Electronic Journal of Statistics
Vol. 13 (2019) 3485–3512
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1616

Rates of contraction with respect to L_2-distance for Bayesian nonparametric regression

Fangzheng Xie, Wei Jin and Yanxun Xu

Department of Applied Mathematics and Statistics, Johns Hopkins University
e-mail: yanxun.xu@jhu.edu

Abstract: We systematically study the rates of contraction with respect to the integrated L_2-distance for Bayesian nonparametric regression in a generic framework, and, notably, without assuming the regression function space to be uniformly bounded. The generic framework is very flexible and can be applied to a wide class of nonparametric prior models. Three non-trivial applications of the framework are provided: The finite random series regression of an α-Hölder function, with adaptive rates of contraction up to a logarithmic factor; The un-modified block prior regression of an α-Sobolev function, with adaptive-and-exact rates of contraction; The Gaussian spline regression of an α-Hölder function, with near optimal rates of contraction. These applications serve as generalization or complement of their respective results in the literature. Extensions to the fixed-design regression problem and sparse additive models in high dimensions are discussed as well.

Keywords and phrases: Bayesian nonparametric regression, block prior, finite random series, Gaussian splines, integrated L_2-distance, rate of contraction.

Received October 2018.

1. Introduction

Consider the standard nonparametric regression problem $y_i = f(\mathbf{x}_i) + e_i$, $i = 1, \cdots, n$, where the set of predictors $(\mathbf{x}_i)_{i=1}^n$ are referred to as design (points) and take values in $[0,1]^p \subset \mathbb{R}^p$, e_i's are independent and identically distributed (i.i.d.) mean-zero Gaussian noise with $\text{var}(e_i) = \sigma^2$, and y_i's are the responses. We follow the popular Bayesian approach by assigning f a prior distribution, and perform inference tasks by finding the posterior distribution of f given the observations $(\mathbf{x}_i, y_i)_{i=1}^n$.

In this paper we systematically study the rates of contraction with respect to the integrated L_2-distance

$$\|f - g\|_2 = \left\{ \int_{[0,1]^p} [f(\mathbf{x}) - g(\mathbf{x})]^2 \, \mathrm{d}\mathbf{x} \right\}^{1/2}$$

for Bayesian nonparametric regression in a generic framework that can be applied to a wide class of nonparametric prior models. In particular, we emphasize

that it allows the space of regression functions to be unbounded, including the renowned Gaussian process priors as special examples.

Rates of contraction of posterior distributions for Bayesian nonparametric priors have been studied extensively. Following the earliest framework on studying generic rates of contraction with i.i.d. data proposed by [13], specific examples for density estimation via Dirichlet process mixture models [3, 14, 17, 38] and location-scale mixture models [23, 47] are discussed. For nonparametric regression, the rates of contraction had not been discussed until [15], who develop a generic framework for fixed-design nonparametric regression to study rates of contraction with respect to the empirical L_2-distance. There are extensive studies for various priors that fall into this framework, including location-scale mixture priors [10], conditional Gaussian tensor-product splines [11], and Gaussian processes [43, 45], among which adaptive rates are obtained in [10, 11, 45].

Although it is interesting to achieve adaptive rates of contraction with respect to the empirical L_2-distance for nonparametric regression, this might be restrictive since the empirical L_2-distance quantifies the convergence of functions only at the given design points. In nonparametric regression, one also expects that the error between the estimated function and the true function can be globally small over the whole design space [48], *i.e.*, small mean-squared error for the out-of-sample prediction. Therefore the integrated L_2-distance is a natural choice. For Gaussian processes, [41, 50] provide contraction rates for nonparametric regression with respect to the integrated L_2 and L_∞-distance, respectively. A novel spike-and-slab wavelet series prior is constructed in [53] to achieve adaptive contraction with respect to the stronger L_∞-distance. These examples however, take advantage of their respective prior structures and may not be easily generalized. In particular, in [53] the authors tackle the more challenging posterior contraction problem with regard to the L_∞-distance, in which case the generic framework proposed in [15] is known to fail. A closely related reference is [26], where the authors discuss the rates of contraction of the rescaled-Gaussian process prior for the nonparametric random-design regression problem with respect to the integrated L_1-distance, which is weaker than the integrated L_2-distance. Although a generic framework for the integrated L_2-distance is presented in [21], the prior there is imposed on a uniformly bounded function space and hence rules out some popular priors, *e.g.*, the widely-adopted Gaussian process prior [29].

It is therefore natural to ask the following fundamental question: for Bayesian nonparametric regression, can one systematically study rates of contraction for various priors with respect to the integrated L_2-distance without assuming the uniform boundedness of the regression function space? In this paper we provide a positive answer to this question. The major contribution of this work is that we prove the existence of an ad-hoc test function that is required in the generic rates of contraction framework in [13] by leveraging Bernstein's inequality and imposing certain structural assumption on the sieves with large prior probabilities. This is made clear in Section 2. Furthermore, we do not require the prior to be supported on a uniformly bounded function space. Consequently, we are able to establish a general rate of contraction theorem with respect to the integrated

L_2-distance for Bayesian nonparametric regression. Examples of applications falling into this framework include the finite random series prior [30, 36], the (un-modified) block prior [12], and the Gaussian splines prior [11]. In particular, for the block prior regression, rather than modifying the block prior by conditioning on a truncated function space as in [12] with a known upper bound for the unknown true regression function, we prove that the un-modified block prior automatically yields rate-exact Bayesian adaptation for nonparametric regression without such a truncation. We further extend the proposed framework to the fixed-design regression and sparse additive models in high dimensions. The analyses of the above applications and extensions also generalize their respective results in the literature. These improvements and generalizations are made clear in Sections 3 and 4.

The layout of this paper is as follows. In Section 2 we introduce the main generic result for studying rates of contraction for Bayesian nonparametric regression. As applications of the main result, we derive the rates of contraction of various popular priors for nonparametric regression in the literature with substantial improvements in Section 3. Section 4 elaborates on extensions of the proposed framework to the fixed-design regression problem and sparse additive models in high dimensions. The technical proofs of the main result are deferred to Section 5 and to supplementary material [49].

Notations

For $1 \leq r \leq \infty$, we use $\|\cdot\|_r$ to denote both the ℓ_r-norm on any finite dimensional Euclidean space and the integrated L_r-norm of a measurable function (with respect to the Lebesgue measure). In particular, for any function $f \in L_2([0,1]^p)$, we use $\|f\|_2$ to denote the integrated L_2-norm defined to be $\|f\|_2^2 = \int_{[0,1]^p} f^2(\mathbf{x})\mathrm{d}\mathbf{x}$. We follow the convention that when $r = 2$, the subscript is omitted, *i.e.*, $\|\cdot\| = \|\cdot\|_2$. The Hilbert space l^2 denotes the space of sequences that are squared-summable. Given $x \in \mathbb{R}$, we use $\lfloor x \rfloor$ to denote the maximal integer no greater than x, and $\lceil x \rceil$ to denote the minimum integer no less than x. The notations $a \lesssim b$ and $a \gtrsim b$ denote the inequalities up to a positive multiplicative constant, and we write $a \asymp b$ if $a \lesssim b$ and $a \gtrsim b$. Throughout capital letters $C, C_1, \tilde{C}, C', D, D_1, \tilde{D}, D', \cdots$ are used to denote generic positive constants and their values might change from line to line unless particularly specified, but are universal and unimportant for the analysis.

We refer to $\mathcal{P}$ as a sampling model if it consists of a class of densities on a sample space $\mathcal{X}$ with respect to some underlying σ-finite measure. Given a sampling model $\mathcal{P}$ and the i.i.d. data $(\mathbf{w}_i)_{i=1}^n$ from some $P \in \mathcal{P}$, the prior and the posterior distribution on $\mathcal{P}$ are always denoted by $\Pi(\cdot)$ and $\Pi(\cdot \mid \mathbf{w}_1, \cdots, \mathbf{w}_n)$, respectively. Given a function $f : \mathcal{X} \to \mathbb{R}$, we denote $\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(\mathbf{x}_i)$ and $\mathbb{G}_n f = n^{-1/2} \sum_{i=1}^n [f(\mathbf{x}_i) - \mathbb{E}f(\mathbf{x}_i)]$, given the i.i.d. data $(\mathbf{x}_i)_{i=1}^n$. With a slight abuse of notations, when applying to a set of design points $(\mathbf{x}_i)_{i=1}^n$, we also denote $\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(\mathbf{x}_i)$ and $\mathbb{G}_n f = n^{-1/2} \sum_{i=1}^n [f(\mathbf{x}_i) - \mathbb{E}f(\mathbf{x}_i)]$, regardless of whether these design points are random or fixed. We use ϕ to denote the

probability density function of the (univariate) standard normal distribution, and we use the shorthand notation $\phi_\sigma(y) = \phi(y/\sigma)/\sigma$ to denote the density of $\mathrm{N}(0, \sigma^2)$. For a metric space $(\mathcal{F}, d)$, for any $\epsilon > 0$, the ϵ-covering number of $(\mathcal{F}, d)$, denoted by $\mathcal{N}(\epsilon, \mathcal{F}, d)$, is defined to be the minimum number of ϵ-balls of the form $\mathcal{B}(f, \epsilon) := \{g \in \mathcal{F} : d(f, g) < \epsilon\}$ that are needed to cover $\mathcal{F}$.

2. The framework and main results

Consider the nonparametric regression model: $y_i = f(\mathbf{x}_i) + e_i$, where $(e_i)_{i=1}^n$ are i.i.d. mean-zero Gaussian noise with $\mathrm{var}(e_i) = \sigma^2$, and $(\mathbf{x}_i)_{i=1}^n$ are design points taking values in $[0, 1]^p$. Unless otherwise stated, the design points $(\mathbf{x}_i)_{i=1}^n$ are assumed to be independently and uniformly sampled for simplicity throughout the paper. Our framework naturally adapts to the case where the design points are independently sampled from a density function that is bounded away from 0 and ∞. We assume that the responses y_i's are generated from $y_i = f_0(\mathbf{x}_i) + e_i$ for some unknown $f_0 \in L_2([0, 1]^p)$. Thus the data $\mathcal{D}_n = (\mathbf{x}_i, y_i)_{i=1}^n$ can be regarded as i.i.d. samples from a distribution $\mathbb{P}_0$ with joint density $p_0(\mathbf{x}, y) = \phi_\sigma(y - f_0(\mathbf{x}))$. Throughout we assume that the variance σ^2 of the noise is known, but our framework can be easily extended to the case where σ is unknown by placing a prior on σ that is supported on a compact interval contained in $(0, \infty)$ with a density bounded away from 0 and ∞ (see, for example, Section 2.2.1 in [10] and Theorem 3.3 in [43]).

Before presenting the main result, let us first introduce the basic framework for studying convergence of Bayesian nonparametric regression. In the context of the aforementioned nonparametric regression, by assigning a prior Π on the regression function f, one obtains the posterior distribution $\Pi(f \in \cdot \mid \mathcal{D}_n)$ defined through

$$\Pi(f \in A \mid \mathcal{D}_n) = \frac{\int_A \prod_{i=1}^n [p_f(\mathbf{x}_i, y_i)/p_0(\mathbf{x}_i, y_i)]\Pi(\mathrm{d}f)}{\int \prod_{i=1}^n [p_f(\mathbf{x}_i, y_i)/p_0(\mathbf{x}_i, y_i)]\Pi(\mathrm{d}f)}$$

for any measurable function class A, where $p_f(\mathbf{x}, y) = \phi_\sigma(y - f(\mathbf{x}))$. In order that the posterior distribution $\Pi(\cdot \mid \mathcal{D}_n)$ contracts to f_0 at rate ϵ_n with respect to a distance d, *i.e.*, $\Pi(d(f, f_0) > M\epsilon_n \mid \mathcal{D}_n) \to 0$ in $\mathbb{P}_0$-probability for some large constant $M > 0$, the authors of [13] proposed the following sufficient conditions, referred to as the prior-concentration-and-testing framework: There exist some constants $D, D' > 0$, such that for sufficiently large n:

1. The prior concentration condition holds:

$$\Pi\left(\mathbb{E}_0\left(\log\frac{p_0}{p_f}\right) \le \epsilon_n^2, \mathbb{E}_0\left[\left(\log\frac{p_0}{p_f}\right)^2\right] \le \epsilon_n^2\right) \ge \mathrm{e}^{-Dn\epsilon_n^2}. \tag{2.1}$$

2. There exists a sequence $(\mathcal{F}_n)_{n=1}^\infty$ of subsets of $L_2([0, 1]^p)$ (often referred to as the sieves) and test functions $(\phi_n)_{n=1}^\infty$ such that $\Pi(\mathcal{F}_n^c) \le \mathrm{e}^{-(D+4)n\epsilon_n^2}$,

$$\mathbb{E}_0\phi_n \to 0, \text{ and } \sup_{f \in \mathcal{F}_n \cap \{d(f, f_0) > M\epsilon_n\}} \mathbb{E}_f(1 - \phi_n) \le \mathrm{e}^{-D'Mn\epsilon_n^2}.$$

However, the above framework is too abstract, and is not instructive for constructing the appropriate sieves $(\mathcal{F}_n)_{n=1}^{\infty}$ nor the desired test functions $(\phi_n)_{n=1}^{\infty}$ for studying the rates of contraction for nonparametric regression with respect to $\|\cdot\|_2$. Specifically, it does not provide a guidance on how to construct the desired sieves, or what their structural features are. The major contribution of this work, in contrast, is that we impose certain structural assumption on the sieves to construct the desired test functions. By doing so, we are able to make the framework more concrete and instructive for a variety of nonparametric regression priors to derive the corresponding posterior contraction rates.

