On the Convergence of Bayesian Regression Models

Yuao Hu
Division of Mathematical Sciences, SPMS
Nanyang Technological University
Singapore 637371

October 7, 2010

Abstract

We consider heteroscedastic nonparametric regression models, when both the mean function and variance function are unknown and to be estimated with nonparametric approaches. We derive convergence rates of posterior distributions for this model with different priors, including splines and Gaussian process priors. The results are based on the general ones on the rates of convergence of posterior distributions for independent, non-identically distributed observations, and are established for both of the cases with random covariates, and deterministic covariates. We also illustrate that the results can be achieved for all levels of regularity, which means they are adaptive.

1 Introduction

The posterior distribution is said to be consistent if the posterior probability of any small neighborhood of the true parameter value converges to one. In recent years, many results, giving condition, under which the posterior distribution is consistent have appeared, especially under the situation that the parameter spaces are in finite-dimensional. For example, Barron et al [1] gave necessary and sufficient conditions for the posterior consistency , and results were then specialized to weak and L_1 neighborhoods from Kullback-Leibler neighborhoods. For details, we refer the reader to [1]; [2]; The consistency of posterior distributions in nonparametric Bayesian inference has received quite a lot of attention ever since 1986, when Diaconis and Freedman gave counterexample to argue that Bayesian methods sometimes can not work. On the positive side, consistency has been demonstrated on many models [3, 4, 5, 6, 7, 8, 9, 10, 11]

In nonparametric Bayesian analysis, we have an independent sample $Y_1, \cdots, Y_n$ from a distribution P_0 with density p_0 with respect to some measure on the sample space $(\mathcal{Y}, \mathcal{B})$. The model space is denoted by $\mathcal{P}$ which is known to contain the true distribution P_0. Given some prior distribution Π on $\mathcal{P}$, the posterior is a random measure given by

$$\Pi_n(A|Y_1, \cdots, Y_n) = \frac{\int_A \Pi_{i=1}^n p(Y_i) d\Pi(P)}{\int \Pi_{i=1}^n p(Y_i) d\Pi(P)}.$$

For ease of notation, we will omit the explicit conditioning and write $\Pi(A)$ for the posterior distribution. We say that the posterior is consistent if

$$\Pi_n(P : d(P, P_0) > \epsilon) \to 0 \; in \; P_0^n \; probability,$$

for any $\epsilon > 0$, where d is some suitable distance function between probability measures.

Furthermore, issues of rates of convergence are of interests on. We say the rate is at least ϵ_n if for a sufficiently large constant M

$$\Pi_n(P : d(P, P_0) > M\epsilon_n) \to 0 \; in \; P_0^n \; probability,$$

where ϵ_n is a positive sequence decreasing to zero. Ghosal and van der Vaart [12]; presented general results on the rates of convergence of the posterior measure , and [13] then generalized the results to case even the observations are not i.i.d, which is useful for the model considered in this article.

For Bayesian nonparametric regression models, one of the common approaches is through the splines basis expansion for regression functions, Ghosal and van der Vaart [12] gave the posterior consistency rate for regression model with unknown mean function and normal distributed error variable with zero means and known variances σ^2, using this approach. T.Choi and M. Schervish[14] provided sufficient conditions for posterior consistency in nonparametric regression problems with homogenous Gaussian errors with unknown level by constructing tests that separate from the outside of the suitable neighborhoods of the parameter. Amewou-Atisso, Ghosal, Ghosh and Ramamoorthi [15] presented a posterior consistency analysis for linear regression problems with an unknown error distribution which is symmetric about zero. Besides, both papers did not consider the rates of convergence.

In this paper, we give the convergence rates for heteroscedastic nonparametric regression models, when we use nonparametric methods to estimate the unknown variance function and the unknown mean function simultaneously. Besides, as in [2], we also deal with two types of covariates either randomly sampled from a probability distribution or fixed in advance. When the covariate values in one-dimensional, we use the approach of splines basis expansion for regression functions and give the convergence rate. For high dimensional cases, we use rescaled smooth Gaussian fields as priors for multidimensional functions to get the result.

Using Gaussian process in the context of density estimation is another common approach in Bayesian nonparametric analysis. It is first used by Leonard[16] and Lenk [17]. Recently, many results on posterior consistency are induced by the Gaussian process prior, such as in [16], and [18]. Van der Vaart and van Zanten [19] derived the rates of contraction of posterior distributions on nonparametric or semiparametric models based on Gaussian processes and showed that the rates depend on the position of the true parameter associated with the reproducing kernel Hilbert space of the Gaussian process and the small ball probabilities of the Gaussian process. With rescaled smooth Gaussian fields as priors, they[20] extended the results to be fully adaptive to the smoothness.

The rest of the paper is organized as follows. In Section 2, we describe the regression model. In Section 3, we give the main results of the posterior convergence rates with splines basis expansion approach and Gaussian process approach. Section 4 contains proofs with some lemma left to the Appendix. We discuss about the results and some directions on future work in section 5.

2 The model

We consider the heteoscedastic nonparametric regression model, where a random response y corresponding to a covariate vector $\mathbf{x}$ taking values in a compact set $T \subset R^d$, without loss of generality, we assume that $T = [0, 1]^d$.

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

$$\begin{aligned}
&y_i = \eta(\mathbf{x_i}) + V^{1/2}(\mathbf{x_i})\epsilon_i, \\
&\epsilon_i \sim N(0,1), \\
&\eta(.) \sim \Pi_1, \\
&f(.) = \log V(.) \sim \Pi_2,
\end{aligned} \tag{1}$$

when $\eta(.)$ is the mean function, and $V(.)$ is the variance function. Let $\Theta^{(g)}$ be the abstract measure space where the function $g(x)$ ($g(x)$ indicates $\eta(x)$, or $f(x)$) belongs to, with respect to a common σ-finite measure. With the assumption $\Theta^{(\eta)}$ and $\Theta^{(f)}$ are independent, we define the jointly parameter space Θ the product space of $\Theta^{(\eta)}$ and $\Theta^{(f)}$. Since the parameter space is infinite-dimensional, we consider a sieve Θ_n growing eventually to the space of Θ,with $\Theta_1 \subseteq \cdots, \subseteq \Theta_n \subseteq \Theta$ and $\cup\Theta_n = \Theta$. We model the unknown function $\eta(.)$ and $f(.)$ with suitable prior distributions $\Pi_n^{(\eta)}$ and $\Pi_n^{(f)}$ on their parameter sieve spaces, respectively.

