)^T$-banded and $\boldsymbol{B}$ is $\boldsymbol{w} = (w_1, \ldots, w_d)^T$-banded, where both $\boldsymbol{A}$ and $\boldsymbol{B}$ are $J \times J$ symmetric and positive definite matrices, then we will have $(\boldsymbol{AB})_{\boldsymbol{i},\boldsymbol{j}} = \sum_{l_1=1}^{J_1} \cdots \sum_{l_d=1}^{J_d} a_{(i_1,\ldots,i_d),(l_1,\ldots,l_d)} b_{(j_1,\ldots,j_d),(l_1\ldots,l_d)} \neq 0$ only if at least one of the terms in the sum is nonzero. Therefore $a_{(i_1,\ldots,i_d),(l_1,\ldots,l_d)} \neq 0$ and $b_{(j_1,\ldots,j_d),(l_1,\ldots,l_d)} \neq 0$ for some $(l_1, \ldots, l_d)$. Hence $|i_k - l_k| \leq q_k$ and $|j_k - l_k| \leq w_k$ for $k = 1, \ldots, d$. This implies that $|i_k - j_k| \leq q_k + w_k$ for $k = 1, \ldots, d$ by the triangle inequality. We then conclude that $\boldsymbol{AB}$ is $\boldsymbol{q} + \boldsymbol{w}$ banded. Letting $\boldsymbol{B} = \boldsymbol{A}$, we see that $\boldsymbol{A}^2$ is $2\boldsymbol{q}$-banded. Suppose $\boldsymbol{A}^r$ is $r\boldsymbol{q}$-banded, then $\boldsymbol{A}^{r+1} = \boldsymbol{A}^r\boldsymbol{A}$ is $(r+1)\boldsymbol{q}$-banded by the reasoning above. Hence, we conclude that $\boldsymbol{A}^n$ is $n\boldsymbol{q}$-banded for all $n \in \mathbb{N}$ by induction.

Since we can scale $\boldsymbol{A}$ by τ_m such that its eigenvalues are in $[a, b]$, we set $\tau_m = 1$ without loss of generality. Let $p_n(\cdot)$ be a polynomial of degree n, then $p_n(\boldsymbol{A})$ is $n\boldsymbol{q}$-banded. Since $\Lambda(\boldsymbol{A}) \subseteq [a, b]$ by assumption, spectral theorem and Proposition 2.1 of Demko et al. (1984) implies that

$$\|\boldsymbol{A}^{-1} - p_n(\boldsymbol{A})\|_{(2,2)} = \max_{x \in \Lambda(\boldsymbol{A})} |1/x - p_n(x)| \leq C_0[(\sqrt{b/a} - 1)/(\sqrt{b/a} + 1)]^{n+1} \tag{D.1}$$

for $C_0 = (1 + \sqrt{b/a})^2/(2b)$. For any $n \in \mathbb{N}$, $p_n(\boldsymbol{A})_{\boldsymbol{i},\boldsymbol{j}} = 0$ if $|i_k - j_k| > nq_k$ for some $1 \leq k \leq d$. Suppose $\boldsymbol{i} \neq \boldsymbol{j}$, choose n to satisfy $n < \max_{1 \leq k \leq d} |i_k - j_k| q_k^{-1} \leq n+1$. Therefore by (D.1) above,

$$\begin{aligned} |\boldsymbol{A}^{-1}(\boldsymbol{i}, \boldsymbol{j})| &= |\boldsymbol{A}^{-1}(\boldsymbol{i}, \boldsymbol{j}) - p_n(\boldsymbol{A})_{\boldsymbol{i},\boldsymbol{j}}| \leq \|\boldsymbol{A}^{-1} - p_n(\boldsymbol{A})\|_{(2,2)} \\ &\leq C_0[(\sqrt{b/a} - 1)/(\sqrt{b/a} + 1)]^{\max_{1 \leq k \leq d} |i_k - j_k|/q_k} \\ &\leq C_0 \lambda^{\sum_{k=1}^d |i_k - j_k|}, \end{aligned} \tag{D.2}$$

for $\lambda = [(\sqrt{b/a}-1)/(\sqrt{b/a}+1)]^{1/\sum_{k=1}^d q_k}$. When $i_k = j_k$ for all $k = 1, \ldots, d$, we have $\boldsymbol{A}^{-1}(\boldsymbol{i}, \boldsymbol{j}) \leq \|\boldsymbol{A}^{-1}\|_{(2,2)} = 1/\lambda_{\min}(\boldsymbol{A}) \leq 1/a$. Combining this case with (D.2), we conclude that

$$|\boldsymbol{A}^{-1}(\boldsymbol{i}, \boldsymbol{j})| \leq C\lambda^{\sum_{k=1}^d |i_k - j_k|}, \tag{D.3}$$

for $C = \max\{C_0, 1/a\}$. Now using (D.3) above and observing that $0 < \lambda < 1$,

$$\begin{aligned}
\|\boldsymbol{A}^{-1}\|_{(\infty,\infty)} &= \max_{1\le i_k\le J_k, k=1,\ldots,d} \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} |\boldsymbol{A}^{-1}\{(i_1\ldots,i_d),(j_1,\ldots,j_d)\}| \\
&\le C \max_{1\le i_k\le J_k, k=1,\ldots,d} \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \prod_{k=1}^{d} \lambda^{|i_k-j_k|} \\
&\le C \max_{1\le i_k\le J_k, k=1,\ldots,d} \prod_{k=1}^{d} \left(1 + 2\sum_{j_k=1}^{J_k} \lambda^{j_k}\right) \\
&\le C \left(1 + 2\sum_{j=1}^{\infty} \lambda^{j}\right)^{d} < \infty. \qquad \square
\end{aligned}$$

Remark The crux of the lemma's proof lies in (D.1), where we used polynomials to approximate $1/x$ on a bounded interval, and relate this approximation error on the real line to that on matrices by the spectral theorem. Theorem 60 of Meinardus (1967) and the fact $(n+1)! > 2^n$ for $n \in \mathbb{N}$ imply that any analytic function on a compact interval can be approximated by polynomials with exponential error rate as $n \to \infty$. Hence, the above lemma applies to $g(\boldsymbol{A})$ where g is analytic over $[a\tau_m, b\tau_m]$, i.e., $\|g(\boldsymbol{A})\|_{(\infty,\infty)} = O(\tau_m^{-1})$. For instance, the result could also be applied on square root of $\boldsymbol{A}$.