The following local testing lemma is the first technical contribution of this work. It also serves as a building block to construct the desired test functions required in the prior-concentration-and-testing framework.

Lemma 2.1. *Let $\eta > 0$ be a constant. For any $m \in \mathbb{N}_+$ and $\delta > 0$, assume the class of functions $\mathcal{F}_m(\delta)$ satisfies*

$$f \in \mathcal{F}_m(\delta) \Longrightarrow \|f - f_0\|_\infty^2 \le \eta(m\|f - f_0\|_2^2 + \delta^2). \tag{2.2}$$

Then for any $f_1 \in \mathcal{F}_m(\delta)$ with $\sqrt{n}\|f_1 - f_0\|_2 > 1$, there exists a test function $\phi_n : (\mathcal{X} \times \mathcal{Y})^n \to [0,1]$ such that

$$\begin{aligned} \mathbb{E}_0\phi_n &\le \exp\left(-Cn\|f_1 - f_0\|_2^2\right), \\ \sup_{\{f \in \mathcal{F}_m(\delta) : \|f - f_1\|_2 \le \xi\|f_0 - f_1\|_2\}} \mathbb{E}_f(1 - \phi_n) &\le \exp\left(-Cn\|f_1 - f_0\|_2^2\right) \\ &\quad + 2\exp\left(-\frac{Cn\|f_1 - f_0\|_2^2}{m\|f_1 - f_0\|_2^2 + \delta^2}\right) \end{aligned}$$

for some constant $C > 0$ and $\xi \in (0,1)$.

The key ingredient of Lemma 2.1 is the condition (2.2) on the sieve $\mathcal{F}_m(\delta)$. By requiring that functions in $\mathcal{F}_m(\delta)$ cannot explode in $\|f - f_0\|_\infty$ when $\|f - f_0\|_2$ is small, we can apply Bernstein's inequality to the likelihood ratio test statistic and obtain exponentially small type I and type II error probability bounds. Based on Lemma 2.1, we are able to establish the following global testing lemma.

Lemma 2.2. *Suppose that $\mathcal{F}_m(\delta)$ satisfies (2.2) for $m \in \mathbb{N}_+$ and $\delta > 0$. Let $(\epsilon_n)_{n=1}^{\infty}$ be a sequence with $n\epsilon_n^2 \to \infty$. Then for any $M \ge 0$, there exists a sequence of test functions $(\phi_n)_{n=1}^{\infty}$ such that*

$$\begin{aligned} \mathbb{E}_0\phi_n &\le \sum_{j=M}^{\infty} N_{nj} \exp\left(-Cnj^2\epsilon_n^2\right), \\ \sup_{\{f \in \mathcal{F}_m(\delta) : \|f - f_0\|_2 > M\epsilon_n\}} \mathbb{E}_f(1 - \phi_n) &\le \exp(-CM^2n\epsilon_n^2) + 2\exp\left(-\frac{CM^2n\epsilon_n^2}{mM^2\epsilon_n^2 + \delta^2}\right), \end{aligned}$$

where $N_{nj} = \mathcal{N}(\xi j\epsilon_n, \mathcal{S}_{nj}(\epsilon_n), \|\cdot\|_2)$ is the covering number of

$$\mathcal{S}_{nj}(\epsilon_n) = \{f \in \mathcal{F}_m(\delta) : j\epsilon_n < \|f - f_0\|_2 \le (j+1)\epsilon_n\},$$

and C is some positive constant.

The prior concentration condition (2.1) is very important in the study of Bayes theory. It guarantees that the denominator appearing in the posterior distribution $\int[p_f(\mathbf{x}_i, y_i)/p_0(\mathbf{x}_i, y_i)]\Pi(\mathrm{d}f)$ can be lower bounded by $\mathrm{e}^{-D'n\epsilon_n^2}$ for some constant $D' > 0$ with large probability (see, for example, Lemma 8.1 in [13]). In the context of normal regression, the Kullback-Leibler divergence is proportional to the integrated L_2-distance between two regression functions. Motivated by this observation, we establish the following lemma that yields an exponential lower bound for the denominator $\int[p_f(\mathbf{x}_i, y_i)/p_0(\mathbf{x}_i, y_i)]\Pi(\mathrm{d}f)$ in the posterior distribution under the current framework.

Lemma 2.3. *Let $\eta' > 0$ be a constant. Denote*

$$B(m, \epsilon, \omega) = \left\{f : \|f - f_0\|_2 < \epsilon, \|f - f_0\|_\infty^2 \le \eta'(m\|f - f_0\|_2^2 + \omega^2)\right\}$$

for any $\epsilon, \delta > 0$ and $m \in \mathbb{N}_+$. Suppose sequences $(\epsilon_n)_{n=1}^\infty$ and $(k_n)_{n=1}^\infty$ satisfy $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$, $k_n\epsilon_n^2 = O(1)$, and ω is some constant. Then for any constant $C > 0$,

$$\mathbb{P}_0\left(\int\prod_{i=1}^n \frac{p_f(\mathbf{x}_i, y_i)}{p_0(\mathbf{x}_i, y_i)}\Pi(\mathrm{d}f) \le \Pi\left(B(k_n, \epsilon_n, \omega)\right)\exp\left[-\left(C + \frac{1}{\sigma^2}\right)n\epsilon_n^2\right]\right) \to 0.$$

In some cases it is also straightforward to consider the prior concentration with respect to the stronger $\|\cdot\|_\infty$-norm. For example, for a wide class of Gaussian process priors, the prior concentration $\Pi(\|f - f_0\|_\infty < \epsilon)$ has been extensively studied (see, for example, [16, 43, 45] for more details).

Lemma 2.4. *Suppose the sequence $(\epsilon_n)_{n=1}^\infty$ satisfies $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$. Then for any constant $C > 0$,*

$$\mathbb{P}_0\left(\int\prod_{i=1}^n \frac{p_f(\mathbf{x}_i, y_i)}{p_0(\mathbf{x}_i, y_i)}\Pi(\mathrm{d}f) \le \Pi(\|f - f_0\|_\infty < \epsilon_n)\exp\left[-\left(C + \frac{1}{\sigma^2}\right)n\epsilon_n^2\right]\right) \to 0.$$

Now we present the main result regarding the rates of contraction for Bayesian nonparametric regression. The proof is based on the modification of the prior-concentration-and-testing procedure. We also remark that the prior Π on f is not necessarily supported on a uniformly bounded function space, which is also a part of the major contribution of this work.

Theorem 2.1 (Generic Contraction)**.** *Let $(\epsilon_n)_{n=1}^\infty$ and $(\underline{\epsilon}_n)_{n=1}^\infty$ be sequences such that $\min(n\epsilon_n^2, n\underline{\epsilon}_n^2) \to \infty$ as $n \to \infty$, with $0 \le \underline{\epsilon}_n \le \epsilon_n \to 0$. Assume that the sieve $(\mathcal{F}_{m_n}(\delta))_{n=1}^\infty$ satisfies (2.2) with $m_n\epsilon_n^2 \to 0$ for some constant δ. In addition, assume that there exists another sequence $(k_n)_{n=1}^\infty \subset \mathbb{N}_+$ such that $k_n\underline{\epsilon}_n^2 = O(1)$. Suppose the following conditions hold for some constants $\omega, D > 0$ and sufficiently large n and M:*

$$\sum_{j=M}^\infty N_{nj}\exp\left(-Dnj^2\epsilon_n^2\right) \to 0, \tag{2.3}$$

$$\Pi(\mathcal{F}_{m_n}^c(\delta)) \lesssim \exp\left[-\left(2D + \frac{1}{\sigma^2}\right) n\underline{\epsilon}_n^2\right], \tag{2.4}$$

$$\Pi\left(B(k_n, \underline{\epsilon}_n, \omega)\right) \geq \exp\left(-Dn\underline{\epsilon}_n^2\right), \tag{2.5}$$

where $N_{nj} = \mathcal{N}(\xi j\epsilon_n, \mathcal{S}_{nj}(\epsilon_n), \|\cdot\|_2)$ *is the covering number of*

$$\mathcal{S}_{nj} = \{f \in \mathcal{F}_{m_n}(\delta) : j\epsilon_n < \|f - f_0\|_2 \leq (j+1)\epsilon_n\},$$

and $B(k_n, \underline{\epsilon}_n, \omega)$ *is defined in Lemma 2.3. Then* $\mathbb{E}_0\left[\Pi(\|f - f_0\|_2 > M\epsilon_n \mid \mathcal{D}_n)\right] \to 0$.

Remark 2.1. Recall that the idea of the sieve $\mathcal{F}_m(\delta)$ is that $\|f - f_0\|_\infty^2$ can be upper bounded by a constant multiple of $\|f - f_0\|_2^2 + \delta^2$ for any function $f \in \mathcal{F}_m(\delta)$. When $m_n\epsilon_n^2 \to 0$ and δ is a constant, $\|f - f_0\|_\infty^2$ is uniformly bounded for any $f \in \mathcal{F}_{m_n}(\delta)$.

Remark 2.2. In light of Lemma 2.4, by exploiting the proof of Theorem 2.1 we remark that when the assumptions and conditions in Theorem 2.1 hold with (2.5) replaced by $\Pi(\|f - f_0\|_\infty < \underline{\epsilon}_n) \geq \exp(-Dn\underline{\epsilon}_n^2)$, the same rate of contraction also holds: $\mathbb{E}_0\left[\Pi(\|f - f_0\|_2 > M\epsilon_n \mid \mathcal{D}_n)\right] \to 0$ for sufficiently large $M > 0$.

Remark 2.3. When the underlying true regression function f_0 is one-dimensional (*i.e.*, $p = 1$) and has certain smoothness level α in the Hölder or Sobolev sense (which will be made clear in Section 3) with $\alpha > 1/2$, m_n is typically chosen to be $m_n \asymp n^{1/(2\alpha+1)}$, possibly up to a logarithmic factor, and ϵ_n is typically $\epsilon_n \asymp n^{-\alpha/(2\alpha+1)}$, possibly up to a logarithmic factor. It follows immediately that $m_n\epsilon_n^2 \asymp n^{(1-2\alpha)/(1+2\alpha)} \to 0$, potentially up to a logarithmic factor, due to the requirement that $\alpha > 1/2$.

Remark 2.4. As pointed out by one of the referees, when σ is unknown, it is possible to extend Theorem 2.1 to the case where σ is assigned a prior density on $(0, \infty)$ and is positive in a neighborhood of σ_0 by leveraging the technique in [5]. The major complication comes from constructing suitable test functions when σ is unknown. The test function adopted in the proof of Theorem 2.1 is based on the likelihood ratio statistic, whereas that used in [5] is more mathematically involved.

3. Applications

In this section we consider three concrete priors on f for the nonparametric regression problem $y_i = f(x_i) + e_i$, $i = 1, \cdots, n$. For simplicity the design points are assumed to independently follow the one-dimensional uniform distribution $\text{Unif}(0, 1)$. For some of the examples, the results can be easily generalized to the case where the design points are multi-dimensional by considering the tensor-product basis functions. These results also generalize their respective counterparts in the literature.