3 The main results

In this section, we give the rates of convergence of the nonparametric regression model described in Section 1, The parameter is $\theta = (\eta, V)$ with $\theta_0 = (\eta_0, V_0)$ being the true functions.

Let $\mathcal{P}_{\eta,V}$ be the distribution of y. To be specific, for our model

$$\mathcal{P}_{\eta,V}(y|x) = \frac{1}{\sqrt{2\pi V(\mathbf{x})}} exp(-\frac{(y-\eta(\mathbf{x}))^2}{2V(\mathbf{x})}). \tag{2}$$

We use d_n^2 to denote the squares of the Hellinger distances. It means, for random covariates and fixed covariates

$$d_n^2(\mathcal{P}_{\eta_1,V_1}, \mathcal{P}_{\eta_2,V_2}) = \int\int (\mathcal{P}_{\eta_1,V_1}^{\frac{1}{2}} - \mathcal{P}_{\eta_2,V_2}^{\frac{1}{2}})^2 \; dy \; dQ(\mathbf{x}). \tag{3}$$

For random covariates, $Q(\mathbf{x})$ denotes the distribution function of $\mathbf{x}$, and for fixed covariates, it is the empirical probability measure of the design points, which is defined by$P_n^{\mathbf{x}} = n^{-1}\sum_{i=1}^n \delta_{\mathbf{x_i}}$.

The Kullback-Leibler divergence and variance divergence of P_{η_1,V_1} and P_{η_2,V_2} for fixed $\mathbf{x}$ are defined in the following way:

$$\begin{aligned} K_{\mathbf{x}}(\mathcal{P}_{\eta_1,V_1},\mathcal{P}_{\eta_2,V_2}) &= \int \mathcal{P}_{\eta_1,V_1}\ \log(\frac{\mathcal{P}_{\eta_1,V_1}}{\mathcal{P}_{\eta_2,V_2}})\ dy; \\ Var_{\mathbf{x}}(\mathcal{P}_{\eta_1,V_1},\mathcal{P}_{\eta_2,V_2}) &= \int \mathcal{P}_{\eta_1,V_1}\ (\log(\frac{\mathcal{P}_{\eta_1,V_1}}{\mathcal{P}_{\eta_2,V_2}}) - K_{\mathbf{x}}(\mathcal{P}_{\eta_1,V_1},\mathcal{P}_{\eta_2,V_2}))^2\ dy. \end{aligned} \tag{4}$$

For the specific model in section 2:

$$\begin{aligned} K_{\mathbf{x}}(\mathcal{P}_{\eta_1,V_1},\mathcal{P}_{\eta_2,V_2}) &= \frac{1}{2}\log\frac{V_2(\mathbf{x})}{V_1(\mathbf{x})} - \frac{1}{2}(1-\frac{V_1(\mathbf{x})}{V_2(\mathbf{x})}) + \frac{1}{2}\frac{[\eta_1(\mathbf{x})-\eta_2(\mathbf{x})]^2}{V_2(\mathbf{x})}; \\ Var_{\mathbf{x}}(\mathcal{P}_{\eta_1,V_1},\mathcal{P}_{\eta_2,V_2}) &= 2[-\frac{1}{2}+\frac{1}{2}\frac{V_1(\mathbf{x})}{V_2(\mathbf{x})}]^2 + [\frac{V_1(\mathbf{x})}{V_2(\mathbf{x}))}[\eta_1(\mathbf{x})-\eta_2(\mathbf{x})]]^2. \end{aligned} \tag{5}$$

Correspondently, the average Kullback-Leibler divergence and variance divergence are in the forms of

$$\begin{aligned} K(\mathcal{P}_{\eta_1,V_1},\mathcal{P}_{\eta_2,V_2}) &= \int K_{\mathbf{x}}(\mathcal{P}_{\eta_1,V_1},\mathcal{P}_{\eta_2,V_2})\ dQ(\mathbf{x}); \\ Var(\mathcal{P}_{\eta_1,V_1},\mathcal{P}_{\eta_2,V_2}) &= \int Var_{\mathbf{x}}(\mathcal{P}_{\eta_1,V_1},\mathcal{P}_{\eta_2,V_2})\ dQ(\mathbf{x}). \end{aligned} \tag{6}$$

In the remainder of the article, let $||.||_n$ stand for the norm on $L_2(Q)$, $||.||_\infty$ denotes the supreme norm .

3.1 Splines

In this section we give the convergence rates to prior distributions on spline models for regression functions. We restrict ourselves to the one-dimensional case here, though for higher dimensions case, tensor splines can be used.

The basic assumption for the true densities of the mean function and variance function is that they belong to the Hölder spaces $C^\alpha[0,1]$ and $C^\gamma[0,1]$, respectively, where $\alpha,\gamma > 0$ could be fractional. The Hölder space $C^\alpha[0,1]$ is constructed by all functions that have α_0 derivatives, with α_0 being the greatest integer less than α and α_0th derivative being Lipschitz of order $\alpha-\alpha_0$.

Throughout this article, we fix an order q, which is a natural number satisfied $q \geq max\{\alpha,\gamma\}$. A B-spline basis function of order q consists of q

polynomial pieces of degree $q-1$, which are $q-2$ times continuously differentiable throughout [0,1]. To approximate a function on [0,1], we partition the interval [0,1] into K_n subintervals $((k-1)/K_n, k/K_n]$ for $k = 1, 2, \cdots, K_n$, with $\{K_n\}$ being a sequence of natural numbers increasing to infinity as n goes to infinity. Each subinterval $((k-1)/K_n, k/K_n]$ is approximated by a polynomials of degree strictly less than q. The number of basis functions needed is $J_n = (q + K_n - 1)$. The basis functions can be denoted as B_j, with $j = 1, 2, \cdots J_n$. Thus, the space of splines of order q is a J_n-dimensional linear space, consisted by all functions from $[0,1]$ to $\mathbb{R}$ in form of $g = \sum_{j=1}^{J^n} \beta_j B_j$. As in [13], the B-splines satisfy (i) $B_j \geq 0$, $j = 1, 2, \cdots J_n$, (ii) $\sum_{j=1}^{J_n} B_j = 1$, (iii) B_j is supported inside an interval of length q/K_n and (iv) at most q of $B_1, B_2, \cdots, B_{J_n}$ are nonzero at any given x.

