The Annals of Statistics
2016, Vol. 44, No. 1, 318–345
DOI: 10.1214/15-AOS1368

RATE EXACT BAYESIAN ADAPTATION WITH MODIFIED BLOCK PRIORS[1]

BY CHAO GAO AND HARRISON H. ZHOU

Yale University

A novel block prior is proposed for adaptive Bayesian estimation. The prior does not depend on the smoothness of the function or the sample size. It puts sufficient prior mass near the true signal and automatically concentrates on its effective dimension. A rate-optimal posterior contraction is obtained in a general framework, which includes density estimation, white noise model, Gaussian sequence model, Gaussian regression and spectral density estimation.

1. Introduction. Bayesian nonparametric estimation is attracting more and more attention in a wide range of applications. We consider a fundamental question in Bayesian nonparametric estimation: is it possible to construct a prior such that the posterior contracts to the truth with the exact optimal rate and at the same time is adaptive regardless of the unknown smoothness? We provide a positive answer to this question by designing a block prior on coefficients of orthogonal series expansion of the function.

Specifically, we obtain adaptive Bayesian estimation under a Sobolev ball assumption. Assume that f is a function on the unit interval $[0, 1]$. Let $\{\phi_j\}$ be the trigonometric orthogonal basis of $L^2[0, 1]$, and define $\theta_j = \int f\phi_j$ for each j. The Sobolev ball is specified as

$$E_\alpha(Q) = \left\{ f \in L^2[0,1] : \sum_{j=1}^{\infty} j^{2\alpha}\theta_j^2 \le Q^2, \text{ with } \theta_j = \int f\phi_j \text{ for each } j \right\}.$$

Under a general framework, we construct a prior Π, which satisfies the Kullback–Leibler (KL) property and it automatically concentrates on the effective dimension of the signal f_0, then as a consequence, the minimax posterior contraction rate is obtained, that is,

$$P_{f_0}^{(n)}\Pi\bigl(\|f - f_0\| > Mn^{-\alpha/(2\alpha+1)}|X^n\bigr) \longrightarrow 0, \tag{1}$$

where the loss function $\|\cdot\|$ is the l^2-norm.

Adaptive Bayesian estimators over Sobolev balls or Hölder balls are considered in the literature. There are two main approaches in these works. The first one is to put a hyper-prior on the smoothness index α. As is shown in Scricciolo (2006) and

Received August 2014; revised July 2015.

[1] Supported in part by NSF Grant DMS-12-09191.

MSC2010 subject classifications. 62G07, 62G20.

Key words and phrases. Bayesian nonparametrics, adaptive estimation, block prior.

Ghosal, Lember and van der Vaart (2008), minimax rate can be achieved, but the set of α is restricted to be countable or even finite. The second approach is to put a prior on k, where k is the number of basis functions for approximation, or the model dimension. This is called sieve prior in Shen and Wasserman (2001). Examples of using sieve prior include Kruijer and van der Vaart (2008) and Rivoirard and Rousseau (2012). Their procedures are adaptive over all α, but the rates have extra logarithmic terms. Other recent works in Bayesian adaptive estimation include van der Vaart and van Zanten (2007, 2009), de Jonge and van Zanten (2010), Kruijer, Rousseau and van der Vaart (2010), Rousseau (2010), Shen, Tokdar and Ghosal (2013) and Castillo, Kerkyacharian and Picard (2014), but the posterior contraction rates in these works all miss a logarithmic factor.

The investigation of whether a logarithmic term is necessary in the posterior contraction rate has fundamental implications. The results can lead to answers to two important questions. First, is the presence of a logarithmic term an intrinsic problem to Bayesian adaptive nonparametric estimation? Second, is the presence of a logarithmic term an artifact due to the current proof technique? The answer to the first question should have an impact on statisticians' views of the frequentist/Bayesian debate. The answer to the second question will provide a better understanding on the famous "prior mass and testing" framework [Barron, Schervish and Wasserman (1999); Ghosal, Ghosh and van der Vaart (2000)] that is widely used to establish posterior contraction results.

Compared to the previous results in the literature, the proposed block prior is adaptive over a continuum of smoothness, and its posterior contraction is exactly rate-optimal. The framework for the applications of the block prior is very general. It includes density estimation, white noise, Gaussian sequence, regression and spectral density estimation.

At the point when the first draft of the paper was finished, we received a manuscript by Hoffmann, Rousseau and Schmidt-Hieber (2015) on Bayes adaptive estimation. They considered the similar problem as ours and obtain the exact minimax rate by using a spike and slab prior. However, their adaptation result for the l_2 loss only holds for the white noise model. Since their proof technique takes advantage of the Gaussian sequence structure, it cannot be immediately extended to other model settings. In contrast, by designing a block prior that especially works under the "prior mass and testing" framework, we are able to establish results for models including density estimation, nonparametric regression and spectral density estimation.

The major difficulty of adaptation with the exact rate in various model settings is the design of a prior distribution that satisfies the conditions of the general prior mass and testing framework, which can be applied to a wide range of models. This framework was pioneered by LeCam (1973) and Schwartz (1965), and was later extended to the nonparametric setting by Barron (1988), Barron, Schervish and Wasserman (1999) and Ghosal, Ghosh and van der Vaart (2000). They proved as long as the prior satisfies a Kullback–Leibler property and there exists a testing procedure on the essential support of the prior, the posterior distribution contracts to the truth with certain rate of convergence. Though it is possible to analyze the posterior distribution according to the Bayes formula directly as in Hoffmann, Rousseau and Schmidt-Hieber (2015), the prior mass and testing framework imposes the weakest assumption on the likelihood function, which makes it flexible to various model settings. The price of such flexibility to model settings is the rather strong requirements on the prior. In our opinion, the design of a prior that satisfies the prior mass and testing framework is the major difficulty of achieving rate-optimal adaptation over various model settings. The block prior we propose in this paper gives a solution to this problem. We show that it possesses the strong properties required by the prior mass and testing framework. Therefore, not only does it give rate-optimal adaptation, the good posterior behavior also extends to the settings beyond the white noise model.

The paper is organized as follows. In Section 2, we first introduce a preliminary block prior $\bar{\Pi}$, which satisfies the Kullback–Leibler property and concentrates on the effective dimension of the truth, and then we present the key result of this paper, adaptive rate-optimal posterior contraction for a slightly modified prior Π under a general framework. As applications of the main results, we study adaptive Bayesian estimation of various nonparametric models in Section 3. Section 4 discusses the posterior tail probability bound and an extension of the theory to Besov balls. It also includes discussion on why a logarithmic factor is usually present in the Bayes nonparametric literature. The main body of the proofs are presented in Section 5. Simulation and some auxiliary results of the proofs are given in the supplement [Gao and Zhou (2015)].

1.1. *Notations.* Throughout the paper, $\mathbb{P}$ and $\mathbb{E}$ are generic probability and expectation operators, which are used whenever the distribution is clear in the context. Small and big case letters denote constants which may vary from line to line. We will not pay attention to the values of constants which do not affect the result, unless otherwise specified. Notice these constants may or may not be universal, which we shall make clear in the context. The function f and its Fourier coefficients $\theta = \{\theta_j\}$ are used interchangeably. We say f is distributed by Π if the corresponding $\theta \sim \Pi$. In the same way, the function space and the parameter space of f and θ will not be distinguished. The norm $\|\cdot\|$ denotes both the l^2-norm of f and the l^2-norm of θ. For two probabilities P_1 and P_2 with densities p_1 and p_2, we use the following divergences throughout the paper:

$$D(P_1, P_2) = P_1 \log \frac{p_1}{p_2},$$

$$V(P_1, P_2) = P_1\left(\log \frac{p_1}{p_2} - D(P_1, P_2)\right)^2,$$

$$H(P_1, P_2) = \left(\int (\sqrt{p_1} - \sqrt{p_2})^2\right)^{1/2}.$$

We use θ_j and θ_{0j} to indicate the jth entries of vectors $\theta = \{\theta_j\}$ and $\theta_0 = \{\theta_{0j}\}$, respectively. The bold notation $\boldsymbol{\theta}_k$ represents the vector $\{\theta_j\}_{j \in B_k}$ for the kth block. The rate ε_n is always the minimax rate $\varepsilon_n^2 = n^{-2\alpha/(2\alpha+1)}$.