3.1. Finite random series regression with adaptive rate

The finite random series prior [1, 30, 36, 46] is popular in the literature of Bayesian nonparametric theory. It is a class of hierarchical priors that first draw an integer-valued random variable serving as the number of "terms" to be used in a finite sum, and then sample the "term-wise" parameters given the number of "terms". The finite random series prior typically does not depend on the smoothness level of the true function, and often yields minimax-optimal rates of contraction (up to a logarithmic factor) in many nonparametric problems (*e.g.*, density estimation [30, 36] and fixed-design regression [1]). However, the adaptive rates of contraction for the finite random series prior in the random-design regression with respect to the integrated L_2-distance has not been established. In this subsection we address this issue by leveraging the technique developed in Section 2.

We first introduce the finite random series prior. Let $(\psi_k)_{k=1}^{\infty}$ be the Fourier basis in $L_2([0,1])$, *i.e.*, $\psi_1(x) = 1, \psi_{2k}(x) = \sin k\pi x$, and $\psi_{2k+1}(x) = \cos k\pi x$, $k \in \mathbb{N}_+$. Writing f in terms of the Fourier series expansion $f(\mathbf{x}) = \sum_{k=1}^{\infty} \beta_k \psi_k(\mathbf{x})$, we then assign the finite random series prior Π on f by considering the following prior distribution on the coefficients $(\beta_k)_{k=1}^{\infty}$: first sample an integer-valued random variable N from a density function π_N (with respect to the counting measure on $\mathbb{N}_+$), and then given $N = \widetilde{m}$, the coefficients β_k's are independently sampled according to

$$\Pi(\mathrm{d}\beta_k \mid N = \widetilde{m}) = \begin{cases} g(\beta_k)\mathrm{d}\beta_k, & \text{if } 1 \le k \le \widetilde{m}, \\ \delta_0(\mathrm{d}\beta_k), & \text{if } k > \widetilde{m}, \end{cases}$$

where g is an exponential power density $g(x) \propto \exp(-\tau_0|x|^{\tau})$ for some $\tau, \tau_0 > 0$ [37]. We further require that

$$\pi_N(\widetilde{m}) \ge \exp(-b_0 \widetilde{m} \log \widetilde{m}) \quad \text{and} \quad \sum_{N=\widetilde{m}+1}^{\infty} \pi_N(N) \le \exp(-b_1 \widetilde{m} \log \widetilde{m}) \tag{3.1}$$

for some constants b_0, b_1. The zero-truncated Poisson distribution $\pi_N(\widetilde{m}) = (\mathrm{e}^{\lambda} - 1)^{-1} \lambda^{\widetilde{m}} / \widetilde{m}! \mathbb{1}(\widetilde{m} \ge 1)$ satisfies condition (3.1) [47].

The true regression function f_0 is assumed to yield a Fourier expansion $f_0(x) = \sum_{k=1}^{\infty} \beta_{0k} \psi_k(x)$ with regard to $(\psi_k)_{k=1}^{\infty}$, and be in the α-Hölder ball

$$\mathfrak{C}_{\alpha}(Q) = \left\{ f(x) = \sum_{k=1}^{\infty} \beta_k \psi_k(x) : \sum_{k=1}^{\infty} k^{\alpha} |\beta_k| \le Q \right\},$$

where $\alpha > 1/2$ is the smoothness level, and $Q > 0$ is the α-Hölder radius. Note that the construction of the aforementioned finite random series prior does not require the knowledge of the smoothness level α. In the literature of Bayes theory, such a procedure is referred to as *adaptive*.

The following theorem shows that the constructed finite random series prior is adaptive and the rate of contraction $n^{-\alpha/(2\alpha+1)}(\log n)^t$ with respect to the integrated L_2-distance is minimax-optimal up to a logarithmic factor [39].

Theorem 3.1. *Suppose the true regression function $f_0 \in \mathfrak{C}_\alpha(Q)$ for some $\alpha > 1/2$ and $Q > 0$, and f is imposed the prior Π given above. Then there exists some sufficiently large constant $M > 0$ such that*

$$\mathbb{E}_0 \left[\Pi \left(\|f - f_0\|_2 > M n^{-\alpha/(2\alpha+1)} (\log n)^t \mid \mathcal{D}_n \right) \right] \to 0$$

for any $t > \alpha/(2\alpha + 1)$.

3.2. Block prior regression with adaptive and exact rate

In the literature of adaptive Bayesian procedure, the minimax-optimal rates of contraction are often obtained with an extra logarithmic factor. It typically requires extra work to obtain the exact minimax-optimal rate. Gao and Zhou [12] elegantly construct a modified block prior that yields rate-adaptive (*i.e.*, the prior does not depend on the smoothness level) and rate-exact contraction (*i.e.*, the contraction rate does not involve an extra logarithmic factor) for a wide class of nonparametric problems. Nevertheless, for nonparametric regression, [12] modifies the block prior by conditioning on the space of uniformly bounded functions. Requiring a known upper bound for the unknown f_0 when constructing the prior is restrictive, and it eliminates the popular Gaussian process priors. Besides the theoretical concern, the block prior itself is also a conditional Gaussian process and such a modification is inconvenient for implementation. In this section, we address this issue by showing that for nonparametric regression such a modification is not necessary.

Recall that in Section 3.1 we have introduced the Fourier basis functions $(\psi_k)_{k=1}^\infty$ in $L_2([0,1])$ with $\psi_1(x) = 1$, $\psi_{2k}(x) = \sqrt{2}\sin \pi k x$, and $\psi_{2k+1}(x) = \sqrt{2}\cos \pi k x$, $k \in \mathbb{N}_+$. The true regression function f_0 is also assumed to yield a Fourier expansion $f_0(x) = \sum_{k=1}^\infty \beta_{0k}\psi_k(x)$, and to be in the α-Sobolev ball

$$\mathcal{H}_\alpha(Q) = \left\{ f(x) = \sum_{k=1}^\infty \beta_k \psi_k(x) : \sum_{k=1}^\infty k^{2\alpha}\beta_k^2 \le Q \right\}$$

with radius $Q > 0$. Write $f(x) = \sum_{k=1}^\infty \beta_k \psi_k(x)$ in terms of the Fourier expansion. Similar to the finite random series prior, the block prior is constructed by assigning a prior distribution on the coefficients $(\beta_k)_{k=1}^\infty$ as follows. Given a sequence $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots)$ in the squared-summable sequence space l^2, define the ℓth block B_ℓ to be the integer index set $B_\ell = \{k_\ell, \cdots, k_{\ell+1} - 1\}$ and with size $n_\ell = |B_\ell| = k_{\ell+1} - k_\ell$, where $k_\ell = \lceil \mathrm{e}^\ell \rceil$. We use $\boldsymbol{\beta}_\ell = (\beta_j : j \in B_\ell) \in \mathbb{R}^{n_\ell}$ to denote the coefficients with indices lying in the ℓth block B_ℓ. The block prior then assigns the following distribution on the coefficients $(\beta_k)_{k=1}^\infty$:

$$\boldsymbol{\beta}_\ell \mid A_\ell \sim \mathrm{N}(\mathbf{0}, A_\ell \mathbf{I}_{n_\ell}), \quad A_\ell \sim g_\ell, \quad \text{independently for each } \ell,$$

where $(g_\ell)_{\ell=0}^\infty$ is a sequence of densities satisfying the following properties:

1. There exists $c_1 > 0$ such that for any ℓ and $t \in [\mathrm{e}^{-\ell^2}, \mathrm{e}^{-\ell}]$,

$$g_\ell(t) \ge \exp(-c_1 \mathrm{e}^\ell). \tag{3.2}$$

2. There exists $c_2 > 0$ such that for any ℓ,

$$\int_0^\infty t g_\ell(t)\mathrm{d}t \le 4\exp(-c_2\ell^2). \tag{3.3}$$

3. There exists $c_3 > 0$ such that for any ℓ,

$$\int_{\mathrm{e}^{-\ell^2}}^\infty g_\ell(t)\mathrm{d}t \le \exp(-c_3\mathrm{e}^\ell). \tag{3.4}$$

The existence of a sequence of densities $(g_\ell)_{\ell=0}^\infty$ satisfying (3.2), (3.3), and (3.4) is verified in [12] (see Proposition 2.1 in [12]).

Our major improvement for the block prior regression is the following theorem, which shows that the (un-modified) block prior yields rate-exact Bayesian adaptation for nonparametric regression.

Theorem 3.2. *Suppose the true regression function $f_0 \in \mathcal{H}_\alpha(Q)$ for some $\alpha > 1/2$ and $Q > 0$, and $f(x) = \sum_{k=1}^\infty \beta_k\psi_k(x)$ is imposed the block prior Π as described above. Then*

$$\mathbb{E}_0\left[\Pi\left(\|f - f_0\|_2 > Mn^{-\alpha/(2\alpha+1)} \mid \mathcal{D}_n\right)\right] \to 0$$

for some sufficiently large constant $M > 0$.

Rather than using the sieve $\mathcal{F}_n$ proposed in Theorem 2.1 in [12], which does not necessarily satisfy (2.2), we construct $\mathcal{F}_{m_n}(\delta)$ in a slightly different fashion:

$$\mathcal{F}_{m_n}(Q) = \left\{ f(x) = \sum_{k=1}^\infty \beta_k\psi_k(x) : \sum_{k=1}^\infty (\beta_k - \beta_{0k})^2 k^{2\alpha} \le Q^2 \right\}$$

with $m_n \asymp n^{1/(2\alpha+1)}$ and $\delta = Q$. The covering number N_{nj} can be bounded using the metric entropy for Sobolev balls (see, for example, Lemma 6.4 in [2]), and the rest conditions in Theorem 2.1 can be verified using similar techniques as in [12].

As discussed in Section 4.2 in [12], the block prior can be conveniently extended to the wavelet basis functions and wavelet series. The wavelet basis functions are another widely-used class of orthonormal basis functions for $L_2([0,1])$. Let $(\psi_{jk})_{j\in\mathbb{N},k\in I_j}$ be an orthonormal basis of compactly supported wavelets for $L_2([0,1])$, with j referring to the so-called "resolution level", and k is the "translation" index (see, for example, Section E.3 in [16]). We adopt the convention that the index set I_j for the jth resolution level runs through $\{0, 1, \cdots, 2^j - 1\}$. The exact definition and specific formulas for the wavelet basis are not of great interest to us, and for a complete and thorough review of wavelets from a statistical perspective, we refer to [18]. We shall assume that the wavelet basis ψ_{jk}'s are appropriately selected such that for any $f(x) = \sum_{j=0}^\infty \sum_{k\in I_j} \beta_{jk}\psi_{jk}(x)$, the following inequality hold [6, 7, 20]: $\|f\|_\infty \le \sum_{j=0}^\infty 2^{j/2} \max_{k\in I_j} |\beta_{jk}|$.

Write f in terms of the wavelet series expansion $f(x) = \sum_{j=0}^\infty \sum_{k\in I_j} \beta_{jk}\psi_{jk}(x)$. The block prior for the wavelet series is then introduced through the wavelet

coefficients β_{jk}'s as follows:

$$\boldsymbol{\beta}_j \mid A_j \sim \mathrm{N}(\mathbf{0}, A_k \mathbf{I}_{n_k}), \quad A_j \sim g_j, \quad \text{independently for each } j,$$

where $\boldsymbol{\beta}_j = (\beta_{jk} : k \in I_j)$, $n_k = |I_k| = 2^j$, and g_j is given by

$$g_j(t) = \begin{cases} \mathrm{e}^{j^2 \log 2}(\mathrm{e}^{-2^j \log 2} - T_j)t + T_j, & 0 \le t \le \mathrm{e}^{-j^2 \log 2}, \\ \mathrm{e}^{-2^j \log 2}, & \mathrm{e}^{-j^2 \log 2} < t \le \mathrm{e}^{-j \log 2}, \\ 0, & t > \mathrm{e}^{-j \log 2}, \end{cases}$$

$$T_j = \exp\left[(1 + j^2)\log 2\right] - \exp\left[(-2^j + j^2 - j)\log 2\right] + \mathrm{e}^{-2^j \log 2}.$$

We further assume that f_0 is an α-Sobolev function. For the block prior regression via wavelet series, the rate-exact Bayesian adaptation also holds.