We denote g= f or η, and put prior on g by a prior on $\beta = (\beta_1, \cdots, \beta_{J_n})^T$, the spline coefficients, where g is represented as $g_\beta(x) = \beta^T B(x)$. Let $\Pi_n^{(g)}$ be priors induced by a multivariate normal distribution $N_{J_n}(0, I)$ on the spline coefficients.

We also assume the regressors are sufficiently regularly distributed, by satisfying the condition expressed in the following term

$$J_n^{-1}||\beta||^2 \lesssim \beta^T \Sigma \beta \lesssim J_n^{-1}||\beta||^2, \tag{7}$$

where $\Sigma = (\int B_i B_j \; d\, Q)$, $||.||$ is the Euclidean norm on $\mathbb{R}^{J_n}$.

Theorem 1. *Assume that* $\eta_0 \in C^\alpha[0,1], V_0 \in C^\gamma[0,1]$ *for some* $\alpha, \gamma \geq \frac{1}{2}$, V_0 *is away from 0, and (7) holds. Let* $\Pi_n^{(\eta)}$ *and* $\Pi_n^{(f)}$ *be priors of* η *and* f *both induced by* $N_{J_n}(0, I)$ *on the spline coefficients. If*

$$J_n \sim \min\{(n/\log n)^{1/(1+2\alpha)}, n^{1/(2+2\gamma)}\},$$

then the posterior converges at the rate

$$\epsilon_n \sim \max\{(n/\log n)^{-\alpha/(1+2\alpha)}, n^{-\gamma/(2+2\gamma)}\},$$

relative to d_n.

Usually, we can view J_n to be a sequence of random variables with a prior distributions. It can be prove that the posterior can convergence at the same rate.

Corollary 1. *Assume that* $\eta_0 \in C^\alpha[0,1], V_0 \in C^\gamma[0,1]$ *for some* $\alpha, \gamma \geq \frac{1}{2}$, V_0 *is away from 0, and (7) holds. Let* $\Pi_n^{(\eta)}$ *and* $\Pi_n^{(f)}$ *be priors of* η *and* f *both induced by* $N_{J_n}(0,I)$ *on the spline coefficients.* J_n *is a sequence of geometric distributed random variables with successful probability* p_n *satisfying* $p_n^{k_n-1}(1-p_n) = e^{-n\epsilon_n^2}$, *with* $k_n = \lfloor \min\{(n/\log n)^{1/(2\alpha+1)}, n^{1/(2+2\gamma)}\}\rfloor$, $\epsilon_n \sim \max\{(n/\log n)^{-\alpha/(1+2\alpha)}, n^{-\gamma/(2+2\gamma)}\}$.*Then, the posterior convergence rate is* ϵ_n, *relative to* d_n.

3.2 Gaussian process prior

For higher dimensional case, we employ prior distributions, constructed by rescaling smooth Gaussian random field. Let Θ be $C[0,1]^d$, the space of all continuous functions defined on $[0,1]^d$. As in [20], we set $W^{(g)} = (W_{\mathbf{x}}^{(g)} : \mathbf{x} \in \mathbb{R}^d)$ to be a centered, homogeneous Gaussian random field with covariance function of the form, for a given continuous function ϕ:

$$EW_s^{(g)}W_t^{(g)} = \phi(s-t).$$

To be specific, we choose $W^{(g)} = (W_x^{(g)} : \mathbf{x} \in \mathbb{R}^d)$ to be the squared exponential process, which is the centered Gaussian process with covariance function

$$EW_s^{(g)}W_t^{(g)} = exp(-||s-t)||^2),$$

where $||.||$ is the Euclidean norm on $\mathbb{R}^d$.

Let A be a random variable defined on the same probability space as $W^{(g)}$ and independent of $W^{(g)}$. Here we assume A^d possesses a Gamma distribution. $W^{(g)A}$ is used to denote the rescaled process $x \to W_{Ax}$ restricted on $[0,1]^d$, which can be considered as a Borel measurable map in the space $C[0,1]^d$, with the uniform norm $||.||_\infty$, as showed in [20].

Theorem 2. *Assume that* $\eta_0 \in C^\alpha[0,1]^d, V_0 \in C^\gamma[0,1]^d$ *for some* $\alpha, \gamma \geq \frac{1}{2}$, V_0 *is away from 0. We consider the prior on* g *is (*g *denotes* f *or* η*)* $W^{(g)A}$, *which is the restricted and rescaled squared exponential process with* A^d *a Gamma distributed random variable. Then, the posterior converges at the rate*

$$\epsilon_n = \max\{n^{-\alpha/(d+2\alpha)}(\log n)^{(d+1)\alpha/(2\alpha+d)}, n^{-\gamma/(d+2\gamma)}(\log n)^{(d+1)\gamma/(2\gamma+d)}\}$$

relative to d_n.

The proof can be found in Section 4. Also, this rate of contraction is not minimax. By choosing a different prior for A, the power $(d+1)\alpha/(2\alpha+d)$ of the logarithmic factor can be improved. Though the prior does not depend on α and γ, the convergence rate is true for any level of α, and γ. In this sense, it is rate-adaptive.

If we do not consider about the property of adaption or the regularity levels are known, we can find the minimax rate by using proper priors.

Corollary 2. *Assume that $\eta_0 \in C^\alpha[0,1], V_0 \in C^\gamma[0,1]$ for some $\alpha, \gamma \geq \frac{1}{2}$, V_0 is away from 0. For simplicity, we only consider the one-dimensional situation for simplicity. We denote $W^{(g)}$ to be a standard Brownian motion and $Z_0, \cdots Z_{k_g}$ independent standard normal random variables. We consider the prior on g is the process $x \to I_{0+}^{k_g} W_x^{(g)} + \sum_{i=1}^{k_g} Z_i x^i/i!$, where $I_{0+}W$ denotes $x \to \int_0^x W(x)dx$, and $I_{0+}^k W$ denotes $I_{0+}^1(I_{0+}^{k-1}W)$. Then, the posterior converges at the rate*

$$\max\{n^{-\alpha/(2k_\eta+2)}, n^{-\gamma/2k_f+2}\}$$

When $\gamma = k_f + 1/2$ and $\alpha = k_\eta + 1/2$,

$$\epsilon_n = \max\{n^{-\alpha/(1+2\alpha)}, n^{-\gamma/(1+2\gamma)}\}$$

which is the minimax rate.