2. Main results.

In this section, we first give some necessary background of Bayes nonparametric estimation, then introduce a block prior and the result of adaptive posterior contraction.

2.1. *Background.*

Suppose we have data $X^n \sim P_{f_0}^{(n)}$, and the distribution $P_{f_0}^{(n)}$ has density $p_{f_0}^{(n)}$ with respect to a dominating measure. The posterior distribution for a prior Π is defined to be

$$\Pi(A|X^n) = \frac{\int_A (p_f^{(n)}/p_{f_0}^{(n)})(X^n)\, d\Pi(f)}{\int (p_f^{(n)}/p_{f_0}^{(n)})(X^n)\, d\Pi(f)} \qquad \text{where } X^n \sim P_{f_0}^{(n)}.$$

We need to bound the expectation of $\Pi(d(f, f_0) > M\varepsilon_n|X^n)$ in this paper. To bound this quantity, it is sufficient to upper bound the numerator and lower bound the denominator. Following Barron, Schervish and Wasserman (1999) and Ghosal, Ghosh and van der Vaart (2000), this involves three steps:

1. Show the prior Π puts sufficient mass near the truth, that is, we need

$$\Pi(K_n) \geq \exp(-C_1 n\varepsilon_n^2),$$

where $K_n = \{D(P_{f_0}^{(n)}, P_f^{(n)}) \leq n\varepsilon_n^2, V(P_{f_0}^{(n)}, P_f^{(n)}) \leq n\varepsilon_n^2\}$.

2. Choose an appropriate set $\mathcal{F}_n$, and show the prior is essentially supported on $\mathcal{F}_n$ in the sense that

$$\Pi(\mathcal{F}_n^c) \leq \exp(-C_2 n\varepsilon^2).$$

This controls the complexity of the prior.

3. Construct a testing function ϕ_n for the following testing problem:

$$H_0: f = f_0 \quad \text{vs.} \quad H_1: f \in \operatorname{supp}(\Pi) \cap \mathcal{F}_n \quad \text{and} \quad d(f, f_0) > M\varepsilon_n.$$

The testing error needs to be well controlled in the sense that

$$P_{f_0}^{(n)}\phi_n \vee \sup_{f \in H_1} P_f^{(n)}(1 - \phi_n) \leq \exp(-C_3 n\varepsilon^2).$$

Note that the constants C_1, C_2 and C_3 are different in these three steps above. Step 1 lower bounds the prior concentration near the truth, which leads to a lower bound for the denominator $\int \frac{p_f^{(n)}}{p_{f_0}^{(n)}}(X^n)\, d\Pi(f)$. It is originated from Schwartz (1965). Steps 2 and 3 are mainly for upper bounding the numerator $\int_A \frac{p_f^{(n)}}{p_{f_0}^{(n)}}(X^n)\, d\Pi(f)$. The testing idea in step 3 is initialized by LeCam (1973) and

Schwartz (1965). Step 2 goes back to Barron (1988), who proposes the idea to choose an appropriate $\mathcal{F}_n$ to regularize the alternative hypothesis in the test, otherwise the testing function for step 3 may never exist [see LeCam (1973) and Barron (1989)].

2.2. *The block prior* $\bar{\Pi}$. Given a sequence $\theta = (\theta_1, \theta_2, \dots)$ in the Hilbert space l^2. Define the blocks to be $B_k = \{l_k, \dots, l_{k+1} - 1\}$, and $\{1, 2, 3, \dots\} = \bigcup_{k=0}^{\infty} B_k$. Define the block size of the kth block to be $n_k = l_{k+1} - l_k = |B_k|$. Remember the notation $\boldsymbol{\theta}_k$ represents the vector $\{\theta_j\}_{j \in B_k}$. The block prior $\bar{\Pi}$ on the function f is induced by a distribution on its Fourier sequence $\{\theta_j\}$. For each k, let g_k be a one-dimensional density function on $\mathbb{R}^+$.

We describe $\bar{\Pi}$ as follows:

$$A_k \sim g_k \qquad \text{independently for each } k,$$

$$\boldsymbol{\theta}_k | A_k \sim N(0, A_k I_{n_k}) \qquad \text{independently for each } k,$$

where I_{n_k} is the $n_k \times n_k$ identity matrix. In this work, we specify l_k to be $l_k = [e^k]$. The sequence of densities $\{g_k\}$ is used to mix the scale parameter A_k for each block, and we call them mixing densities. Our theory covers a class of mixing densities. The mixing density class $\mathcal{G}$ contains all $\{g_k\}$ satisfying the following properties:

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

$$g_k(t) \geq \exp(-c_1 e^k). \tag{2}$$

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

$$\int_0^\infty t g_k(t)\, dt \leq 4 \exp(-c_2 k^2). \tag{3}$$

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

$$\int_{e^{-k^2}}^\infty g_k(t)\, dt \leq \exp(-c_3 e^k). \tag{4}$$

For a function $f_0 \in E_\alpha(Q)$, define the set

$$\mathcal{F}_n = \mathcal{F}_n(\beta) = \left\{\theta : \sum_{j > (n\beta^{-1})^{1/(2\alpha+1)}} (\theta_j - \theta_{0j})^2 \leq \varepsilon_n^2\right\}. \tag{5}$$

We have the following theorem characterizing the property of $\bar{\Pi}$.

THEOREM 2.1. *For the block prior* $\bar{\Pi}$ *with mixing densities* $\{g_k\} \in \mathcal{G}$, *let* $f_0 \in E_\alpha(Q)$ *for some* α, $Q > 0$, *then there exists a constant* $C > 0$ *such that*

$$\bar{\Pi}\left\{\sum_{j=1}^\infty (\theta_j - \theta_{0j})^2 \leq \varepsilon_n^2\right\} \geq \exp(-Cn\varepsilon_n^2), \tag{6}$$

and

$$\bar{\Pi}(\mathcal{F}_n^c) \le 2\exp(-(C+4)n\varepsilon_n^2), \tag{7}$$

for sufficiently large n whenever $\beta \le (\min\{\frac{c_3}{2(C+4)}, (4Q^2)^{-2\alpha}\})^{2\alpha+1}$, *with c_3 defined in* (4).

REMARK 2.1. The theorem presents two properties of the block prior $\bar{\Pi}$. Property (6) says the prior gives sufficient mass near the true signal f_0. This is also recognized as the K–L condition once the Kullback–Leibler divergence is upper bounded by the l^2-norm in the support of the prior. Property (7) says the prior concentrates on the effective dimension of the true signal f_0 automatically. In Bayesian nonparametric theory, a testing argument is needed to prove posterior contraction rate. Such test can be established on a sieve receiving most of the prior mass. In (7), the set $\mathcal{F}_n$ can be used as such a sieve.

REMARK 2.2. When the smoothness α is known, a well-known prior $\Pi_\alpha = \bigotimes_{j=1}^{\infty} N(0, j^{-2\alpha-1})$ is used in the literature. It can be shown that this prior satisfies (6). The block prior $\bar{\Pi}$ satisfies (6) and (7), and it does not depend on the smoothness α. Thus, it is fully adaptive.

We claim that the mixing density class $\mathcal{G}$ is not empty by presenting an example (Figure 1):

$$g_k(t) = \begin{cases} e^{k^2}(\exp(-e^k) - T_k)t + T_k, & 0 \le t \le e^{-k^2}; \\ \exp(-e^k), & e^{-k^2} < t \le e^{-k}; \\ 0, & t > e^{-k}. \end{cases} \tag{8}$$

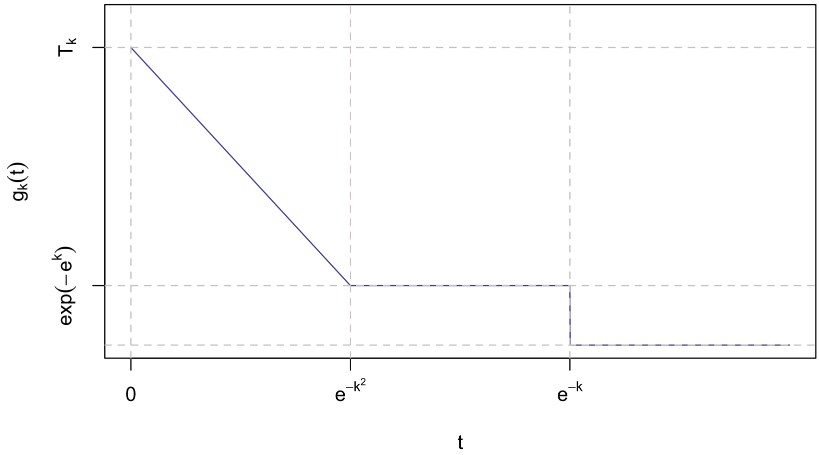

FIG. 1. *The plot of the mixing density function $A_k \sim g_k$ defined in* (8).