Theorem 3.3. *Suppose the true regression function $f_0 \in \mathcal{H}_\alpha(Q)$ for some $\alpha > 1/2$ and $Q > 0$, and $f(x) = \sum_{j=0}^{\infty} \sum_{k \in I_j} \beta_{jk} \psi_{jk}(x)$ is imposed the block prior for wavelet series Π as described above. Then there exists some sufficiently large constant $M > 0$ such that*

$$\mathbb{E}_0 \left[\Pi \left(\|f - f_0\|_2 > M n^{-\alpha/(2\alpha+1)} \mid \mathcal{D}_n \right) \right] \to 0.$$

3.3. Beyond Fourier series: Gaussian spline regression

The previous two examples show that Theorem 2.1 can be applied to priors through Fourier series expansions. It turns out that this theorem also works for prior distributions beyond Fourier basis, and we provide an example in this subsection.

Spline functions or splines are defined piecewise by polynomials on subintervals $[t_0, t_1), [t_1, t_2), \ldots, [t_{K-1}, t_K]$ that are also globally smooth on the entire domain $[t_0, t_K]$, where K is the number of subintervals. Without loss of generality, we may further require that $t_0 = 0$ and $t_K = 1$. A spline function is said to be of order q for some positive integer q, if the involved polynomials are of degrees at most $q - 1$. Given the order q and the number of subintervals K, the space of spline functions forms a linear space with dimension $m_0 = q + K - 1$, and a basis for this linear space is also a set of spline functions, referred to as *B-splines* and denoted as $(B_k)_{k=1}^{m_0}$. We refer the readers to [9] and [35] for a systematic introduction of the spline functions. We present the following facts regarding the approximation property of splines, and the norm equivalence between spline functions and the coefficients of the corresponding B-splines. These results can be found in [9].

Lemma 3.1. *Let f_0 be an α-Hölder function with $\alpha > 0$, and $q \ge \alpha$. Then there exists a constant C depending on f_0, q, and α, and $(\beta_{0k})_{k=1}^{m_0} \subset \mathbb{R}$ with $m_0 = q + K - 1$, such that*

$$\left\| \sum_{k=1}^{m_0} \theta_k B_k - f_0 \right\|_\infty \le C m_0^{-\alpha}.$$

Furthermore, for any $\beta_1, \ldots, \beta_{m_0} \in \mathbb{R}$,

$$\max_{1\le k\le m_0} |\beta_k| \asymp \left\| \sum_{k=1}^{m_0} \beta_k B_k \right\|_\infty, \quad \left(\sum_{k=1}^{m_0} \beta_k^2 \right)^{1/2} \asymp \sqrt{m_0} \left\| \sum_{k=1}^{m_0} \beta_k B_k \right\|_2.$$

Assume that the true regression function f_0 is an α-Hölder function with $\alpha > 1/2$. We now present the Gaussian spline prior, which simplifies the version presented in [11]. Assume that f is a spline function of order $q \ge \alpha$ on $[0,1]$, and the number of subintervals is K. Then we write f in terms of a linear combination of the B-splines $f(x) = \sum_{k=1}^{m_0} \beta_k B_k(x)$. The Gaussian spline prior assigns a prior distribution on f by letting the coefficients $\beta_1, \ldots, \beta_{m_0}$ independently follow the standard normal distribution $\mathrm{N}(0,1)$. Allowing m_0 grows moderately with the sample size n, we show in the following theorem that the posterior contraction rate with respect to $\|\cdot\|_2$ is minimax-optimal up to a logarithmic factor.

Theorem 3.4. *Suppose the true regression function* $f_0 \in \mathfrak{C}^\alpha(Q)$ *for some* $\alpha > 1/2$ *and* $Q > 0$, *and* $f(x)$ *is assigned the Gaussian spline prior* Π *as described above with order* $q \ge \alpha$ *and the number of subintervals* K. *If* $m_0 = q + K - 1 \asymp n^{1/(2\alpha+1)}$, *then there exists some sufficiently large constant* $M > 0$ *such that*

$$\mathbb{E}_0 \left[\Pi \left(\|f - f_0\|_2 > M n^{-\alpha/(2\alpha+1)} (\log n)^{1/2} \mid \mathcal{D}_n \right) \right] \to 0.$$

We briefly compare the result of Theorem 3.4 with that in [11], which considered the empirical L_2-distance and obtained the minimax-optimal rate up to a logarithmic factor in the fixed-design regression problem. In contrast, we put the prior models in the context of the random-design regression and consider the integrated L_2-distance, which can be viewed as a complement of the contraction result presented in [11]. Although the posterior contraction in Theorem 3.4 is not adaptive to the smoothness level α, one can further assign a prior distribution to m_0 in a similar fashion to that in [11] to make it rate-adaptive.

4. Extensions

4.1. Extension to the fixed-design regression

So far, the design points $(\mathbf{x}_i)_{i=1}^n$ in this paper are assumed to be randomly sampled from $[0,1]^p$. This is referred to as the random-design regression problem. There are, however, many cases where the design points $(\mathbf{x}_i)_{i=1}^n$ are fixed and can be controlled. One of the examples is the design and analysis of computer experiments [8, 34]. To emulate a computer model, the design points are typically manipulated so that they are reasonably spread. In some physical experiments the design points can also be required to be fixed [40]. In this subsection we show that by slightly extending Theorem 2.1, the integrated L_2-distance contraction is also obtainable under similar conditions when the design points are fixed but reasonably selected.

Suppose that the design points $(x_i)_{i=1}^n \subset [0,1]$ are one-dimensional and fixed. Intuitively, the design points need to be relatively "spread" so that the global behavior of the true signal f_0 can be recovered as much as possible. Formally, we require that the design points satisfy

$$\sup_{x\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i \le x) - x \right| = O\left(\frac{1}{n}\right). \tag{4.1}$$

A simple example of such design is the univariate equidistance design, *i.e.*, $x_i = (i-1/2)/n$ (see, for example, [52, 53]).

Now we extend the framework in Section 2 to the (one-dimensional) fixed-design regression problem. Specifically, this amounts to modifying the requirement for the sieve in (2.2): Let $\eta > 0$ be a constant, and for any $m \in \mathbb{N}_+$ and $\delta > 0$, assume that $\mathcal{F}_m(\delta)$ satisfies

$$\begin{aligned} &f, f_1 \in \mathcal{F}_m(\delta) \Rightarrow \left|\mathbb{P}_n(f-f_0)^2 - \|f-f_0\|_2^2\right| \le \eta\left(\frac{m}{n}\|f-f_0\|_2^2 + \frac{\delta}{\sqrt{n}}\|f-f_0\|_2\right), \\ &\text{and } \left|\mathbb{P}_n(f-f_1)^2 - \|f-f_1\|_2^2\right| \le \eta\left(\frac{m}{n}\|f-f_1\|_2^2 + \frac{\delta}{\sqrt{n}}\|f-f_1\|_2\right). \end{aligned} \tag{4.2}$$

With the above ingredients, we present the following modification of Theorem 2.1 for the fixed-design regression, which might be of independent interest as well.

Theorem 4.1 (Generic Contraction, Fixed-design)**.** *Suppose the design points $(x_i)_{i=1}^n$ are fixed and satisfy (4.1). Let $(\epsilon_n)_{n=1}^\infty$ and $(\underline{\epsilon}_n)_{n=1}^\infty$ be sequences such that $\min(n\epsilon_n^2, n\underline{\epsilon}_n^2) \to \infty$ as $n \to \infty$ with $0 \le \underline{\epsilon}_n \le \epsilon_n \to 0$. Let the sieves $(\mathcal{F}_{m_n}(\delta))_{n=1}^\infty$ satisfy (4.2) for some constant $\delta > 0$, where $m_n \to \infty$ and $m_n/n \to 0$. Suppose the conditions (2.3), (2.4), and $\Pi(\|f-f_0\|_\infty < \underline{\epsilon}_n) \ge \exp(-Dn\underline{\epsilon}_n^2)$ hold for some constant $D > 0$ and sufficiently large n and M. Then*

$$\mathbb{E}_0\left[\Pi(\|f-f_0\|_2 > M\epsilon_n \mid \mathcal{D}_n)\right] \to 0.$$

As a sample application of Theorem 4.1, we consider one of the most popular Gaussian processes $\mathrm{GP}(0, K)$ with the covariance function of the squared-exponential form $K(x,x') = \exp\left[-(x-x')^2\right]$ [29]. We show that optimal rates of contraction with respect to the integrated L_2-distance is also attainable when the design points are reasonably selected, in contrast to most Bayesian literatures that obtain rates of contraction with respect to the empirical L_2-distance.

Given constants $c, Q > 0$, we assume that the underlying true regression function f_0 lies in the following function class

$$\mathcal{A}_c(Q) = \left\{ f(x) = \sum_{k=1}^{\infty} \beta_k \psi_k(x) : \sum_{k=1}^{\infty} \beta_k^2 \exp\left(\frac{k^2}{c}\right) \le Q^2 \right\}.$$

The function class $\mathcal{A}_c(Q)$ is closely related to the reproducing kernel Hilbert space (RKHS) associated with $\mathrm{GP}(0, K)$. For a complete and thorough review

of RKHS from a Bayesian perspective, we refer to [44]. A key feature of the functions in $\mathcal{A}_c(Q)$ is that they are "supersmooth", *i.e.*, they are infinitely differentiable. For the squared-exponential Gaussian process prior, the following property regarding the corresponding RKHS is available by applying Theorem 4.1 in [44].

Lemma 4.1. *Let $\mathbb{H}$ be the RKHS associated with the squared-exponential Gaussian process* $\mathrm{GP}(0, K)$*, where $K(x, x') = \exp[-(x - x')^2]$. Then $\mathcal{A}_4(Q) \subset \mathbb{H}$ for any $Q > 0$.*

Under the squared-exponential Gaussian process prior Π, the rate of contraction of a supersmooth $f_0 \in \mathcal{A}_4(Q)$ is $1/\sqrt{n}$ up to a logarithmic factor.

Theorem 4.2. *Assume that the design points $(x_i)_{i=1}^n$ are fixed and satisfy (4.1). Suppose the true regression function $f_0 \in \mathcal{A}_4(Q)$ for some $Q > 0$, and f follows the squared-exponential Gaussian process prior Π. Then there exists some sufficiently large constant $M > 0$ such that*

$$\mathbb{E}_0 \left[\Pi \left(\|f - f_0\|_2 > M n^{-1/2} (\log n) \mid \mathcal{D}_n \right) \right] \to 0.$$

Remark 4.1. For the squared-exponential Gaussian process regression with random design, the rate of contraction with respect to the integrated L_2-distance for $f_0 \in \mathcal{A}_4(Q)$ has been studied in the literature. In contrast, we remark that for the fixed-design regression problem, Theorem 4.2 is new and original, and provides a stronger result compared to the existing literature (see, for example, Theorem 10 in [41]).

4.2. Extension to sparse additive models in high dimensions

We have so far considered that the design space is low dimensional with fixed p. Nonetheless, the rapid development of technology has been enabling scientists to collect data with high-dimensional covariates, where the number of covariates p can be much larger than the sample size n, to explore the potentially nonlinear relationship between these covariates and certain outcome of interest. The emergence of high dimensional prediction problems naturally motivates the study of nonparametric regression in high dimensions [51]. In this section, we focus on one class of high-dimensional nonparametric regression problem, known as *sparse additive models*, and illustrate that with suitable prior specification, the framework for low-dimensional Bayesian nonparametric regression naturally extends to such a high-dimensional scenario.

We first review some background regarding the sparse additive models. Consider the additive regression model $y_i = f(\mathbf{x}_i) + e_i$, where the regression function $f(\mathbf{x}_i)$ is of an additive structure of the covariates $f(\mathbf{x}_i) = \mu + \sum_{j=1}^p f_j(x_{ij})$. Without loss of generality, one can assume that each component $f_j(x_j)$ is centered: $\int_0^1 f_j(x_j) \mathrm{d}x_j = 0$, $j = 1, \cdots, p$. For sparse additive models in high dimensions, the number of covariates p is typically much larger than the sample size n, and the underlying true regression function f_0 depends only on a small

number of covariates, say, $x_{j_1}, \ldots, x_{j_q}$, *i.e.*, f_0 is of a sparse additive structure $f_0(\mathbf{x}_i) = \mu_0 + \sum_{r=1}^{q} f_{0j_r}(x_{ij_r})$, where each $f_{0r} : [0,1] \to \mathbb{R}$ is a univariate function, and q is the number of active covariates that does not change with sample size. Furthermore, the indices of these active covariates $\{j_1, \ldots, j_q\}$ and q are unknown. This is referred to as the sparse additive models in high dimensions in the literature [19, 22, 25, 27, 28]. There have also been several works regarding Bayesian modeling of sparse additive models in high dimensions, see, for example, [24, 33, 51].