This example shows, for the case α, and γ are known, we can use the above specific Gaussian process prior to get the minimax rate. However, this is not optimal for all level of α and γ, so other choice of k_g will corresponds to under-or-over-smoothed prior.

4 The proofs for the main results

In preparation for the proofs of the main results, we first collect some lemmas, which are used to bound the average hellinger distance entropy, Kullback-Leibler divergence and variance divergence with the L_2 norm of the regression functions.

Lemma 1. *The average hellinger distance entropy of the product space* Θ_n *can be bounded by a multiple of the summation of* $||.||_n$*-entropy of* $\Theta_n^{(\eta)}$ *and* $\Theta_n^{(f)}$*, reminding that* $f = \log V$*, which means*

$$\log N(3\epsilon, \Theta_n, d_n) \lesssim \log N(\epsilon/e^{N_n}, \Theta_n^{(\eta)}, ||.||_n) + \log N(\epsilon, \Theta_n^{(f)}, ||.||_n). \quad (8)$$

With this lemma, the ϵ-covering number relative to d_n-metric can be estimated that with relative to L_2-metric.

Lemma 2. *Under the assumption that both* f_1 *and* f_2 *are uniformly bounded by a constant N,*

$$\begin{aligned} &K(\mathcal{P}_{\eta_1,V_1}, \mathcal{P}_{\eta_2,V_2}) \leq (1+e^{2N})(||\eta_1-\eta_2||_n^2 + ||(f_1-f_2)||_n^2); \\ &Var(\mathcal{P}_{\eta_1,V_1}, \mathcal{P}_{\eta_2,V_2}) \leq e^{4N}(||\eta-\eta_0||_n^2 + ||f-f_0||_n^2). \end{aligned} \quad (9)$$

We use this lemma to estimate the prior concentration probability. The proofs can be found in the Appendix.

4.1 Proof for theorem 1

We consider sieve $\Theta_n = \Theta_n^{(f)} \times \Theta_n^{(\eta)}$ where

$$\Theta_n^{(f)} = \{f_\beta \in supp\{\Pi_n^{(f)}\}, ||f_\beta|| \leq N_n\};$$

$$\Theta_n^{(\eta)} = \{\eta_\beta \in supp\{\Pi_n^{(\eta)}\}, ||\eta_\beta|| \leq M_n\},$$

where $supp\{\Pi_n\}$ means the support of Π_n and M_n, N_n are sequence of real numbers goes to infinity as n goes to infinity. Since we suppose $\eta_0 \in C^\alpha[0,1], V_0 \in C^\gamma[0,1]$, and V_0 is away from 0, we have $\log V_0 \in C^\gamma[0,1]$, too. By the Lemma 4.1 in [12], there exists some $\beta_{\eta_0}, \beta_{f_0} \in \mathbb{R}^{J_n}$(dependent on n), for the true density of f_0 and η_0, the basic approximation property of splines are satisfied as

$$\begin{aligned} &||\beta_{f_0}^T B - f_0||_\infty \leq A J_n^{-\gamma} ||f_0||_\gamma; \\ &||\beta_{\eta_0}^T B - \eta_0||_\infty \leq A' J_n^{-\alpha} ||\eta_0||_\alpha, \end{aligned} \quad (10)$$

where A, and A' are constant.

Under the assumption of (7) in Theorem 1, we can use Euclidean norms on the spline coefficients to control the L_2 distance of functions, since for all $\beta, \beta^{'} \in \mathbb{R}^{J_n}$,

$$C^{-1}||\beta - \beta^{'}|| \leq \sqrt{J}||g_\beta - g_{\beta'}||_n \leq (C^{'})^{-1}||\beta - \beta^{'}|| \qquad (11)$$

are satisfied for some constants C and $C^{'}$.

We verify all the conditions of general results on rates of posterior contraction (e.g. Theorem 4 of [13]), except that the local entropy in condition (3.2) is replaced by the global entropy $\log N(\epsilon, \Theta_n, d_n)$ without affection rates. The parameter θ in Theorem 4 of [13] is (η, V) with $\theta_0 = (\eta_0, V_0)$.

We start from the estimation of entropy number. We project g_0 onto the J_n-dimensional space of splines and denote the projection function $g_{\beta_g^{(n)}}$. Using the property of projection combined with (11), we have that $\{\beta : ||g_\beta - g_0||_n \leq \epsilon\} \subset \{\beta : ||\beta - \beta_g^{(n)}|| \leq C\sqrt{J_n}\epsilon\}$ for every $\epsilon > 0$. For details, please refer to [13]. Thus, we can use the $C\sqrt{J_n}\epsilon$-covering numbers relative to Euclidean norm to bound the ϵ-covering number of the set $\{\beta : ||g_\beta - g_0||_n\}$ relative to L_2 norm. Thus, we have

$$N(\epsilon/3, \Theta_n^{(\eta)}, ||.||_n) \lesssim N(C\sqrt{J_n}\epsilon, \Theta_n^{(\eta)}, ||.||) \lesssim (\frac{KM_n}{\epsilon_n})^{J_n}, \qquad (12)$$

where K is a constant, η can be replaced by f with M_n replaced by N_n together. So by lemma 1, the entropy condition $\log N(\epsilon, \Theta_n, d_n) \lesssim n\epsilon^2$ is satisfied, provided $J_n \log M_n \lesssim n\epsilon_n^2$, $J_n N_n \lesssim n\epsilon_n^2$ and $J_n \log \epsilon_n^{-1} \lesssim n\epsilon_n^2$.