The value of T_k is specified as

$$T_k = 2e^{k^2} - 2\exp(-e^k + k^2 - k) + \exp(-e^k). \tag{9}$$

The following proposition is proved in the supplementary material [Gao and Zhou (2015)].

PROPOSITION 2.1. *The densities* $\{g_k\}$ *defined in* (8) *satisfies* (2), (3) *and* (4). *Thus*, $\mathcal{G}$ *is not empty.*

2.3. *Adaptive posterior contraction of the modified block prior* Π. In order to prove posterior contraction rate, it is essential to construct a suitable test. A preliminary test is first constructed in a local neighborhood. Then a global test is established by combining all the local tests when the metric entropy is well controlled. We say the distance d satisfies the testing property with respect to the prior Π and the truth f_0 if and only if there exists some constants $L > 0$ and $\xi \in (0, 1/2)$, such that for any $f_1 \in \operatorname{supp}(\Pi)$ satisfying $d(f_0, f_1) > \varepsilon_n$, we have

$$P_{f_0}^{(n)} \phi_n \leq \exp(-Lnd^2(f_0, f_1)), \tag{10}$$

$$\sup_{\{f \in \operatorname{supp}(\Pi): d(f, f_1) \leq \xi d(f_0, f_1)\}} P_f^{(n)}(1 - \phi_n) \leq \exp(-Lnd^2(f_0, f_1)), \tag{11}$$

for some testing function ϕ_n. Then a global test can be constructed for $H_0 : f = f_0$ against $H_1 = \{f \in \mathcal{F}_n \cap \operatorname{supp}(\Pi) : d(f, f_0) > M\varepsilon_n\}$ as long as $d(f_1, f_2) \asymp \|f_1 - f_2\|$ for any f_1 and f_2. The equivalence of d and $\|\cdot\|$ may not be true for d being Hellinger distance or total variation. We thus consider a modification of the block prior $\bar{\Pi}$, denoted as Π, so that d and $\|\cdot\|$ are equivalent in the support of the modified block prior Π. Define

$$\Pi(A) = \frac{\bar{\Pi}(D \cap A)}{\bar{\Pi}(D)},$$

where the constraint set D needs to be designed case by case such that

$$D(P_{f_1}^{(n)}, P_{f_2}^{(n)}) \leq bn\|f_1 - f_2\|^2, \qquad V(P_{f_1}^{(n)}, P_{f_2}^{(n)}) \leq bn\|f_1 - f_2\|^2,$$
$$b^{-1} d(f_1, f_2) \leq \|f_1 - f_2\| \leq bd(f_1, f_2),$$

for some constant $b > 1$. We give a specific choice of D for each model considered in this paper. Another crucial property of D we need is that Π inherits properties (6) and (7) from $\bar{\Pi}$. It is obvious that (7) is still true for Π as long as $\bar{\Pi}(D) > 0$. Therefore, one only needs to check (6), which is usually not hard as we will see in all the examples in Section 3. A general theorem covers all examples in Section 3 is stated as follows.

THEOREM 2.2. *For the block prior $\bar{\Pi}$ with mixing densities $\{g_k\} \in \mathcal{G}$, define $\Pi(A) = \frac{\bar{\Pi}(D \cap A)}{\bar{\Pi}(D)}$ with the constraint set D satisfying the properties above. Let the distance d satisfy the testing property* (10) *and* (11). *Assume that, for any $f_0 \in E_\alpha(Q) \cap D$ with $\alpha \in (\alpha^*, \infty)$ and $Q \in (0, Q^*)$, the prior Π inherits properties* (6) *and* (7) *from $\bar{\Pi}$ for some $C > 0$. Then, for any such f_0, there exists $M > 0$, such that*

$$P_{f_0}^{(n)} \Pi\big(d(f, f_0) > M n^{-\alpha/(2\alpha+1)} | X^n\big) \longrightarrow 0.$$

REMARK 2.3. We note that the range $\alpha \in (\alpha^*, \infty)$ and $Q \in (0, Q^*)$ is the adaptive region for the prior Π. It is determined by the constraint set D and by whether properties (6) and (7) can be inherited from $\bar{\Pi}$ to Π. In some examples such as the white noise model, the modification by D is not needed, so that we have $\Pi = \bar{\Pi}$. This will result in $\alpha^* = 0$ and $Q^* = \infty$, and thus the prior may adapt to all Sobolev balls. In the regression and the density estimation models, α^* needs to be larger than $1/2$, and Q^* can be chosen arbitrarily large by properly picking the corresponding D. For the spectral density estimation, we need $\alpha^* > 3/2$. See Section 3 for details.

REMARK 2.4. Theorem 2.2 requires the assumption $f_0 \in E_\alpha(Q) \cap D$. In all the nonparametric estimation examples we consider in Section 3, we consider very specific forms of D and we are going to show that such D can be removed from the assumption because of the relation $E_\alpha(Q) \subset D$ for $\alpha > \alpha^*$. This implies $E_\alpha(Q) \cap D = E_\alpha(Q)$ and we only need $f_0 \in E_\alpha(Q)$ in the assumption.

3. Applications.

Given the experiment $((\mathcal{X}^{(n)}, \mathcal{A}^{(n)}, P_f^{(n)}) : f \in E_\alpha(Q))$, and observation $X^n \sim P_{f_0}^{(n)}$, we estimate the function f_0 by an adaptive Bayesian procedure. The goal is to achieve the minimax posterior contraction rate without knowing the smoothness α. In this section, we consider the following examples:

1. *Density estimation*. The observations $X_1, \dots, X_n$ are i.i.d. distributed according to the density

$$p_f(t) = \frac{e^{f(t)}}{\int e^{f(t)}\,dt},$$

for some function f in a Sobolev ball.

2. *White noise*. The observation $Y_t^{(n)}$ is from the following process:

$$dY_t^{(n)} = f(t)\,dt + \frac{1}{\sqrt{n}}\,dW_t,$$

where W_t is the standard Wiener process.

3. *Gaussian sequence*. We have independent observations

$$X_i = \theta_i + n^{-1/2} Z_i, \qquad i \in \mathbb{N},$$

where $\{\theta_i\}$ are Fourier coefficients of f, and $\{Z_i\}$ are i.i.d. standard Gaussian variables.

4. *Gaussian regression*. The design is uniform $X \sim U[0, 1]$. Given X, $Y|X \sim N(f(X), 1)$. The observations are i.i.d. pairs $(X_1, Y_1), \dots, (X_n, Y_n)$.

5. *Spectral density*. The observations are stationary Gaussian time series $X_1, \dots, X_n$ with mean 0 and auto-covariance $\eta_h(g) = \int_{-\pi}^{\pi} e^{ih\lambda} g(\lambda)\, d\lambda$. The spectral density g is modeled by $g = \exp(f)$ for some symmetric f in a Sobolev ball.

The above models have similar frequentist estimation procedures, which is due to the deep fact that they are asymptotically equivalent to each other under minor regularity assumptions. References for asymptotic equivalence theory include Brown and Low (1996), Nussbaum (1996), Brown et al. (2002) and Golubev, Nussbaum and Zhou (2010).