To model the sparsity occurring in the high-dimensional additive regression model, we consider the following parameterization of f by introducing the binary covariate-selection variables $z_1, \ldots, z_p \in \{0,1\}$:

$$f(\mathbf{x}) = \mu + \sum_{j=1}^{p} z_j f_j(x_j), \quad z_j \in \{0,1\}, \quad j = 1, \cdots, p, \tag{4.3}$$

where $z_j = 1$ indicates that the jth covariate is active and $z_j = 0$ otherwise. Following the strategy in Section 2, each component function f_j is assigned a prior distribution independently across $j = 1, \ldots, p$. We complete the prior distribution Π by imposing the selection variables z_j with a Bernoulli distribution $z_j \sim \text{Bernoulli}(1/p)$. The Bernoulli prior for sparsity has been widely adopted in other high-dimensional Bayesian models with variable selection structures (see, for example, [4, 31, 32]).

We now extend Theorem 2.1 to sparse additive models by modifying the sieve property (2.2). Denote $\mathbf{z} = [z_1, \ldots, z_p]^{\mathrm{T}} \in \{0,1\}^p$ and let A be a positive integer. Let $\eta > 0$ be a constant. For any $m \in \mathbb{N}_+$ and $\delta > 0$, we consider the sieve $\mathcal{G}_m^A(\delta) = \bigcup_{\|\mathbf{z}\|_1 \le Aq} \mathcal{G}_m(\delta, \mathbf{z})$, where $\mathcal{G}_m(\delta, \mathbf{z})$ with $\|\mathbf{z}\|_1 \le A$ satisfies the following condition: there exists some constant $\eta > 0$ such that

$$f \in \mathcal{G}_m(\delta, \mathbf{z}) \Longrightarrow \|f - f_0\|_\infty^2 \le \eta(A^2 m \|f - f_0\|_2^2 + \delta^2). \tag{4.4}$$

Theorem 4.3 (Generic Contraction, Sparse Additive Models)**.** *Consider the aforementioned sparse additive model in high dimensions. Let* $(\epsilon_n)_{n=1}^\infty$ *and* $(\underline{\epsilon}_n)_{n=1}^\infty$ *be sequences such that* $\min(n\epsilon_n^2, n\underline{\epsilon}_n^2) \to \infty$ *as* $n \to \infty$*, with* $0 \le \underline{\epsilon}_n \le \epsilon_n \to 0$*. Assume that there exist sieves of the form* $\mathcal{G}_{m_n}^{A_n}(\delta) = \bigcup_{\|\mathbf{z}\|_1 \le A_n q} \mathcal{G}_{m_n}(\delta, \mathbf{z})$*, where* $\mathcal{G}_m(\delta, \mathbf{z})$ *satisfies* (4.4)*,* $(m_n)_{n=1}^\infty$*,* $(A_n)_{n=1}^\infty$ *are sequences such that* $A_n m_n \epsilon_n^2 \to 0$*, and* δ *is some constant. Let* $(k_n)_{n=1}^\infty$ *be another sequence such that* $k_n \underline{\epsilon}_n^2 = O(1)$*. Suppose the following conditions hold for some constants* $\omega, D > 0$ *and sufficiently large* n *and* M*:*

$$\sum_{j=M}^{\infty} N_{nj}^{A_n} \exp\left(-Dnj^2\epsilon_n^2\right) \to 0, \tag{4.5}$$

$$\Pi\left(\mathcal{G}_{m_n}^{A_n}(\delta)^c\right) \lesssim \exp\left[-\left(2D + \frac{1}{\sigma^2}\right) n\underline{\epsilon}_n^2\right], \tag{4.6}$$

$$\Pi\left(\widetilde{B}_n(k_n, \underline{\epsilon}_n, \omega)\right) \ge \exp\left(-Dn\underline{\epsilon}_n^2\right), \tag{4.7}$$

where for any $m \in \mathbb{N}_+$ *and* $\epsilon, \omega > 0$,

$$\widetilde{B}(m, \epsilon, \omega) = \left\{ \|f - f_0\|_2 \le \epsilon, \|f - f_0\|_\infty^2 \le \eta'(m\|f - f_0\|_2^2 + \delta^2) \right\}$$

for some constant $\eta' > 0$, *and* $N_{nj}^{A_n} = \mathcal{N}(\xi j \epsilon_n, \mathcal{S}_{nj}^{A_n}(\epsilon_n), \|\cdot\|_2)$ *is the covering number of*

$$\mathcal{S}_{nj}^{A_n} = \{ f \in \mathcal{G}_{m_n}^{A_n}(\delta) : j\epsilon_n < \|f - f_0\|_2 \le (j+1)\epsilon_n \}.$$

Then $\mathbb{E}_0 \left[\Pi(\|f - f_0\|_2 > M\epsilon_n \mid \mathcal{D}_n) \right] \to 0$.

The above framework is quite flexible and can be applied to a variety of prior models. For example, let us extend the finite random series prior discussed in Section 3.1 to the sparse additive models as follows: We model each component $f_j(x_j)$ via a Fourier series $f_j(x_j) = \sum_{k=1}^{\infty} \beta_{jk} \psi_k(x_j)$, where $(\psi_k)_{k=1}^{\infty}$ are the Fourier basis introduced in Section 3.1. Then the coefficients $(\beta_{jk} : j = 1, \cdots, p, k = 2, 3, \cdots)$ are assigned the following prior distributions: First sample an integer-valued random variable N with density π_N satisfying (3.1), and then given $N = \widetilde{m}$, sample the coefficients $(\beta_{jk} : j = 1, \cdots, p, k = 2, 3, \cdots)$ with

$$\Pi(\mathrm{d}\beta_{jk} \mid N = \widetilde{m}) = \begin{cases} g(\beta_{jk})\mathrm{d}\beta_{jk}, & \text{if } 2 \le k \le \widetilde{m}, \\ \delta_0(\mathrm{d}\beta_{jk}), & \text{if } k > \widetilde{m}, \end{cases}$$

independently for all $j = 1, \cdots, p$ and $k = 2, 3, \cdots$, where $g(x) \propto \exp(-\tau_0 |x|^\tau)$ for some $\tau, \tau_0 > 0$. Finally let μ follow a prior distribution with density $\pi(h)$ supported on $\mathbb{R}$, and set $\beta_{j1} = -\sum_{k=2}^{\widetilde{m}} \beta_{jk} \int_0^1 \psi_k(x_j)\mathrm{d}x_j$, $j = 1, \cdots, p$ to ensure each component f_j integrates to 0. The prior specification is completed by combining the aforementioned Bernoulli prior on $\mathbf{z}$.

Theorem 4.4. *Consider the sparse additive model* (4.3) *with the above prior distribution, and the dimension of the design space* $p \ge 2$ *possibly grows with the sample size* n. *Suppose the true regression function yields an additive structure:* $f_0(\mathbf{x}) = \mu + \sum_{r=1}^{q} f_{0j_r}(x_{j_r})$, *where each* $f_{0j} \in \mathfrak{C}_\alpha(Q)$ *for some* $\alpha > 1/2$ *and* $Q > 0$ *for all* $j \in \{j_1, \cdots, j_q\}$, *and* q *does not change with* n. *Assume the dimension* p *satisfies* $\log p \lesssim \log n$. *Let* f *be imposed the prior* Π *given above. Then there exists some sufficiently large constant* $M > 0$ *such that*

$$\mathbb{E}_0 \left\{ \Pi \left[\|f - f_0\|_2 > M n^{-\alpha/(2\alpha+1)} (\log n)^t \;\middle|\; \mathcal{D}_n \right] \right\} \to 0$$

for any $t > \alpha/(2\alpha + 1)$.

Remark 4.2. The minimax rate of convergence with respect to $\|\cdot\|_2$ for sparse additive models is $n^{-\alpha/(2\alpha+1)} + (\log p)/n$, provided that $\log p \lesssim n^c$ for some $c < 1$ [51]. The first term $n^{-\alpha/(2\alpha+1)}$ is the usual rate for estimating a one-dimensional α-smooth function, and the second term $(\log p)/n$ comes from the complexity of selecting the q active covariates $x_{j_1}, \cdots, x_{j_q}$ among $x_1, \cdots, x_p$. Under the assumption that $\log p \lesssim \log n$, the minimax rate of convergence is dominated by the first term $n^{-\alpha/(2\alpha+1)}$. Thus Theorem 4.4 states that with the

aforementioned finite random series prior for the sparse additive model in high dimensions, the rate of contraction is adaptive and minimax-optimal modulus a logarithmic factor, which can be also viewed as a non-trivial application of the result in Section 3.1 to sparse additive model in high dimensions.

5. Proofs of the main results

Proof of Lemma 2.1. Recall the assumption

$$\|f - f_0\|_\infty^2 \lesssim m\|f - f_0\|_2^2 + \delta^2. \tag{5.1}$$

Let us take $\xi = 1/(4\sqrt{2})$. Define the test function to be $\phi_n = \mathbb{1}\{T_n > 0\}$, where

$$T_n = \sum_{i=1}^n y_i(f_1(\mathbf{x}_i) - f_0(\mathbf{x}_i)) - \frac{1}{2}n\mathbb{P}_n(f_1^2 - f_0^2) - \frac{\sqrt{n}}{8\sqrt{2}}\|f_1 - f_0\|_2\sqrt{n\mathbb{P}_n(f_1 - f_0)^2}.$$

Before proceeding to the analysis of the type I and type II error probabilities, we introduce the motivation of the choice of T_n as the test statistic. In fact, by the well-known Neyman-Pearson lemma, the most powerful test is the likelihood ratio test. In the context of random-design normal regression, the likelihood ratio test for testing $H_0 : f = f_0$ against $H_A : f = f_1$ is equivalent to rejecting f_0 for large value of

$$\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n[y_i - f_1(\mathbf{x}_i)]^2 + \frac{1}{2\sigma^2}\sum_{i=1}^n[y_i - f_0(\mathbf{x}_i)]^2\right\} = \exp\left\{\frac{1}{\sigma^2}\sum_{i=1}^n y_i(f_1(\mathbf{x}_i) - f_0(\mathbf{x}_i)) - \frac{1}{2\sigma^2}n\mathbb{P}_n(f_1^2 - f_0^2)\right\},$$

which is also equivalent to rejecting f_0 for large value of T_n.

The proof idea is similar to constructing the likelihood ratio test for establishing the posterior contraction with regard to $L_2(\mathbb{P}_n)$-distance in the fixed-design regression problems, which are commonly encountered in the literature. We present the proof here for the sake of mathematical rigor.