Lemma D.3. *Let $\boldsymbol{A}$ be a $n \times n$ symmetric positive definite matrix. Assume that $\|\boldsymbol{A}\|_{(2,2)} \le 1$. Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ such that ε_i are iid mean 0, variance σ_0^2 and has finite fourth moment for $i = 1, \ldots, n$. Then $\mathrm{Var}(\boldsymbol{\varepsilon}^T \boldsymbol{A} \boldsymbol{\varepsilon}) = O(n)$.*

Proof. By eigen-decomposition, $\boldsymbol{A} = \boldsymbol{P}^T \boldsymbol{\Lambda} \boldsymbol{P}$ where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $\boldsymbol{P}$ is an orthogonal matrix, with p_{ij} its (i,j)th element. Define $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T = \boldsymbol{P}\boldsymbol{\varepsilon}$. Then,

$$\mathrm{Var}(\boldsymbol{\epsilon}^T \boldsymbol{A} \boldsymbol{\epsilon}) = \mathrm{Var}\left(\sum_{i=1}^{n} \lambda_i Z_i^2\right) = \sum_{i=1}^{n} \lambda_i^2 \mathrm{Var}(Z_i^2) + \sum_{r\ne s}^{n} \lambda_r \lambda_s \mathrm{Cov}(Z_r^2, Z_s^2). \tag{D.4}$$

For any $1 \le r, s \le n$,

$$\begin{aligned}
\mathrm{E}(Z_r^2 Z_s^2) &= \sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\sum_{j_1'=1}^{n}\sum_{j_2'=1}^{n} p_{rj_1} p_{rj_2} p_{sj_1'} p_{sj_2'} \mathrm{E}(\epsilon_{j_1}\epsilon_{j_2}\epsilon_{j_1'}\epsilon_{j_2'}) \\
&= \mathrm{E}(\epsilon_1^4) \sum_{j=1}^{n} p_{rj}^2 p_{sj}^2 + \sigma_0^4 \sum_{j_1\ne j_2}^{n} p_{rj_1}^2 p_{sj_2}^2 + 2\sigma_0^4 \sum_{j_1\ne j_2}^{n} p_{rj_1} p_{s_1} p_{rj_2} p_{sj_2}. \qquad \text{(D.5)}
\end{aligned}$$

Since $\lambda_i \le 1$ for $i = 1, \ldots, n$ and each row of $\boldsymbol{P}$ has unit norm, the first variance term in (D.4)

is bounded by

$$\sum_{i=1}^{n}\left(\mathrm{E}(\epsilon_1^4)\sum_{j=1}^{n}p_{ij}^4+3\sigma_0^4\sum_{j_1\neq j_2}^{n}p_{ij_1}^2p_{ij_2}^2\right)\lesssim\sum_{i=1}^{n}\left(\sum_{j=1}^{n}p_{ij}^2\right)^2\lesssim n. \tag{D.6}$$

Observe that by orthonormality of $\boldsymbol{P}$, $\mathrm{Var}(\boldsymbol{Z})=\mathrm{Var}(\boldsymbol{\varepsilon})=\sigma_0^2\boldsymbol{I}_n$ and $\mathrm{E}(Z_i^2)=\sigma_0^2$. Since $\sum_{j=1}^{n}p_{rj}p_{sj}=0$, then for $r\neq s$, $\sum_{j_1\neq j_2}^{n}p_{rj_1}p_{sj_1}p_{rj_2}p_{sj_2}=(\sum_{j=1}^{n}p_{rj}p_{sj})^2-\sum_{j=1}^{n}p_{rj}^2p_{sj}^2\leq 0$ and $\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}p_{rj_1}^2p_{rj_2}^2=\sum_{j=1}^{n}p_{rj}^4+\sum_{j_1\neq j_2}^{n}p_{rj_1}^2p_{rj_2}^2=1$. Hence for $r\neq s$ and in view of (D.5),

$$\mathrm{Cov}(Z_r^2,Z_s^2)\leq\mathrm{E}(\epsilon_1^4)\sum_{j=1}^{n}p_{sj}^2p_{rj}^2+\sigma_0^4\left(1-\sum_{j=1}^{n}p_{rj}^4\right)-\mathrm{E}(Z_r^2)\mathrm{E}(Z_s^2)\leq\mathrm{E}(\epsilon_1^4)\sum_{j=1}^{n}p_{sj}^2p_{rj}^2. \tag{D.7}$$

It then follows that the second covariance term in (D.4) is bounded by

$$\sum_{r\neq s}^{n}\mathrm{Cov}(Z_r^2,Z_s^2)\leq\mathrm{E}(\epsilon_1^4)\sum_{r=1}^{n}\sum_{j=1}^{n}\sum_{s=1}^{n}p_{sj}^2p_{rj}^2\lesssim n. \quad \square$$