3.1. *Density estimation*. Let $P_f^{(n)}$ be the product measure $P_f^{(n)} = \bigotimes_{i=1}^n P_f$. The data is i.i.d. $X^n = (X_1, \dots, X_n) \sim \bigotimes_{i=1}^n P_{f_0}$. Let P_f be dominated by Lebesgue measure μ, and it has density function $p_f(t) = \frac{e^{f(t)}}{\int_0^1 e^{f(t)} \mu(dt)}$. Consider the Fourier expansion $f = \sum_j \theta_j \phi_j$, and the density p_f can be written in the form of infinite-dimensional exponential family:

$$p_f(t) = \exp\biggl(\sum_j \theta_j \phi_j(t) - \psi(\theta)\biggr),$$

where

$$\psi(\theta) = \int_0^1 e^{\sum_j \theta_j \phi_j(t)} \mu(dt).$$

Notice the first Fourier base function is $\phi_1(t) = 1$. It is easy to see that different θ_1's correspond to the same p_f. For identifiability, we set $\theta_1 = 0$, so that we have $\int f(t)\mu(dt) = \sum_{j \ge 2} \theta_j \int \phi_j(t)\, dt = 0$. We use the modified block prior $\Pi(A) = \frac{\bar{\Pi}(D \cap A)}{\bar{\Pi}(D)}$ with the constraint set

$$D = \Biggl\{\theta : \sum_{j=1}^{\infty} |\theta_j| < B\Biggr\}, \tag{12}$$

for some constant $B > 0$. The next lemma shows that the modified block prior Π inherits properties (6) and (7) from $\bar{\Pi}$.

LEMMA 3.1. *For $\alpha^* > 1/2$, define the constant*

$$\gamma = \Biggl(\sum_{j=1}^{\infty} j^{-2\alpha^*}\Biggr)^{1/2} < \infty. \tag{13}$$

For any $f_0 \in E_\alpha(Q)$, *with* $\alpha \geq \alpha^*$ *and* $3\gamma Q \leq B$, *there is a constant* $C > 0$, *such that*

$$\Pi\left\{\sum_{j=1}^{\infty}(\theta_{0j} - \theta_j)^2 \leq \varepsilon_n^2\right\} \geq \exp\bigl(-Cn\varepsilon_n^2\bigr),$$

and

$$\Pi\bigl(\mathcal{F}_n^c\bigr) \leq 2\exp\bigl(-(C+4)n\varepsilon_n^2\bigr).$$

For density estimation, it is natural to use Hellinger distance as the testing distance d. According to the testing theory in LeCam (1973) and Ghosal, Ghosh and van der Vaart (2000), it satisfies testing property (10) and (11). The next lemma establishes equivalence among various distances and divergences under D defined in (12).

LEMMA 3.2. *On the set* D, *there exists a constant* $b > 1$, *such that*

$$D(P_{f_1}, P_{f_2}) \leq b\|\theta_1 - \theta_2\|^2, \qquad V(P_{f_1}, P_{f_2}) \leq b\|\theta_1 - \theta_2\|^2,$$
$$b^{-1}H(P_{f_1}, P_{f_2}) \leq \|\theta_1 - \theta_2\| \leq bH(P_{f_1}, P_{f_2}).$$

We will prove the above two lemmas in the supplementary material [Gao and Zhou (2015)]. The main result of posterior contraction for density estimation is stated as follows.

THEOREM 3.1. *Let* $\alpha^* > 1/2$ *be fixed*, *and* γ *is the associated constant defined in* (13). *For any* α, Q *satisfying* $\alpha \geq \alpha^*$ *and* $B \geq 3\gamma Q$, *there is a constant* $M > 0$, *such that*

$$\sup_{f_0 \in E_\alpha(Q)} P_{f_0}^n \Pi\bigl(H(P_f, P_{f_0}) > M\varepsilon_n | X_1, \ldots, X_n\bigr) \longrightarrow 0.$$

REMARK 3.1. The prior Π depends on the value of B, which determines the range of adaptation. For any $\alpha^* > 1/2$ and $Q^* > 0$, we can choose B satisfying $B \geq 3\gamma Q^*$ (γ depends on α^*), such that the prior Π is adaptive for all $E_\alpha(Q)$ with $\alpha \geq \alpha^*$ and $Q \leq Q^*$.

3.2. *White noise.* We let $P_f^{(n)}$ be the distribution of the following process:

$$dY_t^{(n)} = f(t)\,dt + \frac{1}{\sqrt{n}}\,dW_t, \qquad t \in [0,1],$$

where W_t is the standard Wiener process and the signal has Fourier expansion $f = \sum_j \theta_j \phi_j$. This model is the simplest and most studied nonparametric model. It is equivalent to the Gaussian sequence model, and we have

$$D\bigl(P_{f_0}^{(n)}, P_f^{(n)}\bigr) = \tfrac{1}{2}n\|f - f_0\|^2, \qquad V\bigl(P_{f_0}^{(n)}, P_f^{(n)}\bigr) = n\|f - f_0\|^2.$$

In the white noise model, it is natural to use the l^2 norm as the testing distance d. The following lemma is from Lemma 5 in Ghosal and van der Vaart (2007).

LEMMA 3.3. *Let* $\phi_n = \{2\int(f_1(t) - f_0(t))\,dY_t^{(t)} > \|f_1\|^2 - \|f_0\|^2\}$. *Then we have*

$$P_{f_0}^{(n)}\phi_n \le 1 - \Phi(\sqrt{n}\|f_1 - f_0\|/2),$$

$$\sup_{\{f:\|f-f_1\|\le\|f_1-f_0\|/4\}} P_f^{(n)}(1-\phi_n) \le 1 - \Phi(\sqrt{n}\|f_1 - f_0\|/4),$$

where Φ *is the standard Gaussian cumulative distribution function.*

By the property of Gaussian tail, we have

$$1 - \Phi(\sqrt{n}L\|f_1 - f_0\|) \le e^{-(1/2)L^2 n\|f_1-f_0\|^2},$$

provided $\sqrt{n}L\|f_1 - f_0\| > 1$, which is true because we only need to test those f_1 with $\|f_1 - f_0\| > M\varepsilon_n$, and we have $\sqrt{n}\varepsilon_n \to \infty$. Therefore, in the white noise model, the distance satisfying (10) and (11) is the l^2 norm. Considering that the divergence $D(P_{f_0}^{(n)}, P_f^{(n)})$ and $V(P_{f_0}^{(n)}, P_f^{(n)})$ are also l^2 norm, we reach the following conclusion.

THEOREM 3.2. *In the white noise model, for any* $\alpha > 0$ *and* $Q > 0$, *there exists a constant* $M > 0$, *such that*

$$\sup_{f_0\in E_\alpha(Q)} P_{f_0}^{(n)}\bar{\Pi}(\|f - f_0\| > M\varepsilon_n | Y_t^{(n)}) \longrightarrow 0.$$

Hence, this is a case that we have adaptation for all Sobolev balls.

3.3. *Gaussian sequence.* The Gaussian sequence model is equivalent to the while noise model. We present this case just for illustration of the theory. Given $f = \sum_j \theta_j\phi_j$, the model $P_f^{(n)}$ is in a product form

$$P_f^{(n)} = \bigotimes_{i=1}^{\infty} P_{\theta_i}^{(n)} = \bigotimes_{i=1}^{\infty} N(\theta_i, n^{-1}). \tag{14}$$

Thus, the observations are independent Gaussian variables in the form

$$X_i = \theta_i + n^{-1/2}Z_i, \qquad i \in \mathbb{N},$$

where $\{Z_i\}$ are i.i.d. standard Gaussian variables. The divergence in this case is easy to calculate. That is, $D(P_{f_0}^{(n)}, P_f^{(n)}) = \frac{n}{2}\|\theta_0 - \theta\|^2$ and $V(P_{f_0}^{(n)}, P_f^{(n)}) = n\|\theta_0 - \theta\|^2$, and they are exactly the l^2 norm. Define

$$\phi_n(X) = \{\|X - \theta_1\|^2 < \|X - \theta_0\|^2\} = \{X^T(\theta_1 - \theta_0) > \|\theta_1\|^2 - \|\theta_0\|^2\}.$$

We observe this is exactly the same test in the white noise model, and thus Lemma 3.3 applies here. Therefore,

$$P_{f_0}^{(n)}\phi_n \le e^{-(1/8)n\|\theta_0-\theta_1\|^2},$$

$$\sup_{\{\theta:\|\theta-\theta_1\|\le\|\theta_1-\theta_0\|/4\}} P_f^{(n)}(1-\phi_n) \le e^{-(1/32)n\|\theta_0-\theta_1\|^2}.$$

The d satisfying the testing property (10) and (11) can be chosen as the l^2 norm. We thus reach the following conclusion.