We first consider the type I error probability. Under $\mathbb{P}_0$, we have $y_i = f_0(\mathbf{x}_i) + e_i$, where e_i's are i.i.d. $\mathrm{N}(0, \sigma^2)$ noise. Therefore, there exists a constant $C_1 > 0$ such that $\mathbb{P}_0(e_i > t) \le \exp(-4C_1t^2)$ for all $t > 0$. Then for a sequence $(a_i)_{i=1}^n \in \mathbb{R}^n$, the Chernoff bound yields $\mathbb{P}_0\left(\sum_{i=1}^n a_ie_i \ge t\right) \le \exp\left(-4C_1t^2/\sum_{i=1}^n a_i^2\right)$. Now we set $a_i = f_1(\mathbf{x}_i) - f_0(\mathbf{x}_i)$ and

$$t = \frac{1}{2}n\mathbb{P}_n(f_1 - f_0)^2 + \frac{\sqrt{n}}{8\sqrt{2}}\|f_1 - f_0\|_2\sqrt{n\mathbb{P}_n(f_1 - f_0)^2}.$$

Clearly,

$$t^2 \ge n\mathbb{P}_n(f_1 - f_0)^2\left[\frac{1}{4}n\mathbb{P}_n(f_1 - f_0)^2 + \frac{1}{128}n\|f_1 - f_0\|_2^2\right]$$

$$\geq n\mathbb{P}_n(f_1 - f_0)^2 \left[\frac{1}{128} n\|f_1 - f_0\|_2^2\right].$$

Then under $\mathbb{P}_0(\cdot \mid \mathbf{x}_1, \cdots, \mathbf{x}_n)$, we have

$$\mathbb{E}_0(\phi_n \mid \mathbf{x}_1, \cdots, \mathbf{x}_n) \leq \exp\left(-\frac{C_1}{32} n\|f_1 - f_0\|_2^2\right).$$

It follows that the unconditioned error can be bounded:

$$\mathbb{E}_0\phi_n \leq \exp\left(-\frac{C_1}{32} n\|f_1 - f_0\|_2^2\right).$$

We next consider the type II error probability. Under $\mathbb{P}_f$, we have $y_i = f(\mathbf{x}_i) + e_i$ with e_i's being i.i.d. mean-zero Gaussian. Then for any f with $\|f - f_1\|_2 \leq \|f_0 - f_1\|_2/(4\sqrt{2}) \leq \|f_0 - f_1\|_2/4$, we have

$$\begin{aligned}\mathbb{E}_f(1 - \phi_n) \leq \mathbb{E}\left[\mathbb{1}\left\{\mathbb{P}_n(f - f_1)^2 \leq \frac{1}{16}\mathbb{P}_n(f_1 - f_0)^2\right\} \mathbb{E}_f(1 - \phi_n \mid \mathbf{x}_1, \cdots, \mathbf{x}_n)\right] \\ + \mathbb{P}\left(\mathbb{P}_n(f - f_1)^2 > \frac{1}{16}\mathbb{P}_n(f_1 - f_0)^2\right).\end{aligned}$$

When $\mathbb{P}_n(f - f_1)^2 \leq \mathbb{P}_n(f_1 - f_0)^2/16$, we have

$$\begin{aligned}&T_n + \frac{\sqrt{n}}{8\sqrt{2}}\|f_1 - f_0\|_2\sqrt{n\mathbb{P}_n(f_1 - f_0)^2} \\ &\quad= \sum_{i=1}^n e_i\left[f_1(\mathbf{x}_i) - f_0(\mathbf{x}_i)\right] + n\mathbb{P}_n(f - f_1)(f_1 - f_0) + \frac{1}{2}n\mathbb{P}_n(f_1 - f_0)^2 \\ &\quad\geq \sum_{i=1}^n e_i\left[f_1(\mathbf{x}_i) - f_0(\mathbf{x}_i)\right] + \frac{1}{4}n\mathbb{P}_n(f_1 - f_0)^2.\end{aligned}$$

Now set

$$R = R(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \frac{1}{4}n\mathbb{P}_n(f_1 - f_0)^2 - \frac{\sqrt{n}}{8\sqrt{2}}\|f_1 - f_0\|_2\sqrt{n\mathbb{P}_n(f_1 - f_0)^2}.$$

Then given $R \geq \sqrt{n}\|f_1 - f_0\|_2\sqrt{n\mathbb{P}_n(f_1 - f_0)^2}/(8\sqrt{2})$, we use the Chernoff bound to obtain

$$\begin{aligned}\mathbb{P}_f\left(T_n < 0 \mid \mathbf{x}_1, \cdots, \mathbf{x}_n\right) &\leq \mathbb{P}\left(\sum_{i=1}^n e_i[f_1(\mathbf{x}_i) - f_0(\mathbf{x}_i)] \leq -R \mid \mathbf{x}_1, \cdots, \mathbf{x}_n\right) \\ &\leq \exp\left(-\frac{4C_1R^2}{n\mathbb{P}_n(f_1 - f_0)^2}\right) \leq \exp\left(-\frac{C_1 n\|f_1 - f_0\|_2^2}{32}\right).\end{aligned}$$

On the other hand,

$$\mathbb{P}\left(R < \frac{\sqrt{n}}{8\sqrt{2}}\|f_1 - f_0\|_2\sqrt{n\mathbb{P}_n(f_1 - f_0)^2}\right)$$

$$= \mathbb{P}\left(\mathbb{G}_n(f_1 - f_0)^2 < -\frac{\sqrt{n}}{2}\|f_1 - f_0\|_2^2\right).$$

It follows that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left\{\mathbb{P}_n(f - f_1)^2 \le \frac{1}{16}\mathbb{P}_n(f_1 - f_0)^2\right\}\mathbb{E}_f(1 - \phi_n \mid \mathbf{x}_1, \cdots, \mathbf{x}_n)\right] \\
&\quad \le \mathbb{E}\left[\mathbb{1}\left\{R \ge \frac{\sqrt{n}}{8\sqrt{2}}\|f_1 - f_0\|_2\sqrt{n\mathbb{P}_n(f_1 - f_0)^2}, \right.\right. \\
&\qquad\qquad \left.\left.\mathbb{P}_n(f - f_1)^2 \le \frac{1}{16}\mathbb{P}_n(f_1 - f_0)^2\right\}\mathbb{P}_f(T_n < 0 \mid \mathbf{x}_1, \cdots, \mathbf{x}_n)\right] \\
&\quad + \mathbb{P}\left(R < \frac{\sqrt{n}}{8\sqrt{2}}\|f_1 - f_0\|_2\sqrt{n\mathbb{P}_n(f_1 - f_0)^2}\right) \\
&\quad \le \exp\left(-\frac{C_1}{32}n\|f_1 - f_0\|_2^2\right) + \mathbb{P}\left(\mathbb{G}_n(f_1 - f_0)^2 < -\frac{\sqrt{n}}{2}\|f_1 - f_0\|_2^2\right).
\end{aligned}$$

Using Bernstein's inequality (Lemma 19.32 in [42]), we obtain the tail probability of the empirical process $\mathbb{G}_n(f_1 - f_0)^2$

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{G}_n(f_1 - f_0)^2 < -\frac{\sqrt{n}}{2}\|f_1 - f_0\|_2^2\right) \\
&\quad \le \exp\left(-\frac{1}{4}\frac{n\|f_1 - f_0\|_2^4/4}{\mathbb{E}(f_1 - f_0)^4 + \|f_1 - f_0\|_2^2\|f_1 - f_0\|_\infty^2/2}\right) \\
&\quad \le \exp\left(-\frac{C'n\|f_1 - f_0\|_2^2}{m\|f_1 - f_0\|_2^2 + \delta^2}\right),
\end{aligned}$$

for some constant $C' > 0$, where we use the relation (5.1). On the other hand, when $\mathbb{P}_n(f - f_1)^2 > \mathbb{P}_n(f_1 - f_0)^2/16$, we again use Bernstein's inequality and the fact that $f \in \{f \in \mathcal{F}_m(\delta) : \|f - f_1\|_2^2 \le 2^{-5}\|f_0 - f_1\|_2^2\}$ to compute

$$\mathbb{P}\left(\mathbb{P}_n(f - f_1)^2 > \frac{1}{16}\mathbb{P}_n(f_1 - f_0)^2\right) \le \exp\left(-\frac{1}{4}\frac{n\|f_1 - f_0\|_2^4/1024}{\|g\|_2^2 + \|f_1 - f_0\|_2^2\|g\|_\infty/32}\right),$$

where $g = (f - f_1)^2 - (f_1 - f_0)^2/16$. We further compute

$$\begin{aligned}
\|g\|_2^2 &\le \left(\|(f - f_1)^2\|_2 + \frac{1}{16}\|(f_1 - f_0)^2\|_2\right)^2 \\
&\le \left(\|f - f_1\|_\infty\|f - f_1\|_2 + \frac{1}{16}\|f_1 - f_0\|_\infty\|f_1 - f_0\|_2\right)^2 \\
&\lesssim \|f - f_1\|_\infty^2\|f - f_1\|_2^2 + \|f_1 - f_0\|_\infty^2\|f_1 - f_0\|_2^2 \\
&\lesssim (m\|f_1 - f_0\|_2^2 + \delta^2)\|f_0 - f_1\|_2^2,
\end{aligned}$$

where we use (5.1), the fact that $\|f - f_1\|_2 \lesssim \|f_0 - f_1\|_2$, and that

$$\|f - f_1\|_\infty^2 \le 2\|f - f_0\|_\infty^2 + 2\|f_0 - f_1\|_\infty^2 \lesssim m\|f_1 - f_0\|_2^2 + \delta^2.$$

Similarly, we obtain on the other hand,

$$\|g\|_\infty = \|f - f_1\|_\infty^2 + \frac{1}{16}\|f_1 - f_0\|_\infty^2 \lesssim m\|f_0 - f_1\|_2^2 + \delta^2.$$

Therefore, we end up with

$$\mathbb{P}\left(\mathbb{P}_n(f - f_1)^2 > \frac{1}{16}\mathbb{P}_n(f_1 - f_0)^2\right) \le \exp\left(-\frac{\tilde{C}_2 n\|f_1 - f_0\|_2^2}{m\|f_1 - f_0\|_2^2 + \delta^2}\right),$$

where $\tilde{C}_2 > 0$ is some constant. Hence we obtain the following exponential bound for type I and type II error probabilities:

$$\begin{aligned}\mathbb{E}_0\phi_n &\le \exp(-Cn\|f_1 - f_0\|_2^2),\\ \mathbb{E}(1 - \phi_n) &\le \exp(-Cn\|f_1 - f_0\|_2^2) + 2\exp\left(-\frac{Cn\|f_1 - f_0\|_2^2}{m\|f_1 - f_0\|_2^2 + \delta^2}\right)\end{aligned}$$

for some constant $C > 0$ whenever $\|f - f_1\|_2^2 \le \|f_1 - f_0\|_2^2/32$. Taking the supremum of the type II error over $f \in \{f \in \mathcal{F}_m(\delta) : \|f - f_1\|_2^2 \le \|f_1 - f_0\|_2^2/32\}$ completes the proof. □

Proof of Lemma 2.2. We partition the alternative set into disjoint unions

$$\begin{aligned}&\{f \in \mathcal{F}_m(\delta) : \|f - f_0\|_2 > M\epsilon_n\}\\ &\quad\subset \bigcup_{j=M}^{\infty}\{f \in \mathcal{F}_m(\delta) : j\epsilon_n < \|f - f_0\|_2 \le (j+1)\epsilon_n\} := \bigcup_{j=M}^{\infty}\mathcal{S}_{nj}(\epsilon_n).\end{aligned}$$

For each $S_{nj}(\epsilon_n)$, we can find $N_{nj} = \mathcal{N}(\xi j\epsilon_n, \mathcal{S}_{nj}(\epsilon_n), \|\cdot\|_2)$-many functions $f_{njl} \in \mathcal{S}_{nj}(\epsilon_n)$ such that

$$\mathcal{S}_{nj}(\epsilon_n) \subset \bigcup_{l=1}^{N_{nj}}\{f \in \mathcal{F}_m(\delta) : \|f - f_{njl}\|_2 \le \xi j\epsilon_n\}.$$

Since for each f_{njl}, we have $f_{njl} \in \mathcal{S}_{nj}(\epsilon_n)$, implying that $\|f_{njl} - f_0\|_2 > j\epsilon_n$, we obtain the final decomposition of the alternative

$$\mathcal{S}_{nj}(\epsilon_n) \subset \bigcup_{l=1}^{N_{nj}}\{f \in \mathcal{F}_m(\delta) : \|f - f_{njl}\|_2 \le \xi\|f_0 - f_{njl}\|_2\}.$$

Now we apply Lemma 2.1 to construct individual test function ϕ_{njl} for each f_{njl} satisfying the following property

$$\begin{aligned}\mathbb{E}_0\phi_{njl} &\le \exp(-Cnj^2\epsilon_n^2),\\ \sup_{\{f\in\mathcal{F}_m(\delta):\|f-f_{njl}\|_2^2\le\xi^2\|f_0-f_{njl}\|_2^2\}}\mathbb{E}_f(1 - \phi_{njl}) &\le \exp(-Cnj^2\epsilon_n^2)\\ &\quad + 2\exp\left(-\frac{Cnj^2\epsilon_n^2}{mj^2\epsilon_n^2 + \delta^2}\right),\end{aligned}$$

where we have used the fact that $\|f_{njl} - f_0\|_2 > j\epsilon_n$. Now define the global test function to be $\phi_n = \sup_{j \geq M} \max_{1 \leq l \leq N_{nj}} \phi_{njl}$. Then the type I error probability can be upper bounded using the union bound

$$\mathbb{E}_0 \phi_n \leq \sum_{j=M}^{\infty} \sum_{l=1}^{N_{nj}} \mathbb{E}_0 \phi_{njl} \leq \sum_{j=M}^{\infty} \sum_{l=1}^{N_{nj}} \exp(-Cnj^2\epsilon_n^2) = \sum_{j=M}^{\infty} N_{nj} \exp(-Cnj^2\epsilon_n^2).$$