Then, we turn to estimate the prior concentration probability for the true density, which is in form of

$$\Pi_n(B_n((\eta_0, f_0), \epsilon_n; 2)) = \left\{(\eta, V) : K(\mathcal{P}_{\eta,V}, \mathcal{P}_{\eta_0,V_0}) \leq \epsilon^2, Var(\mathcal{P}_{\eta,V}, \mathcal{P}_{\eta_0,V_0}) \leq \epsilon^2\right\}. \qquad (13)$$

We denote $N/2$ to be $||f_0||_\infty$. Under the assumption that $||f||_\infty \leq N$ and

$||f_0||_\infty \leq N$, when n is sufficiently large,

$$
\begin{aligned}
&\Pi_n(B_n((\eta_0, f_0), \epsilon_n; 2)) \\
&\geq \left\{(\eta, V) : K(P_{\eta,V}, P_{\eta_0,V_0}) \leq \epsilon^2, Var(P_{\eta,V}, P_{\eta_0,V_0}) \leq \epsilon^2, ||f||_\infty < N\right\} \\
&\geq \Pi_n(||f - f_0||_n^2 + ||\eta - \eta_0||_n^2 \leq e^{-4N}\epsilon_n^2, ||f||_\infty < N) \\
&\geq \Pi_n^{(f)}(f : ||f - f_0||_n^2 \leq \frac{e^{-4N}}{2}\epsilon_n^2, ||f||_\infty < N) \times \Pi_n^{(\eta)}(\eta : ||\eta - \eta_0||_n^2 \leq \frac{e^{-4N}}{2}\epsilon_n^2) \\
&\geq \Pr_{\beta^T B \in \Theta_n^{(f)}}(\beta : ||\beta - \beta_f^{(n)}|| \leq e^{-2N}C'\sqrt{J_n}\epsilon_n, |\beta_j^{(n)}| < N) \\
&\times \Pr_{\beta^T B \in \Theta_n^{(\eta)}}(\beta : ||\beta - \beta_\eta^{(n)}|| \leq e^{-2N}C'\sqrt{J_n}\epsilon_n, |\beta_j^{(n)}| < N) \\
&\geq (\inf_{\beta_1 \in [-2N, 2N]} \phi(\beta_1))^2 Vol(\beta : ||\beta - \beta_f^{(n)}|| \leq e^{-2N}C'\epsilon_n) Vol(\beta : ||\beta - \beta_\eta^{(n)}|| \leq e^{-2N}C'\epsilon_n) \\
&\gtrsim \epsilon_n^{2J_n}
\end{aligned}
\tag{14}
$$

where vol denotes the volume in Euclidean space and $\inf_{\beta_1 \in [-2N, 2N]} \phi(\beta_1)$ represents the infimum value of density function ϕ, which is the density function of normal distribution, constrained on the open set $[-2N, 2N]$. The second inequality is derived from lemma 2. $\inf_{\beta_1 \in (-2N, 2N)} \phi(\beta_1)$ is a real number away from zero, which can be derived from the facts that ϕ is nonzero at any point belongs to $\mathbb{R}$ alone with its continuity, and [-2N, 2N] is a compact set in $\mathbb{R}$.

To satisfy the entropy and the prior concentration conditions, it is necessary that $J_n N_n \lesssim n\epsilon_n^2$, $J_n \log M_n \lesssim n\epsilon_n^2$, and $J_n \log \epsilon_n^{-1} \lesssim n\epsilon_n^2$ together with $\epsilon_n \gtrsim 2J_n^{-\nu}$, where $\nu = \min\{\alpha, \gamma\}$. When we set $N_n \sim n^{1/(2\nu+2)}$, $M_n \sim n$, all conditions of above are satisfied, with

$$J_n \sim \min\{(n/\log n)^{1/(1+2\alpha)}, n^{1/(2+2\gamma)}\},$$

and

$$\epsilon_n \sim \max\{(n/\log n)^{-\alpha/(1+2\alpha)}, n^{-\gamma/(2+2\gamma)}\}.$$

The left is to get the condition on which the probability assigned by prior to Θ_n complement is exponentially small. As we mentioned, $\eta_\beta = \beta^T B(x)$ for all $x \in [0, 1]^d$, and $|\sum_j^{J_n} \beta_j B_j| \leq \max_{j=1}^{J_n} |\beta_j|$. Then for $t_n > 0$, by Markov's inequality and Chernoff Bounds, we have

$$\Pr\Big\{\sup_{x\in[0,1]}|\sum_{j}^{J_n}\beta_j B_j| > M_n\Big\} \leq J_n \exp\Big(-t_n M_n + \frac{1}{2}t_n^2\Big)2\Phi(t_n), \tag{15}$$

where Φ is the standard normal distribution function. By taking $t_n = M_n$, we have

$$\Pr\Big\{\sup_{x\in[0,1]}|\sum_{j}^{J_n}\beta_j B_j| > M_n\Big\} \lesssim J_n exp\Big(-\frac{M_n^2}{2}\Big). \tag{16}$$

With the M_n, N_n, J_n and ϵ_n defined as above, and n sufficiently large,

$$J_n \exp\Big(-\frac{M_n^2}{2}\Big) \lesssim \exp\Big(-n\epsilon_n^2\Big), \tag{17}$$

and the formula replacing M_n with N_n are also satisfied. Thus,

$$\begin{aligned}&\Pi_n(\Theta\setminus\Theta_n) \leq \Pi_n^{(f)}(\Theta^{(f)}\setminus\Theta_n^{(f)}) + \Pi_n^{(\eta)}(\Theta^{(\eta)}\setminus\Theta_n^{(\eta)})\\ &= \Pr\Big\{\sup_{x\in[0,1]}|\sum_{j}^{J_n}\beta_j B_j| > M_n\Big\} + \Pr\Big\{\sup_{x\in[0,1]}|\sum_{j}^{J_n}\beta_j B_j| > N_n\Big\}\\ &\lesssim \exp\Big(-n\epsilon_n^2\Big).\end{aligned} \tag{18}$$

The whole proof is completed.