THEOREM 3.3. *In the Gaussian sequence model, for any $\alpha > 0$ and $Q > 0$, there exists a constant $M > 0$, such that*

$$\sup_{f_0\in E_\alpha(Q)} P_{f_0}^{(n)}\bar{\Pi}\big(\|\theta-\theta_0\| > M\varepsilon_n|X_1, X_2, \dots\big) \longrightarrow 0.$$

We have adaptation for all Sobolev balls.

3.4. *Gaussian regression.* We consider uniform random design instead of fixed design, because the random design allows simple connection between various divergences and the l^2 distance. The model $P_f^{(n)}$ gives i.i.d. observations $(X_1, Y_1), \dots, (X_n, Y_n)$ with distribution

$$X \sim U[0,1], \qquad Y|X \sim N\big(f(X), 1\big).$$

The theory is easily extended to general random design with $X \sim q$ for some density q on [0, 1] bounded from above and below. We choose the uniform design for simplicity of presentation. The function has Fourier expansion $f = \sum_j \theta_j\phi_j$ so that we can apply the modified block prior on f. Let P_f be the distribution of a single observation, and we need to calculate $D(P_{f_0}, P_f)$ and $V(P_{f_0}, P_f)$. Let ϕ be the standard normal density, and it can be shown that $D(P_{f_0}, P_f) \le \frac{1}{2}\|f - f_0\|^2$ and $V(P_{f_0}, P_f) \le (1 + \frac{1}{2}(\|f\|_\infty^2 + \|f_0\|_\infty^2))\|f - f_0\|^2$. As what we have done in the density estimation case, we use the modified block prior $\Pi(A) = \frac{\bar{\Pi}(A\cap D)}{\bar{\Pi}(D)}$ with the constraint set $D = \{\sum_{j=1}^\infty |\theta_j| < B\}$. According to Lemma 3.1, the prior Π inherits properties (6) and (7) from $\bar{\Pi}$. Therefore, for f and $f_0 \in D$, $V(P_{f_0}, P_f) \le (1 + 2B^2)\|f - f_0\|^2$. Next, we deal with the testing procedure. We use the likelihood ratio test as in the white noise and Gaussian sequence model cases, and the error is bounded in the following lemma.

LEMMA 3.4. *There exists a constant $L > 0$, such that for any $f_0, f_1 \in D$ satisfying $\sqrt{n}\|f_1 - f_0\| > 1$, there exits a testing function ϕ_n with error probability bounded as*

$$P_{f_0}^{(n)}\phi_n \le e^{-Ln\|f_0-f_1\|^2},$$

$$\sup_{\{f\in\mathrm{supp}(\Pi):\|f-f_0\|^2\le 1/32\|f_1-f_0\|^2\}} P_f^{(n)}(1-\phi_n) \le e^{-Ln\|f_0-f_1\|^2}.$$

The lemma will be proved in later sections. It says l^2 norm satisfies the testing property (10) and (11). Using Theorem 2.2, we reach the following conclusion.

THEOREM 3.4. *Let $\alpha^* > 1/2$ and γ be the constant defined in* (13). *In the Gaussian regression model with uniform random design, for any α, Q satisfying $\alpha \geq \alpha^*$ and $3\gamma Q \leq B$, there exists a constant $M > 0$, such that*

$$\sup_{f_0 \in E_\alpha(Q)} P_{f_0}^{(n)} \Pi\big(\|f - f_0\| > M\varepsilon_n | X_1, \dots, X_n, Y_1, \dots, Y_n\big) \longrightarrow 0.$$

REMARK 3.2. The prior Π depends on the value of B, which determines the range of adaptation. For any $\alpha^* > 1/2$ and $Q^* > 0$, we can choose B satisfying $B \geq 3\gamma Q^*$ (γ depends on α^*), such that the prior Π is adaptive for all $E_\alpha(Q)$ with $\alpha \geq \alpha^*$ and $Q \leq Q^*$.

3.5. *Spectral density estimation.* Suppose the probability $P_f^{(n)}$ generates stationary Gaussian time series data $X_1, \dots, X_n$ with mean 0 and spectral density $g = e^f$, with $f(t) = f(-t)$. We assume the spectral density to be a function on $[-\pi, \pi]$. The auto-covariance is $\eta_h = \int_{-\pi}^{\pi} e^{iht} g(t)\, dt$. Thus, the observation $(X_1, \dots, X_n)$ follows $P_f^{(n)} = N(0, \Gamma_n(g))$, where the covariance matrix is

$$\Gamma_n(g) = \begin{pmatrix} \eta_0 & \eta_1 & \cdots & \eta_{n-1} \\ \eta_1 & \eta_0 & \cdots & \eta_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_{n-1} & \eta_{n-2} & \cdots & \eta_0 \end{pmatrix}.$$

We model the exponent of the spectral density by $f(t) = \sum_{j=0}^{\infty} \theta_j \cos(jt)$. According to Parseval's identity, we have $2\pi \|g\|^2 = \|\eta\|^2$ and $2\pi \|f\|^2 = \|\theta\|^2$. We use the modified block prior $\Pi(A) = \frac{\bar{\Pi}(D \cap A)}{\bar{\Pi}(D)}$ with the constraint set

$$D = \left\{ \sum_{j=0}^{\infty} j|\theta_j| < B \right\}. \tag{15}$$

The constraint set (15) is stronger than (12). Thus, in order that the modified prior $\bar{\Pi}$ inherits properties (6) and (7) from the block prior Π, we need $\alpha > 3/2$. The following lemma will be proved in the supplementary material [Gao and Zhou (2015)].

LEMMA 3.5. *For an arbitrary $\alpha^* > 3/2$, and the constant γ defined as*

$$\gamma = \sum_{j=1}^{\infty} j^{2-2\alpha^*}. \tag{16}$$

For any $f_0 \in E_\alpha(Q)$, with $\alpha \geq \alpha^*$ and $3\gamma Q \leq B$, there is a constant $C > 0$, such that

$$\Pi\left\{\sum_{j=1}^{\infty}(\theta_{0j} - \theta_j)^2 \leq \varepsilon_n^2\right\} \geq \exp(-Cn\varepsilon_n^2),$$

and

$$\Pi(\mathcal{F}_n^c) \leq 2\exp(-(C+4)n\varepsilon_n^2).$$

The following lemma, comparing the l^2 norm with $D(P_{f_0}^{(n)}, P_f^{(n)})$ and $V(P_{f_0}^{(n)}, P_f^{(n)})$, will be proved in the supplementary material [Gao and Zhou (2015)].

LEMMA 3.6. *For any $f_0, f_1 \in D$, we have*

$$D(P_{f_0}^{(n)}, P_{f_1}^{(n)}) \leq bn\|f_0 - f_1\|^2,$$
$$V(P_{f_0}^{(n)}, P_{f_1}^{(n)}) \leq bn\|f_0 - f_1\|^2,$$

where $b > 1$ is a constant only depending on Π.

The testing distance satisfying the testing properties (10) and (11) is the l^2-norm.

LEMMA 3.7. *There exists constants $L > 0$ and $0 < \xi < 1/2$, such that for any $f_0, f_1 \in D$ with $\|f_0 - f_1\|^2 \geq \varepsilon_n^2$, there exists a testing function ϕ_n such that*

$$P_{f_0}^{(n)}\phi_n \leq \exp(-Ln\|f_0 - f_1\|^2),$$
$$\sup_{\{f\in\operatorname{supp}(\Pi):\|f-f_1\|\leq\xi\|f_1-f_0\|\}} P_f^{(n)}(1-\phi_n) \leq \exp(-Ln\|f_0 - f_1\|^2).$$

The lemma will be proved in later sections. We state the main result of posterior contraction of spectral density estimation as follows.