The type II error probability can also be upper bounded:

$$\begin{aligned}
&\sup_{\{f \in \mathcal{F}_m(\delta): \|f - f_0\|_2 > M\epsilon_n\}} \mathbb{E}_f(1 - \phi_n) \\
&\quad \leq \sup_{j \geq M} \sup_{l=1,\cdots,N_{nj}} \sup_{\{f \in \mathcal{F}_m(\delta): \|f - f_{njl}\|_2 \leq \xi \|f_0 - f_{njl}\|_2\}} \mathbb{E}_f(1 - \phi_{njl}) \\
&\quad \leq \sup_{j \geq M} \sup_{l=1,\cdots,N_{nj}} \left[\exp(-Cj^2 n\epsilon_n^2) + 2\exp\left(-\frac{Cnj^2\epsilon_n^2}{mj^2\epsilon_n^2 + \delta^2} \right) \right] \\
&\quad \leq \exp(-CM^2 n\epsilon_n^2) + 2\exp\left(-\frac{CnM^2\epsilon_n^2}{mM^2\epsilon_n^2 + \delta^2} \right).
\end{aligned}$$

The proof is thus completed. □

Proof of Lemma 2.3. Denote the re-normalized restriction of Π on $B_n = B(k_n, \epsilon_n, \omega)$ to be $\Pi(\cdot \mid B_n)$, and the random variables $(V_{ni})_{i=1}^n$, $(W_{ni})_{i=1}^n$ to be

$$V_{ni} = f_0(\mathbf{x}_i) - \int f(\mathbf{x}_i) \Pi(\mathrm{d}f \mid B_n), \quad W_{ni} = \frac{1}{2} \int (f(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2 \Pi(\mathrm{d}f \mid B_n).$$

Let

$$\mathcal{H}_n := \left\{ \int \prod_{i=1}^n \frac{p_f(\mathbf{x}_i, y_i)}{p_0(\mathbf{x}_i, y_i)} \Pi(\mathrm{d}f) > \Pi(B_n) \exp\left[-\left(C + \frac{1}{\sigma^2} \right) n\epsilon_n^2 \right] \right\}$$

Then by Jensen's inequality

$$\begin{aligned}
\mathcal{H}_n^c &\subset \left\{ \int \prod_{i=1}^n \frac{p_f(\mathbf{x}_i, y_i)}{p_0(\mathbf{x}_i, y_i)} \Pi(\mathrm{d}f \mid B_n) \leq \exp\left[-\left(C + \frac{1}{\sigma^2} \right) n\epsilon_n^2 \right] \right\} \\
&\subset \left\{ \frac{1}{\sigma^2} \sum_{i=1}^n (e_i V_{ni} + W_{ni}) \geq \left(C + \frac{1}{\sigma^2} \right) n\epsilon_n^2 \right\} \\
&\subset \left\{ \frac{1}{\sigma^2} \sum_{i=1}^n e_i V_{ni} \geq Cn\epsilon_n^2 \right\} \cup \left\{ \sum_{i=1}^n W_{ni} \geq n\epsilon_n^2 \right\}.
\end{aligned}$$

Now we use the Chernoff bound for Gaussian random variables to obtain the conditional probability bound for the first event given the design points $(\mathbf{x}_i)_{i=1}^n$:

$$\mathbb{P}_0 \left(\sum_{i=1}^n e_i V_{ni} \geq C\sigma^2 n\epsilon_n^2 \mid \mathbf{x}_1, \cdots, \mathbf{x}_n \right) \leq \exp\left(-\frac{C^2\sigma^4 n\epsilon_n^4}{\mathbb{P}_n V_{ni}^2} \right).$$

Since over the function class B_n, we have $\|f - f_0\|_2 \le \epsilon_n$, $k_n\epsilon_n^2 = O(1)$, and

$$\|f - f_0\|_\infty^2 \lesssim k_n\|f - f_0\|_2^2 + \omega^2 \le k_n\epsilon_n^2 + \omega^2 = O(1),$$

it follows from Fubini's theorem that

$$\begin{aligned}
\mathbb{E}(V_{ni}^2) &\le \int \|f_0 - f\|_2^2 \Pi(\mathrm{d}f \mid B_n) \le \epsilon_n^2, \\
\mathbb{E}\left(V_{ni}^4\right) &\le \mathbb{E}\left[\int (f_0(\mathbf{x}) - f(\mathbf{x}))^4 \Pi(\mathrm{d}f \mid B_n)\right] \\
&\le \int \|f - f_0\|_\infty^2 \|f - f_0\|_2^2 \Pi(\mathrm{d}f \mid B_n) \lesssim \epsilon_n^2.
\end{aligned}$$

Hence by the Chebyshev's inequality,

$$\mathbb{P}\left(\left|\mathbb{P}_n V_{ni}^2 - \mathbb{E}\left(V_{ni}^2\right)\right| > \epsilon_n^2\epsilon\right) \le \frac{1}{n\epsilon^2\epsilon_n^4}\mathrm{var}(V_{ni}^2) \le \frac{1}{n\epsilon_n^4\epsilon^2}\mathbb{E}(V_{ni}^4) \lesssim \frac{1}{n\epsilon_n^2} \to 0$$

for any $\epsilon > 0$, *i.e.*, $\mathbb{P}_n V_{ni}^2 \le \mathbb{E}V_{ni}^2 + o_P(\epsilon_n^2) \le \epsilon_n^2(1 + o_P(1))$, and hence,

$$\exp\left(-\frac{C^2\sigma^4 n\epsilon_n^4}{\mathbb{P}_n V_{ni}^2}\right) = \exp\left(-\frac{C^2\sigma^4 n\epsilon_n^2}{1 + o_P(1)}\right) \to 0$$

in probability. Therefore by the dominated convergence theorem the unconditioned probability goes to 0:

$$\begin{aligned}
\mathbb{P}_0\left(\sum_{i=1}^n e_i V_{ni} \ge C\sigma^2 n\epsilon_n^2\right) &= \mathbb{E}\left[\mathbb{P}_0\left(\sum_{i=1}^n e_i V_{ni} \ge C\sigma^2 n\epsilon_n^2 \mid \mathbf{x}_1, \cdots, \mathbf{x}_n\right)\right] \\
&\le \mathbb{E}\left[\exp\left(-\frac{C^2\sigma^4 n\epsilon_n^4}{\mathbb{P}_n V_{ni}^2}\right)\right] \to 0.
\end{aligned}$$

For the second event we use the Bernstein's inequality. Since

$$\begin{aligned}
\mathbb{E}W_{ni} &= \frac{1}{2}\int \|f - f_0\|_2^2 \Pi(\mathrm{d}f \mid B_n) \le \frac{1}{2}\epsilon_n^2, \\
\mathbb{E}W_{ni}^2 &\le \frac{1}{4}\mathbb{E}\left[\int (f(\mathbf{x}) - f_0(\mathbf{x}))^4 \Pi(\mathrm{d}f \mid B_n)\right] \\
&\le \frac{1}{4}\int \|f - f_0\|_2^2 \|f - f_0\|_\infty^2 \Pi(\mathrm{d}f \mid B_n) \lesssim \epsilon_n^2,
\end{aligned}$$

then

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^n W_{ni} > n\epsilon_n^2\right) &\le \mathbb{P}\left(\sum_{i=1}^n (W_{ni} - \mathbb{E}W_{ni}) > \frac{1}{2}n\epsilon_n^2\right) \\
&\le \exp\left(-\frac{1}{4}\frac{n\epsilon_n^4/4}{\mathbb{E}W_{ni}^2 + \epsilon_n^2\|W_{ni}\|_\infty/2}\right)
\end{aligned}$$

$$\leq \exp\left(-\frac{\hat{C}_1 n\epsilon_n^2}{1+\|W_{ni}\|_\infty}\right),$$

where $\|W_{ni}\|_\infty = \sup_{\mathbf{x}\in[0,1]^p}(1/2)\int(f(\mathbf{x})-f_0(\mathbf{x}))^2\Pi(\mathrm{d}f \mid B_n)$. Since for any $f \in B_n$, $\|f-f_0\|_\infty = O(1)$, it follows that $\|W_{ni}\|_\infty = O(1)$, and hence, $\mathbb{P}(\sum_i W_{ni} > n\epsilon_n^2) \to 0$. To sum up, we conclude that

$$\mathbb{P}(\mathcal{H}_n^c) \leq \mathbb{P}_0\left(\sum_{i=1}^n e_i V_{ni} \geq C\sigma^2 n\epsilon_n^2\right) + \mathbb{P}\left(\sum_{i=1}^n W_{ni} > n\epsilon_n^2\right) \to 0. \qquad \square$$

Proof of Lemma 2.4. Denote $\Pi(\cdot \mid B_n) = \Pi(\cdot \cap B_n)/\Pi(B_n)$ to be the re-normalized restriction of Π on $B_n = \{\|f-f_0\|_\infty < \epsilon_n\}$, and

$$V_{ni} = f_0(\mathbf{x}_i) - \int f(\mathbf{x}_i)\Pi(\mathrm{d}f \mid B_n), \quad W_{ni} = \frac{1}{2}\int (f(\mathbf{x}_i)-f_0(\mathbf{x}_i))^2\Pi(\mathrm{d}f \mid B_n).$$

Similar to the proof of Lemma 2.3, we obtain

$$\begin{aligned}
\mathcal{H}_n^c &= \left\{\int \prod_{i=1}^n \frac{p_f(\mathbf{x}_i,y_i)}{p_0(\mathbf{x}_i,y_i)}\Pi(\mathrm{d}f) \leq \Pi(B_n)\exp\left[-\left(C+\frac{1}{\sigma^2}\right)n\epsilon_n^2\right]\right\} \\
&\subset \left\{\frac{1}{\sigma^2}\sum_{i=1}^n (e_i V_{ni} + W_{ni}) \geq \left(C+\frac{1}{\sigma^2}\right)n\epsilon_n^2\right\} \\
&\subset \left\{\frac{1}{\sigma^2}\sum_{i=1}^n e_i V_{ni} \geq \left(C+\frac{1}{2\sigma^2}\right)n\epsilon_n^2\right\},
\end{aligned}$$

where we use the fact that $W_{ni} \leq (1/2)\int \|f-f_0\|_\infty^2 \Pi(\mathrm{d}f \mid B_n) \leq \epsilon_n^2/2$ for all $f \in B_n$ in the last step. Conditioning on the design points $(\mathbf{x}_i)_{i=1}^n$, we have

$$\begin{aligned}
&\mathbb{P}_0(\mathcal{H}_n^c \mid \mathbf{x}_1, \cdots, \mathbf{x}_n) \\
&\leq \exp\left[-\left(C+\frac{1}{2\sigma^2}\right)^2 \frac{\sigma^4 n\epsilon_n^4}{\mathbb{P}_n V_{ni}^2}\right] \\
&\leq \exp\left\{-\left(C+\frac{1}{2\sigma^2}\right)^2 \sigma^4 n\epsilon_n^4 \left[\int \|f-f_0\|_\infty^2 \Pi(\mathrm{d}f \mid B_n)\right]^{-1}\right\} \\
&\leq \exp\left[-\left(C+\frac{1}{2\sigma^2}\right)^2 \sigma^4 n\epsilon_n^2\right] \to 0.
\end{aligned}$$

The proof is completed by applying the dominated convergence theorem. $\square$

Proof of Theorem 2.1. For convenience denote the log-likelihood ratio function

$$\Lambda_n(f) = \sum_{i=1}^n [\log p_f(\mathbf{x}_i,y_i) - \log p_0(\mathbf{x}_i,y_i)].$$

Let ϕ_n be the test function given by Lemma 2.2 and

$$\mathcal{H}_n = \left\{ \int \exp(\Lambda_n(f \mid \mathcal{D}_n))\Pi(\mathrm{d}f) \geq \exp\left[-\left(\frac{3D}{2} + \frac{1}{\sigma^2} \right) n\underline{\epsilon}_n^2 \right] \right\}.$$