Remark 1. When we generalize the priors of η and f, which are induced by the spline coefficients, with some limitation, the convergence rate will stay unchanged. We assume the same prior Π on each $\beta_j \in \mathbb{R}$, $j = 1, \cdots, J_n$, with density function $d(\beta_j) \in C[\mathbb{R}]$ (the set of continuous functions), which satisfies

$$\begin{aligned}&\Pi(|\beta_j| > M) \lesssim e^{-M^\rho};\\ &d(\beta_j = r) \neq 0 \ for\ any\ r \in \mathbb{R},\end{aligned} \tag{19}$$

where ρ is a real number larger than 1. The normal distribution can be viewed as a special case satisfying (19). Then, with ϵ_n, M_n N_n, and J_n defined as above, 18 are not affected, since

$$\Pr\Big\{\sup_{x\in[0,1]}|\sum_{j}^{J_n}\beta_j B_j| > M_n(N_n, resp.))\Big\} \lesssim J_n \exp\Big(-\frac{M_n^\rho}{2}\Big) \lesssim \exp\Big(-n\epsilon_n^2\Big). \tag{20}$$

The prior concentration probability estimation can also be bounded below by a multiple of the volume of a Euclidean ball. Added with the fact that priors does not affect the entropy, we finish showing that the convergence rate can keep still when we generalize the priors.

4.2 Proof for corollary 1

The proof is almost the same with that for theorem 1. We consider the sieves $\Theta_n = \Theta_n^{(f)} \times \Theta_n^{(\eta)}$ in the form of

$$\Theta_n^{(f)} = \{f_\beta \in supp\{\Pi_n^{(f)}\}, ||f_\beta|| \le N_n, J_n \le k_n\};$$

$$\Theta_n^{(\eta)} = \{\eta_\beta \in supp\{\Pi_n^{(\eta)}\}, ||\eta_\beta|| \le M_n, J_n \le k_n\},$$

where $k_n = \lfloor \min\{(n/\log n)^{1/(2\alpha+1)}, n^{1/(2+2\gamma)}\}\rfloor$ and $\lfloor . \rfloor$ denotes the Integral part. With (12), the ϵ_n-entropy of Θ_n is bounded by a multiple of $(\frac{M_n}{\epsilon_n e^{-N_n}})^{J_n} \times (\frac{N_n}{\epsilon_n})^{J_n}$, which have been proved to be always bounded by a multiple of $e^{n\epsilon^2}$ with $J_n \le \lfloor \min\{(n/\log n)^{1/(2\alpha+1)}, n^{1/(2+2\gamma)}\}\rfloor$, $M_n \sim n$, $N_n \sim n^{1/(2\gamma+2)}$and $\epsilon_n \sim \max\{(n/\log n)^{-\alpha/(1+2\alpha)}, n^{-\gamma/(2+2\gamma)}\}$.

The prior concentration probability (13) can be estimated in the form of

$$\begin{aligned}
&\Pi_n(B_n((\eta_0, f_0), \epsilon; 2)) \\
&= \sum_{k=1}^{k_n} \Pr(J_n = k)\Pi_n(B_n((\eta_0, f_0), \epsilon; 2), J_n = k) \\
&\ge \Pr(J_n = k_n)(\inf_{\beta_1 \in [-2N, 2N]} \phi(\beta_1))^2 Vol(\beta : ||\beta - \beta_f^{(n)}|| \le e^{-2N} C^{'} \epsilon) Vol(\beta : ||\beta - \beta_\eta^{(n)}|| \le e^{-2N} C^{'} \epsilon) \\
&\gtrsim \Pr(J_n = k_n)\epsilon^{2k_n}
\end{aligned}$$

With the assumption for p_n and the fact that we have already proved $\epsilon^{2J_n} \gtrsim e^{-n\epsilon^2}$ with $J_n \sim \min\{(n/\log n)^{1/(1+2\alpha)}, n^{1/(2+2\gamma)}\}$, and $\epsilon_n \sim \max\{(n/\log n)^{-\alpha/(1+2\alpha)}, n^{-\gamma/(2+2\gamma)}\}$. we can guarantee

$$p_n^{k_n - 1}(1 - p_n)\epsilon_n^{2k_n} \gtrsim e^{-n\epsilon_n^2}.$$

We compute the probability of $(\Theta_n^{(\eta)})^c$ as following:

$$\begin{aligned}
&\Pi_n((\Theta_n^{(\eta)})^c) \\
&= \sum_{k=1}^{k_n} \Pr(J_n = k) \Pr\Big\{ \sup_{x\in[0,1]} |\sum_j^k \beta_j B_j| > M_n \Big\} + \sum_{k=k_n+1}^{\infty} \Pr(J_n = k) \\
&\lesssim \sum_{k=1}^{k_n} \Pr(J_n = k) k \exp\Big(- \frac{M_n^2}{2}\Big) + \sum_{k=k_n+1}^{\infty} \Pr(J_n = k) \\
&\lesssim k_n \exp\Big(- \frac{M_n^2}{2}\Big) + \sum_{k=k_n+1}^{\infty} \Pr(J_n = k) \\
&\lesssim e^{-n\epsilon^2}
\end{aligned}$$

We derive the last $\lesssim$ through the facts that $k_n \exp\Big(- \frac{M_n^2}{2}\Big) \lesssim e^{-n\epsilon^2}$, and the assumption $p_n^{k_n-1}(1-p_n) = e^{-n\epsilon_n^2}$.

4.3 Proof for theorem 2

We denote κ to be α or γ. By theorem 3.1 in [20], there exists a Borel measurable subset $B_n^{(g)}$ of $C[0,1]^d$ such that

$$\begin{aligned}
&\Pr(||W^{(g)A} - g_0||_\infty \le \epsilon_n) > e^{-n\epsilon_n^2}; \\
&\Pr(W^{(g)A} \notin B_n^{(g)}) \le e^{-4n\epsilon_n^2}; \qquad (21)\\
&\log N(\epsilon_n, B_n^{(g)}, ||.||_\infty) < K^{(g)} n\epsilon_n^2,
\end{aligned}$$

hold, for every sufficiently large n, and $\epsilon_n = n^{-\kappa/2(\kappa+d)}(\log n)^{(d+1)\kappa/(2\kappa+d)}$, $K^{(g)}$ is a sufficiently large constant. As stated in [20], this power can be improved by using a slightly different prior for A. Then, the final rate of contraction will be improved, too, as which can be seen from the following proof.