THEOREM 3.5. *In the spectral density estimation problem, let $(X_1, \dots, X_n) \sim P_{f_0}^{(n)}$. For any α and Q satisfying Lemma 3.5, there is a constant $M > 0$, such that*

$$\sup_{f_0\in E_\alpha(Q)} P_{f_0}^{(n)}\Pi(\|f - f_0\| > M\varepsilon_n | X_1, \dots, X_n) \longrightarrow 0.$$

REMARK 3.3. The prior Π depends on the value of B, which determines the range of adaptation. For any $\alpha^* > 3/2$ and $Q^* > 0$, we can choose B satisfying $B \geq 3\gamma Q^*$ (γ depends on α^*), such that the prior Π is adaptive for all $E_\alpha(Q)$ with $\alpha \geq \alpha^*$ and $Q \leq Q^*$. Notice the definition of γ in (16) is different from that in (13).

4. Discussion.

4.1. *Exponential tail of the posterior.* The conclusion of the main posterior contraction result in Theorem 2.2 does not specify a decaying rate of the posterior tail. In fact, by scrutinizing the its proof, it has the following polynomial tail:

$$P_{f_0}^{(n)}\Pi\big(\|\theta-\theta_0\|>M\varepsilon_n|X^n\big)\leq\frac{C'}{n\varepsilon_n^2}.$$

However, to obtain a point estimator such as posterior mean with the same rate of convergence as ε_n, faster posterior tail probability is needed [see, e.g., Ghosal, Ghosh and van der Vaart (2000) and Shen and Wasserman (2001)]. In this section, we show that this polynomial tail can be improved to exponential tail in all the examples we consider in Section 3. The critical step is the following lemma, which improves Lemma 5.6 in the proof of the general result of Theorem 2.2.

LEMMA 4.1. *For all statistical models we consider in Section 3 and the corresponding modified block prior Π, let C be the constant with which Π satisfies (6) and (7). Define*

$$\mathcal{H}_n=\left\{\int\frac{p_f^{(n)}}{p_{f_0}^{(n)}}(X^n)\,d\Pi(f)\geq\exp\big(-(C+b+1)n\varepsilon_n^2\big)\right\}. \tag{17}$$

Then we have $P_{f_0}^{(n)}(\mathcal{H}_n^c)\leq\exp(-\bar{C}n\varepsilon^2)$ for $f_0\in E_\alpha(Q)\cap D$ and some $\bar{C}>0$.

From Lemma 4.1, we have the following improved result for posterior contraction.

THEOREM 4.1. *The conclusions of Theorems 3.1, 3.2, 3.3, 3.4 and 3.5 can be strengthened as*

$$P_{f_0}^{(n)}\Pi\big(\|\theta-\theta_0\|>M\varepsilon_n|X^n\big)\leq\exp\big(-C'n\varepsilon_n^2\big),$$

under their corresponding settings.

As a consequence, the posterior mean serves as a rate-optimal point estimator.

COROLLARY 4.1. *Under the setting of Theorems 3.1, 3.2, 3.3, 3.4 and 3.5, we have*

$$P_{f_0}^{(n)}\|\mathbb{E}_{\bar{\Pi}}(\theta|X^n)-\theta_0\|^2\leq M'\varepsilon_n^2,$$

for some constant $M'>0$.

The proofs of Lemma 4.1, Theorem 4.1 and Corollary 4.1 are presented in the supplementary material [Gao and Zhou (2015)].

4.2. *Extension to Besov balls.* Besov balls provides a more flexible collection of functions than Sobolev balls. They are related to wavelet bases. The block prior we propose in this paper naturally takes advantage of the multi-resolution structure of Besov balls. Given a sequence $\{\theta_j\}$, define $\boldsymbol{\theta}_k = \{\theta_{2^k+l}\}_{l=0}^{2^k-1}$ for $k = 0, 1, 2, \dots$. We can view the signals on each resolution level $\boldsymbol{\theta}_k$ as a natural block with size $n_k = 2^k$. The Besov ball is defined as

$$B^{\alpha}_{p,q}(Q) = \left\{\theta : \sum_k 2^{skq} \|\boldsymbol{\theta}_k\|_p^q \le Q^q\right\},$$

where $s = \alpha + \frac{1}{2} - \frac{1}{p}$ and $\|\cdot\|_p$ is the vector l^p-norm. We consider the nonsparse case where the parameters are restricted by

$$(\alpha, p, q, Q) \in (0, \infty) \times [2, \infty] \times [1, \infty] \times (0, \infty). \tag{18}$$

Under such restriction, the block prior is suitable for estimating the signal in $B^{\alpha}_{p,q}(Q)$. We describe the prior $\bar{\Pi}$ as follows:

$$A_k \sim g_k \qquad \text{independently for each } k,$$
$$\boldsymbol{\theta}_k | A_k \sim N(0, A_k I_{n_k}) \qquad \text{independently for each } k,$$

where I_{n_k} is the $2^k \times 2^k$ identity matrix. The mixing densities $\{g_k\}$ are defined through (8) and (9) with the constant e replaced by 2. It is clear that the new mixing densities $\{g_k\}$ satisfies (2), (3) and (4) with every e replaced by 2. Define the new sieve

$$\mathcal{F}_n = \left\{ \sum_{k > (2\alpha+1)^{-1} \log_2(n\beta^{-1})} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{0k}\|^2 \le \varepsilon_n^2 \right\}.$$

We state the property of the block prior $\bar{\Pi}$ targeting at Besov balls below.

THEOREM 4.2. *For the block prior $\bar{\Pi}$ defined above, let $\theta_0 \in B^{\alpha}_{p,q}(Q)$ with (α, p, q, Q) satisfying (18), then there exists a constant $C > 0$ such that*

$$\bar{\Pi}\left\{\sum_{j=1}^{\infty} (\theta_j - \theta_{0j})^2 \le \varepsilon_n^2\right\} \ge 2^{-Cn\varepsilon_n^2}, \tag{19}$$

and

$$\bar{\Pi}(\mathcal{F}_n^c) \le 2^{1-(C+4)n\varepsilon_n^2}, \tag{20}$$

for sufficiently large n whenever $\beta \le (\min\{\frac{c_3}{2(C+4)}, (4Q^2)^{-2\alpha}\})^{2\alpha+1}$, with c_3 defined in (4) where e is replaced by 2.

We apply the prior to the Gaussian sequence model. For other models, some slightly extra works are needed.

THEOREM 4.3. *For the Gaussian sequence model* (14) *with any* $\theta_0 \in B^\alpha_{p,q}(Q)$, *where* (α, p, q, Q) *satisfies* (18), *then there exists* $M > 0$, *such that*

$$\sup_{\theta_0 \in B^\alpha_{p,q}(Q)} P^{(n)}_{\theta_0} \bar{\Pi}\big(\|\theta - \theta_0\| > M\varepsilon_n | X_1, X_2, \ldots\big) \longrightarrow 0.$$

Thus, *the prior is adaptive for all Besov balls satisfying* (18).

We prove the results of the extension in the supplementary material [Gao and Zhou (2015)].

4.3. *Difficulty of achieving the exact rate.* The literature of Bayes nonparametric adaptive estimation usually reports an extra logarithmic term along with the minimax rate ε_n^2. In this section, we provide examples of two priors and illustrate the reasons for them to have the extra logarithmic term. In the first example, the difficulty lies in the prior itself. In the second example, the difficulty lies in the method of proof. The analysis also sheds light on why the block prior is able to achieve the exact minimax rate.

4.3.1. *Difficulty due to the prior.* One of the most elegant priors on f is the rescaled Gaussian process studied by van der Vaart and van Zanten (2007, 2009). Consider the centered Gaussian process $(W_t : t \in [0, 1])$ with the double exponential kernel $\mathbb{E} W_t W_s = \exp(-(s - t)^2)$. The rescaled Gaussian process is defined as $W_{t/c}$ for some c either fixed or sampled from a hyper-prior. The reason for the rescaling is that the original W_t has an infinitely differentiable sample path almost surely. The rescaling step makes it rougher so that it is appropriate for estimating a signal in Sobolev or Hölder balls. In van der Vaart and van Zanten (2007), the number c is fixed as $(n/(\log n)^2)^{-1/(2\alpha+1)}$, and in van der Vaart and van Zanten (2009) c is sampled from a Gamma distribution. The posterior convergence rates are $\varepsilon_n^2(\log n)^{(4\alpha)/(2\alpha+1)}$ and $\varepsilon_n^2(\log n)^{(4\alpha+1)/(2\alpha+1)}$, respectively.