It follows from condition (2.5) that

$$\mathcal{H}_n^c \subset \left\{ \int \exp(\Lambda_n)\Pi(\mathrm{d}f) < \Pi(B_n(k_n, \underline{\epsilon}_n, \omega)) \exp\left[-\left(\frac{D}{2} + \frac{1}{\sigma^2} \right) n\underline{\epsilon}_n^2 \right] \right\},$$

and hence, $\mathbb{P}_0(\mathcal{H}_n^c) = o(1)$ by Lemma 2.3. Now we decompose the expected value of the posterior probability

$$\begin{aligned}
&\mathbb{E}_0 \left[\Pi \left(\|f - f_0\|_2 > M\epsilon_n \mid \mathcal{D}_n \right) \right] \\
&\quad \leq \mathbb{E}_0 \left[(1-\phi_n)\mathbb{1}(\mathcal{H}_n)\Pi(\|f - f_0\|_2 > M\epsilon_n \mid \mathcal{D}_n) \right] + \mathbb{E}_0\phi_n + \mathbb{E}_0[(1-\phi_n)\mathbb{1}(\mathcal{H}_n^c)] \\
&\quad \leq \mathbb{E}_0 \left[(1-\phi_n)\mathbb{1}(\mathcal{H}_n) \frac{\int_{\{\|f-f_0\|_2 > M\epsilon_n\}} \exp(\Lambda_n(f \mid \mathcal{D}_n))\Pi(\mathrm{d}f)}{\int \exp(\Lambda_n(f \mid \mathcal{D}_n))\Pi(\mathrm{d}f)} \right] \\
&\qquad + \mathbb{E}_0\phi_n + \mathbb{P}_0(\mathcal{H}_n^c).
\end{aligned}$$

By (2.3) and Lemma 2.2 the type I error probability $\mathbb{E}_0\phi_n \to 0$. It suffices to bound the first term. Observe that on the event $\mathcal{H}_n$, the denominator in the square bracket can be lower bounded:

$$\begin{aligned}
&\mathbb{E}_0 \left[(1-\phi_n)\mathbb{1}(\mathcal{H}_n) \frac{\int_{\{\|f-f_0\|_2 > M\epsilon_n\}} \exp(\Lambda_n(f \mid \mathcal{D}_n))\Pi(\mathrm{d}f)}{\int \exp(\Lambda_n(f \mid \mathcal{D}_n))\Pi(\mathrm{d}f)} \right] \\
&\quad \leq \exp\left[\left(\frac{3D}{2} + \frac{1}{\sigma^2} \right) n\underline{\epsilon}_n^2 \right] \\
&\qquad \times \mathbb{E}_0 \left[(1-\phi_n) \int_{\mathcal{F}_{m_n}(\delta) \cap \{\|f-f_0\|_2 > M\epsilon_n\}} \exp(\Lambda_n(f \mid \mathcal{D}_n))\Pi(\mathrm{d}f) \right] \\
&\qquad + \exp\left[\left(\frac{3D}{2} + \frac{1}{\sigma^2} \right) n\underline{\epsilon}_n^2 \right] \mathbb{E}_0 \left[\int_{\mathcal{F}_{m_n}^c(\delta)} \exp(\Lambda_n(f \mid \mathcal{D}_n))\Pi(\mathrm{d}f) \right].
\end{aligned}$$

By Fubini's theorem, Lemma 2.2 we have

$$\begin{aligned}
&\mathbb{E}_0 \left[(1-\phi_n) \int_{\mathcal{F}_{m_n}(\delta) \cap \{\|f-f_0\|_2 > M\epsilon_n\}} \exp(\Lambda_n(f \mid \mathcal{D}_n))\Pi(\mathrm{d}f) \right] \\
&\quad \leq \int_{\mathcal{F}_{m_n}(\delta) \cap \{\|f-f_0\|_2 > M\epsilon_n\}} \mathbb{E}_0 \left[(1-\phi_n) \exp(\Lambda_n \mid \mathcal{D}_n) \right] \Pi(\mathrm{d}f) \\
&\quad \leq \sup_{f \in \mathcal{F}_{m_n}(\delta) \cap \{\|f-f_0\|_2 > M\epsilon_n\}} \mathbb{E}_f(1-\phi_n) \\
&\quad \leq \exp(-CM^2 n\epsilon_n^2) + 2\exp\left(-\frac{CM^2 n\epsilon_n^2}{m_n\epsilon_n^2 M^2 + \delta} \right) \\
&\quad \leq \exp(-\tilde{C}M^2 n\epsilon_n^2),
\end{aligned}$$

for some constatn $\tilde{C} > 0$ for sufficiently large n, since $m_n\epsilon_n^2 \to 0$ and $\delta = O(1)$ by assumption. For the integral on $\mathcal{F}_{m_n}^c(\delta)$, we apply Fubini's theorem to obtain

$$\mathbb{E}_0\left[\int_{\mathcal{F}_{m_n}^c(\delta)} \exp(\Lambda_n(f \mid \mathcal{D}_n))\Pi(\mathrm{d}f)\right] = \int_{\mathcal{F}_{m_n}^c(\delta)} \mathbb{E}_0\left[\prod_{i=1}^n \frac{p_f(\mathbf{x}_i, y_i)}{p_0(\mathbf{x}_i, y_i)}\right]\Pi(\mathrm{d}f) \leq \Pi(\mathcal{F}_{m_n}^c(\delta)).$$

Hence we proceed to compute

$$\begin{aligned}&\mathbb{E}_0\left[(1-\phi_n)\mathbb{1}(\mathcal{H}_n)\frac{\int_{\{\|f-f_0\|_2 > M\epsilon_n\}} \exp(\Lambda_n(f \mid \mathcal{D}_n))\Pi(\mathrm{d}f)}{\int \exp(\Lambda_n(f \mid \mathcal{D}_n))\Pi(\mathrm{d}f)}\right] \\ &\lesssim \exp\left[\left(\frac{3D}{2} + \frac{1}{\sigma^2}\right)n\underline{\epsilon}_n^2 - \tilde{C}M^2 n\epsilon_n^2\right] \\ &\quad + \exp\left[\left(\frac{3D}{2} + \frac{1}{\sigma^2}\right)n\underline{\epsilon}_n^2 - \left(2D + \frac{1}{\sigma^2}\right)n\epsilon_n^2\right] \to 0\end{aligned}$$

as long as M is sufficiently large, where (2.4) is applied. □

Supplementary Material

Supplement to "A theoretical framework for Bayesian nonparametric regression"
(doi: 10.1214/19-EJS1616SUPP; .pdf). The supplementary material contains the remaining proofs and additional technical results

References

[1] Arbel, J., Gayraud, G. and Rousseau, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian Journal of Statistics* **40** 549–570. MR3091697

[2] Belitser, E. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *The Annals of Statistics* **31** 536–559. MR1983541

[3] Canale, A. and De Blasi, P. (2017). Posterior asymptotics of non-parametric location-scale mixtures for multivariate density estimation. *Bernoulli* **23** 379–404. MR3556776

[4] Castillo, I. and van der Vaart, A. (2012). Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077

[5] Choi, T. and Schervish, M. J. (2007). On posterior consistency in non-parametric regression problems. *Journal of Multivariate Analysis* **98** 1969–1987. MR2396949

[6] Cohen, A. (2003). *Numerical analysis of wavelet methods* **32**. Elsevier. MR1990555

[7] Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis* **1** 54–81. MR1256527
[8] Currin, C., Mitchell, T., Morris, M. and Ylvisaker, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association* **86** 953–963. MR1146343
[9] de Boor, C. (1978). A practical guide to splines. *Applied Mathematical Sciences, New York: Springer, 1978*. MR0507062
[10] De Jonge, R. and Van Zanten, J. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *The Annals of Statistics* **38** 3300–3320. MR2766853
[11] De Jonge, R. and Van Zanten, J. (2012). Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electronic Journal of Statistics* **6** 1984–2001. MR3020254
[12] Gao, C. and Zhou, H. H. (2016). Rate exact Bayesian adaptation with modified block priors. *The Annals of Statistics* **44** 318–345. MR3449770
[13] Ghosal, S., Ghosh, J. K. and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics* **28** 500–531. MR1790007
[14] Ghosal, S. and Van Der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics* **35** 697–723. MR2336864
[15] Ghosal, S. and Van Der Vaart, A. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics* **35** 192–223. MR2332274
[16] Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference* **44**. Cambridge University Press. MR3587782
[17] Ghosal, S. and Van Der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics* **29** 1233–1263. MR1873329
[18] Giné, E. and Nickl, R. (2015). *Mathematical foundations of infinite-dimensional statistical models* **40**. Cambridge University Press. MR3588285
[19] Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S* 249–307. Routledge.
[20] Hoffmann, M., Rousseau, J. and Schmidt-Hieber, J. (2015). On adaptive posterior concentration rates. *The Annals of Statistics* **43** 2259–2295. MR3396985
[21] Huang, T.-M. (2004). Convergence rates for posterior distributions and adaptive estimation. *The Annals of Statistics* **32** 1556–1593. MR2089134
[22] Koltchinskii, V. and Yuan, M. (2010). Sparsity in multiple kernel learning. *Ann. Statist.* **38** 3660–3695. MR2766864
[23] Kruijer, W., Rousseau, J. and Van Der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics* **4** 1225–1257. MR2735885

[24] Linero, A. R. and Yang, Y. (2017). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *arXiv preprint arXiv:1707.09461*. MR3874311

[25] Meier, L., van de Geer, S. and Bühlmann, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779–3821. MR2572443

[26] Pati, D., Bhattacharya, A. and Cheng, G. (2015). Optimal Bayesian estimation in random covariate design with a rescaled Gaussian process prior. *Journal of Machine Learning Research* **16** 2837–2851. MR3450525

[27] Pradeep, R., John, L., Han, L. and Larry, W. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 1009–1030. MR2750255

[28] Raskutti, G., Wainwright, M. J. and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research* **13** 389–427. MR2913704

[29] Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning* **1**. MIT Press, Cambridge. MR2514435

[30] Rivoirard, V. and Rousseau, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Analysis* **7** 311–334. MR2934953

[31] Rocková, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Ann. Statist.* **46** 401–437. MR3766957

[32] Rocková, V. and George, E. I. (2018). The Spike-and-Slab LASSO. *Journal of the American Statistical Association* **113** 431–444. MR3803476

[33] Rocková, V. and van der Pas, S. (2017). Posterior concentration for Bayesian regression trees and their ensembles. *arXiv preprint arXiv:1708.08734*.

[34] Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science* 409–423. MR1041765

[35] Schumaker, L. (2007). *Spline functions: Basic theory*. Cambridge University Press. MR2348176

[36] Scricciolo, C. (2006). Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *The Annals of Statistics* **34** 2897–2920. MR2329472

[37] Scricciolo, C. (2011). Posterior rates of convergence for Dirichlet mixtures of exponential power densities. *Electronic Journal of Statistics* **5** 270–308. MR2802044

[38] Shen, W., Tokdar, S. T. and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. MR3094441

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

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

[41] Vaart, A. v. d. and Zanten, H. v. (2011). Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research* **12** 2095–2119. MR2819028

[42] Van der Vaart, A. W. (2000). *Asymptotic statistics* **3**. Cambridge University Press. MR1652247

[43] van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics* 1435–1463. MR2418663

[44] van der Vaart, A. W. and van Zanten, J. H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh* 200–222. Institute of Mathematical Statistics. MR2459226

[45] 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* 2655–2675. MR2541442

[46] Weining, S. and Subhashis, G. Adaptive Bayesian procedures using random series priors. *Scandinavian Journal of Statistics* **42** 1194–1213. MR3426318

[47] Xie, F. and Xu, Y. (2019). Bayesian repulsive Gaussian mixture model. *Journal of the American Statistical Association* **0**. 1–29.

[48] Xie, F. and Xu, Y. (2018). Adaptive Bayesian nonparametric regression using a kernel mixture of polynomials with application to partial linear models. *Bayesian Anal.* Advance publication.

[49] Xie, F. Jin, W. and Xu, Y. (2019). Supplement to "Rates of contraction with respect to L_2-distance for Bayesian nonparametric regression". DOI: 10.1214/19-EJS1616SUPP

[50] Yang, Y., Bhattacharya, A. and Pati, D. (2017). Frequentist coverage and sup-norm convergence rate in Gaussian process regression. *arXiv preprint arXiv:1708.04753*.

[51] Yang, Y. and Tokdar, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43** 652–674. MR3319139

[52] Yoo, W. W. and Ghosal, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *The Annals of Statistics* **44** 1069–1102. MR3485954

[53] Yoo, W. W., Rousseau, J. and Rivoirard, V. (2017). Adaptive supremum norm posterior contraction: spike-and-slab priors and anisotropic Besov spaces. *arXiv preprint arXiv:1708.01909*.