We set Θ_n in the following way. Denote $\Theta_n^{(f)} = \{W^A \in B_n^{(f)}, and\, ||W^A||_\infty \le N_n\}$. So, $\Theta_n^{(f)}$ increases to $B_n^{(f)}$ as n increases to infinity. As we assumed, $\{N_n\}$ is a sequence of real numbers increasing to infinity. We choose N_n satisfying

$$\Pr(W^{(f)A} \in B_n^{(f)}) - \Pr(W^A \in \Theta_n^{(f)}) \le e^{-4n\epsilon_n^2}.$$

Then

$$\Pr(W^A \notin D_n^{(f)}) \leq 2e^{-4n\epsilon_n^2}.$$

This can be achieved, since $\Pr(W^{(f)A} \in B_n)$ goes to 1 and $e^{-4n\epsilon_n^2}$ goes to zero. Then we set $\Theta_n = B_n^{(\eta)} \times \Theta_n^{(f)} \subset C[0,1]^d \times C[0,1]^d$.

We start to verify all the conditions of general results on rates of posterior contraction. First, we bound the average hellinger distance entropy of the sieve of parameter space.

$$\begin{aligned}
&\log N(\epsilon_n, \Theta_n, d_n) \\
&\lesssim \log N(\epsilon_n, B_n^{(\eta)}, ||.||_n) + \log N(\epsilon_n, \Theta_n^{(f)}, ||.||_n) \\
&\leq \log N(\epsilon_n/e^{N_n}, B_n^{(\eta)}, ||.||_\infty) + \log N(\epsilon_n, B_n^{(f)}, ||.||_\infty) \\
&\leq Kn\epsilon_n^2.
\end{aligned}$$

The first $\lesssim$ is from Lemma 1, the last $\leq$ is because of the third inequality of (21).

To estimate the prior positivity, we still use Lemma 2. With the assumption that $||f||_\infty \leq N_n$, and $||f_0||_\infty \leq N_n$, for sufficiently large n, we can get

$$\begin{aligned}
&\Pi_n(B_n((\eta_0, f_0), \epsilon_n; 2)) \geq \Pi_n(||f - f_0||_n^2 + ||\eta - \eta_0||_n^2 \leq e^{-4N_n}\epsilon_n^2) \\
&\geq \Pi_n^{(f)}(f : ||f - f_0||_n^2 \leq \frac{e^{-4N_n}}{2}\epsilon_n^2) \times \Pi_n^{(\eta)}(\eta : ||\eta - \eta_0||_n^2 \leq \frac{e^{-4N_n}}{2}\epsilon_n^2) \\
&\geq \Pr(||W^{(f)A} - f_0||_\infty \leq \frac{e^{-2N_n}}{\sqrt{2}}\epsilon_n) \times \Pr(||W^{(\eta)A} - \eta_0|| \leq \frac{e^{-2N_n}}{\sqrt{2}}\epsilon) \\
&\geq e^{-2n\epsilon_n^2}.
\end{aligned}$$

Thus, for $\Theta_n \subset C[0,1]^d \times C[0,1]^d$ defined above, and

$$\epsilon_n = \max\{n^{-\alpha/(d+2\alpha)}(\log n)^{(d+1)\alpha/(2\alpha+d)}, n^{-\gamma/(d+2\gamma)}(\log n)^{(d+1)\gamma/(2\gamma+d)}\},$$

we have proved

$$\begin{aligned}
\log N(\epsilon_n, \Theta_n, d_n) &\leq 2Kn\epsilon_n^2 \\
\Pi_n(B_n((\eta_0, f_0), \epsilon; 2)) &\geq e^{-2n\epsilon_n^2} \\
\Pi_n((f, \eta) \notin \Theta_n) &\leq 3e^{-4n\epsilon_n^2}.
\end{aligned}$$

The three assertions can be matched one-to-one with the assumption of general results on rates of posterior contraction (e.g. Theorem 4 in [12]), so the proof is completed.

The proof Corollary 2 is almost the same, except that the value of ϵ_n is given by Theorem 4.1 of [18].

5 Discussion

In this paper, we investigated the posterior convergence rate for heteroscedastic nonparametric regression model with both mean function and variance function unknown and nonparametric. We considered both of the cases with random covariate $\mathbf{x}$, and deterministic covariates. We also put the high-dimensional case in consideration. Though the rates we gave are not the minimax, they are only different with the optimal ones by a logarithmic factor. Besides, they are optimal for every regularity level. And we gave the minimax rate under the condition with known regularity level.

Whether the logarithmic factor of the posterior convergence rate is necessary for unknown regularity level is not known. To investigate this problem, other kinds of priors must be used, since as van der Vaart and van Zanten have conjectured in [20], the logarithmic factor is necessary with the rescaled Gaussian random field prior, and our current method used in the section of splines cannot give the desired result, either.

6 Appendix A. Proof of Lemma 1

By applying the inequalities $2-2ab \leq 2-2a+2-2b$, when $a \leq 1$ and $b \leq 1$, together with $1-e^{-x} \leq x$ for $x \geq 0$, and $1-\frac{2x}{x^2+1} \leq (2\log x)^2$ for all the x, we have

$$
\begin{aligned}
&2-2exp(-\frac{(\eta_1(x)-\eta_2(x))^2}{4(V_1(x)+V_2(x))}) \times \sqrt{\frac{2\sqrt{V_1(x)V_2(x)}}{V_1(x)+V_2(x)}} \\
&\leq 2(1-\sqrt{\frac{2\sqrt{V_1(\mathbf{x})V_2(\mathbf{x})}}{V_1(\mathbf{x})+V_2(\mathbf{x})}})+2(1-exp\{-\frac{(\eta_1(\mathbf{x})-\eta_2(\mathbf{x}))^2}{4(V_1(x)+V_2(x))}\}) \\
&\leq 2(\log(\frac{V_1(\mathbf{x})}{V_2(\mathbf{x})})^2+2\frac{(\eta_1(\mathbf{x})-\eta_2(\mathbf{x}))^2}{4(V_1(x)+V_2(x))}.
\end{aligned}
$$

Thus ,we have

$$d^2(\mathcal{P}_{\eta_1,V_1},\mathcal{P}_{\eta_2,V_2}) = \int\int(\mathcal{P}^{\frac{1}{2}}_{\eta_1,V_1} - \mathcal{P}^{\frac{1}{2}}_{\eta_2,V_2})^2\ dy\ dQ$$
$$\leq 2\int(\log(\frac{V_1(\mathbf{x})}{V_2(\mathbf{x})})^2 + 2\frac{(\eta_1(\mathbf{x})-\eta_2(\mathbf{x}))^2}{4(V_1(x)+V_2(x))})\ dQ$$

held, which is followed by the result

$$\log N(3\epsilon,\Theta_n,d_n) \lesssim \log N(\epsilon/e^{N_n},\Theta_n^{(\eta)},||.||_n) + \log N(\epsilon,\Theta_n^{(f)},||.||_n).$$

provided $||V_i|| > e^{-N_n}, i = 1,2$.