Recently, this prior was extended by Castillo, Kerkyacharian and Picard (2014) for estimation of a function living on a general manifold $\mathcal{M}$. They constructed a rescaled Gaussian process on $\mathcal{M}$ and obtained an improved posterior convergence rate $\varepsilon_n^2(\log n)^{(2\alpha)/(2\alpha+1)}$. Moreover, they also showed that such a rate cannot further be improved by a rescaled Gaussian process with a reasonable distribution on the rescaling parameter c. To be specific, they proved that under mild conditions, there exists a function $f_0 \in B^\alpha_{2,\infty}(Q)$ and a constant $C > 0$, such that

$$P^{(n)}_{f_0} \Pi\big(\|f - f_0\| \le C\varepsilon_n^2(\log n)^{(2\alpha)/(2\alpha+1)} | X^n\big) \to 0,$$

for a rescaled Gaussian process Π. Hence, the posterior convergence rate cannot be faster than $\varepsilon_n^2(\log n)^{(2\alpha)/(2\alpha+1)}$.

To summarize, in this example, the difficulty lies in the prior. It is shown that a certain class of prior distribution is unable to achieve the exact minimax rate.

4.3.2. *Difficulty due to the proof.* The sieve prior is another popular prior used in Bayes nonparametric estimation. It first samples an integer J, which is the model dimension. Conditioning on J, θ_j is sampled from some distribution p independently for all $j \leq J$ and is set to zero for $j > J$. Rivoirard and Rousseau (2012) considered both fixed $J = [n^{1/(2\alpha+1)}]$ and J sampled from a distribution with exponential tail. In the first case, the posterior convergence rate is $\varepsilon_n^2(\log n)^2$ and a slightly slower rate is obtained for the second case.

We argue that the difficulty for obtaining the exact minimax rate is not due to the sieve prior itself, but due to the technique of the proof. Using the prior mass and testing (see Section 2.1) proof technique developed by Barron, Schervish and Wasserman (1999) and Ghosal, Ghosh and van der Vaart (2000), it is impossible to get the exact minimax rate. Let us consider the Gaussian sequence model. In this case, the prior mass condition for the truth $\theta_0 \in E_\alpha(Q)$ and the rate ε_n^2 is

$$\Pi(\|\theta - \theta_0\|^2 \leq \varepsilon_n^2) \geq \exp(-Cn\varepsilon_n^2), \tag{21}$$

for some constant $C > 0$. Even in the simplest sieve prior where J is chosen to be fixed, (21) cannot hold. This is established in the following lemma.

LEMMA 4.2. *Consider a sieve prior with fixed J and density p. Assume $\|p\|_\infty \leq G$ for some constant $G > 0$. Then, for any $\delta_n \to 0$ satisfying $\log \delta_n^{-1} \asymp \log n$ and any θ_0, we have*

$$\Pi(\|\theta - \theta_0\|^2 \leq \delta_n^2) \leq \exp(-CJ\log n),$$

for some constant $C > 0$.

In the ideal case where $J = [n^{1/(2\alpha+1)}]$, the best possible δ_n^2 for (21) to hold is $\delta_n^2 \asymp n^{-(2\alpha)/(2\alpha+1)}\log n$. The extra $\log n$ term cannot be avoided to establish the desired prior mass condition.

On the other hand, we show that the sieve prior in Lemma 4.2 does achieve the exact minimax rate when p is taken as $N(0,1)$.

LEMMA 4.3. *For Gaussian sequence model, consider the prior distribution $\Pi = \bigotimes_{j=1}^J N(0,1)$, with $J = [n^{1/(2\alpha+1)}]$. Then we have for any $\theta_0 \in E_\alpha(Q)$,*

$$P_{f_0}^{(n)}\Pi(\|\theta - \theta_0\|^2 \geq M\varepsilon_n^2 | X^n) \leq \exp(-Cn\varepsilon_n^2),$$

for some constants $C, M > 0$.

The proof of this results takes advantage of the conjugacy and calculates the posterior probability directly from the posterior distribution formula. Both the proofs of Lemmas 4.2 and 4.3 are stated in the supplementary material [Gao and Zhou (2015)].

Moreover, we also establish an adaptive version of Lemma 4.3. Namely, consider the prior distribution $k \sim \pi$ and conditioning on k, $\sqrt{n}\theta_j \sim g$ i.i.d. for $1 \leq j \leq k$ and $\theta_j = 0$ for $j > k$.

THEOREM 4.4. *Assume* $\max_j \frac{\pi(j)}{\pi(j-1)} \leq c$, $-\log \pi(n^{1/(2\alpha+1)}) \leq Cn^{1/(2\alpha+1)}$, $|\log g(x) - \log g(y)| \leq C(1 + |x - y|)$ *and* $|\log g(0)| \leq C$ *for some constants* $c \in (0, 1)$ *and* $C > 0$. *Then, for Gaussian sequence model with any* $\theta_0 \in E_\alpha(Q)$, *we have*

$$\begin{aligned} &P_{f_0}^{(n)} \Pi\bigl(k > Mn^{1/(2\alpha+1)} | X^n\bigr) \leq \exp\bigl(-C' n\varepsilon_n^2\bigr), \\ &P_{f_0}^{(n)} \Pi\bigl(\|\theta - \theta_0\|^2 \geq M\varepsilon_n^2 | X^n\bigr) \leq \exp\bigl(-C' n\varepsilon_n^2\bigr), \end{aligned} \tag{22}$$

for some constants $M, C' > 0$.

The assumption on the prior distribution in Theorem 4.4 is mild. For example, we may choose $\pi(j) \propto e^{-Dj}$ for some constant $D > 0$ and choose g to be the double exponential density. The resulting posterior distribution contracts to the true signal at the minimax rate adaptively for all $\alpha > 0$. The success of this prior crucially depends on the result (22), which allows us to establish an optimal testing procedure on the set $J \leq Mn^{1/(2\alpha+1)}$. However, the proof of (22) takes advantage of the independence structure of the Gaussian sequence model and we are not able to establish (22) for other models. For the same reason, the block spike and slab prior proposed in Hoffmann, Rousseau and Schmidt-Hieber (2015) works only for the Gaussian sequence model as well. Their argument in establishing (22) also uses the independence structure of Gaussian sequence model and thus does not work in other settings.

To summarize, the sieve prior is an example showing that the current proof technique may result in the sub-optimal posterior convergence rate, while for Gaussian sequence model, special techniques can be used to overcome the difficulty.

4.3.3. *The block prior overcomes both difficulties.* The above discussion leads to two fundamental questions. 1. Is there a prior which can achieve the exact minimax posterior convergence rate without knowing α? 2. Can the prior mass and testing proof technique handle a minimax optimal adaptive prior? While the importance of the first question is evident, the second question seems not that relevant at first thought. However, the prior mass and testing method has a great advantage that it is not specific to the choice of the prior or the form of the model. Though we use direct calculation to show the optimal posterior convergence in Lemma 4.3 and Theorem 4.4, the same proof cannot be extended to a setting beyond Gaussian sequence model. The independence structure of Gaussian sequence model plays an important role in the proof. In contrast, the prior mass and testing method is very general so that it can be applied in various settings.

The block prior provides affirmative answers to both questions. Not only can it achieve the exact minimax rate, its proof also relies on the prior mass and testing method, which makes it easy to apply in many complex settings beyond Gaussian sequence model. We provide various examples in Section 3 including regression, density estimation and spectral density estimation to illustrate the benefit of using the prior mass and testing method. Without the prior mass and testing method, an adaptive prior cannot be easily extended to the case beyond Gaussian sequence model.