7 Appendix B. Proof of Lemma 2

For the Kullback-Leibler divergence, we have,

$$K_{\mathbf{x}}(\mathcal{P}_{\eta_1,V_1},\mathcal{P}_{\eta_2,V_2}) = \frac{1}{2}\log\frac{V_2}{V_1} - \frac{1}{2}(1-\frac{V_1}{V_2}) + \frac{1}{2}\frac{[\eta_1(\mathbf{x})-\eta_2(\mathbf{x})]^2}{V_2(\mathbf{x})}$$
$$= \frac{1}{2}|(f_2(\mathbf{x})-f_1(\mathbf{x})) - \frac{1}{2}(1-e^{f_1(\mathbf{x})-f_2(\mathbf{x})})| + \frac{1}{2}\frac{[\eta_1(\mathbf{x})-\eta_2(\mathbf{x})]^2}{V(\mathbf{x})}.$$

We know that, for $|z| \leq 2N$,

$$|z-1+e^{-z}| \leq |z| + |e^{-z}-1| \leq (e^{2N}+1)|z|;$$

when $z \geq 1$,

$$(e^{2N}+1)|z| \leq (e^{2N}+1)z^2,$$

when $z \leq 1$

$$|z-1+e^{-z}| \leq \sum_{n=2}^{\infty} z^n/2 \leq \frac{|z|^2/2}{1-|z|} \leq (e^{2N}+1)z^2.$$

Thus:

$$K(\mathcal{P}_{\eta_1,V_1},\mathcal{P}_{\eta_2,V_2}) \leq (1+e^{2N})(||\eta_1-\eta_2||_n^2 + ||(f_1-f_2)||_n^2).$$

For the variance divergence, we have

$$Var_x(\mathcal{P}_{\eta_1,V_1},\mathcal{P}_{\eta_2,V_2}) = 2[-\frac{1}{2}+\frac{1}{2}\frac{V_1(x)}{V_2(x)}]^2 + [\frac{V_1(x)}{V_2(x)}[\eta_1(x)-\eta_2(x)]]^2.$$

We can finish the proof with the inequality $|1-e^z|^2 \leq (e^{2N})^2z^2$ for $|z| \leq 2N$.

References

[1] A. Barron, M. J. Schervish, L. Wasserman, The consistency of posterior distributions in nonparametric problems, Annals of Statistics 27 (2) (1999) 536–561.

[2] S. Ghosal, J. K. Ghosh, T. Samanta, Approximation of the posterior distribution in a change-point problem, Annals of the Institute of Statistical Mathematics 51 (3) (1999) 479–497.

[3] S. Walker, On sufficient conditions for bayesian consistency, Biometrika 90 (2) (2003) 482–488.

[4] S. Walker, A. Lijoi, I. Prünster, On rates of convergence for posterior distributions in infinite-dimensional models, Annals of Statistics 35 (2) (2007) 738–746.

[5] S. Ghosal, A. Roy, Posterior consistency of gaussian process prior for nonparametric binary regression, Annals of Statistics 34 (5) (2006) 2423–2429.

[6] H. Lian, Consistency of bayesian estimation of a step function, Statistics & Probability Letters 77 (1) (2007) 19–24.

[7] H. Lian, On the consistency of bayesian function approximation using step functions, Neural Computation 19 (11) (2007) 2871–2880.

[8] H. Lian, On rates of convergence for posterior distributions under misspecification, Communications in Statistics-Theory and Methods 38 (11) (2009) 1893–1900.

[9] H. Lian, Posterior convergence and model estimation in bayesian change-point problems, Electronic Journal of Statistics 4 (2010) 239–253.

[10] Y. Wu, S. Ghosal, Kullback Leibler property of kernel mixture priors in Bayesian density estimation, Electronic Journal of Statistics 2 (2008) 298–331.

[11] S. Ghosal, J. Lember, A. van Der Vaart, Nonparametric Bayesian model selection and averaging, Electronic Journal of Statistics 2 (2008) 63–89.

[12] S. Ghosal, J. K. Ghosh, A. W. van der Vaart, Convergence rates of posterior distributions, Annals of Statistics 28 (2) (2000) 500–531.

[13] S. Ghosal, A. van der Vaart, Convergence rates of posterior distributions for noniid observations, Annals of Statistics 35 (1) (2007) 192–223.

[14] T. Choi, M. J. Schervish, On posterior consistency in nonparametric regression problems, Journal of Multivariate Analysis 98 (2007) 1969–1987.

[15] M. Amewou-Atisso, S. Ghosal, J. Ghosh, R. Ramamoorthi, Posterior consistency for semiparametric regression problems, Bernoulli 9 (2003) 291–312.

[16] T. Leonard, Density estimaiton, stochastic processes and prior information, Journal of the Royal Statistical Society. Series B 40 (1978) 113–146.

[17] P. J. Lenk, The logistic normal distribution for bayesian, nonparametric, predictive densities, Journal of the American Statistical Association 83 (1988) 509–516.

[18] A. W. van der Vaart, J. van der Zanten, Rates of contraction of posterior distributions based on gaussian process priors, Annals of Statistics 36 (3) (2008) 1435–1463.

[19] A. W. van der Vaart, J. van der Zanten, Reproducing kernal hilbert spaces of gaussian priors, in: Pushing the Limits of Contenporary Statistics: Contributions in Honor of Jayanta K.Ghosh, Vol. 3, IMS Collection, Lithuania, 2008, pp. 200–222.

[20] A. W. van der Vaart, J. van der Zanten, Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth, Annals of Statistics 37 (5B) (2000) 2655–2675.