In fact, inequality (22) can be written as

$$P_{f_0}^{(n)}\Pi(\mathcal{F}_n^c|X^n) \le \exp(-C'n\varepsilon_n^2), \tag{23}$$

where $\mathcal{F}_n$ can be of a more general form than that in (22) as long as an optimal testing procedure can be established in $\mathcal{F}_n$. Then both the sieve prior and the block spike and slab prior in Hoffmann, Rousseau and Schmidt-Hieber (2015) satisfy (23). In contrast, the block prior proposed in this paper satisfies

$$\Pi(\mathcal{F}_n^c) \le \exp(-C'n\varepsilon_n^2), \tag{24}$$

which is one of the three conditions required by the prior mass and testing technique. It can be shown that generally (24) is a stronger condition than (23) in the sense that (24) combining the prior mass lower bound imply (23). In this sense, the block prior in this paper is a stronger prior than the sieve prior and the block spike and slab prior in Hoffmann, Rousseau and Schmidt-Hieber (2015). To put it in another way, (23) is not only a condition on the prior distribution, it is also a condition on the likelihood, which imposes certain model structure. On the other hand, (24) is a condition only on the prior. This is why it works in various models besides the Gaussian sequence model.

5. Proofs of main results.

5.1. *Proof of Theorem* 2.1. We first outline the proof and list some preparatory lemmas, and then state the proof in detail. We introduce the notation $\bar{\Pi}^A$ to be defined as

$$\bar{\Pi}^A = \bigotimes_{k=1}^{\infty} N(0, A_k I_{n_k}). \tag{25}$$

Given a scale sequence $A = \{A_k\}$, the random function $f = \sum_j \theta_j \phi_j$ is distributed by $\bar{\Pi}^A$ if for each block B_k, $\boldsymbol{\theta}_k = \{\theta_j\}_{j \in B_k} \sim N(0, A_k I_{n_k})$. Then $\bar{\Pi}^A$ is a Gaussian process for a given A, and the block prior is a mixture of Gaussian process with A distributed by the mixing densities $\{g_k\} \in \mathcal{G}$.

Since $\bar{\Pi}$ itself is not a Gaussian process, the result for the l^2 small ball probability asymptotics for Gaussian process cannot be applied directly. Our strategy is to pick a collection V_α, and by conditioning, we have

$$\bar{\Pi}(\cdot) \ge \mathbb{P}(V_\alpha)\mathbb{E}(\bar{\Pi}^A(\cdot)|A \in V_\alpha). \tag{26}$$

Then as long as for each $A \in V_\alpha$, there is constants $C_1, C_2 > 0$ independent of A, such that

$$\bar{\Pi}^A\left\{\sum_{j=1}^{\infty}(\theta_j - \theta_{0j})^2 \leq \varepsilon_n^2\right\} \geq \exp\bigl(-C_1 n\varepsilon_n^2\bigr), \tag{27}$$

and

$$\mathbb{P}(V_\alpha) \geq \exp\bigl(-C_2 n\varepsilon_n^2\bigr), \tag{28}$$

then the property (6) is a direct consequence with $C = C_1 + C_2$. Thus, picking such V_α is important. Generally speaking, for each $A \in V_\alpha$, we need $\bar{\Pi}^A$ to behave just like a Gaussian prior designed for estimating $f_0 \in E_\alpha(Q)$ when α is known.

The distribution $\bar{\Pi}^A$ may be hard to deal with. Our strategy is to use the following simple comparison result so that we can study a simpler distribution instead. The lemma will be proved in the supplementary material [Gao and Zhou (2015)].

LEMMA 5.1. *For standard i.i.d. Gaussian sequence $\{Z_j\}$ and sequences $\{a_j\}$, $\{b_j\}$ and $\{c_j\}$, suppose there is a constant $R > 0$ such that*

$$R^{-1}a_j \leq b_j \leq Ra_j \qquad \textit{for all } j,$$

then we have

$$\begin{aligned}\mathbb{P}\Bigl(\sum_j b_j(Z_j - c_j)^2 \leq R^{-1}\varepsilon^2\Bigr) &\leq \mathbb{P}\Bigl(\sum_j a_j(Z_j - c_j)^2 \leq \varepsilon^2\Bigr)\\ &\leq \mathbb{P}\Bigl(\sum_j b_j(Z_j - c_j)^2 \leq R\varepsilon^2\Bigr).\end{aligned}$$

Define J_α to be the smallest integer such that $J_\alpha \geq (8Q^2)^{1/(2\alpha)}n^{1/(2\alpha+1)}$. Let K to be the smallest integer such that $e^K > J_\alpha$, and define $J = [e^K]$. Inspired by the comparison lemma, we define

$$V_\alpha = V_{\alpha,R} = \left\{A : R^{-1} \leq \min_{1\leq k\leq K}\frac{A_k}{A_{\alpha,k}} \leq \max_{1\leq k\leq K}\frac{A_k}{A_{\alpha,k}} \leq R\right\}, \tag{29}$$

with

$$A_{\alpha,k} = \frac{l_k^{-2\alpha} - l_{k+1}^{-2\alpha}}{2\alpha(l_{k+1} - l_k)} \qquad \text{for } k = 1, 2, \dots, K.$$

Define the truncated Gaussian process,

$$\bar{\Pi}_K^{A_\alpha} = \bigotimes_{k=1}^{K} N(0, A_{\alpha,k}I_{n_k}). \tag{30}$$

A random function $f = \sum_j \theta_j \phi_j$ is distributed by $\bar{\Pi}_K^{A_\alpha}$ if $\boldsymbol{\theta}_k \sim N(0, A_{\alpha,k} I_{n_k})$ for each $k = 1, \dots, K$ and $\boldsymbol{\theta}_k = 0$ for $k > K$. The comparison lemma implies that we can control $\bar{\Pi}^A$ for each $A \in V_\alpha$ by the truncated Gaussian process $\bar{\Pi}_K^{A_\alpha}$. Additionally, the small ball probability of $\bar{\Pi}_K^{A_\alpha}$ can be established. The argument is separated in the following lemmas, which will be proved in later sections.

LEMMA 5.2. *For any $\alpha > 0$, and $f_0 \in E_\alpha(Q)$, there exists $C_3 > 0$, such that*

$$\bar{\Pi}_K^{A_\alpha}\left\{\sum_{j=1}^{\infty}(\theta_j - \theta_{0j})^2 \le \varepsilon_n^2\right\} \ge \exp(-C_3 n\varepsilon_n^2).$$

LEMMA 5.3. *For each k, let $A_k \sim g_k$, with $\{g_k\} \in \mathcal{G}$, we have*

$$\mathbb{P}(V_\alpha) \ge \exp(-C_2 n\varepsilon_n^2).$$

LEMMA 5.4. *For J defined above, and $f_0 \in E_\alpha(Q)$, we have*

$$\bar{\Pi}\left\{\sum_{j>J}(\theta_j - \theta_{0j})^2 \le \frac{\varepsilon_n^2}{2}\right\} \ge \frac{1}{2},$$

for sufficiently large n.

PROOF OF (6) IN THEOREM 2.1. We first introduce the truncated version of $\bar{\Pi}^A$ to be

$$\bar{\Pi}_K^A = \bigotimes_{k=1}^{K} N(0, A_k I_{n_k}).$$

By Lemma 5.4, we have

$$\begin{aligned}
&\bar{\Pi}\left\{\sum_{j=1}^{\infty}(\theta_j - \theta_{0j})^2 \le \varepsilon_n^2\right\}\\
&\quad \ge \bar{\Pi}\left\{\sum_{j=1}^{J}(\theta_j - \theta_{0j})^2 \le \frac{\varepsilon_n^2}{2}, \sum_{j>J}(\theta_j - \theta_{0j})^2 \le \frac{\varepsilon_n^2}{2}\right\}\\
&\quad = \bar{\Pi}\left\{\sum_{j=1}^{J}(\theta_j - \theta_{0j})^2 \le \frac{\varepsilon_n^2}{2}\right\}\bar{\Pi}\left\{\sum_{j>J}(\theta_j - \theta_{0j})^2 \le \frac{\varepsilon_n^2}{2}\right\}\\
&\quad \ge \frac{1}{2}\bar{\Pi}\left\{\sum_{j=1}^{J}(\theta_j - \theta_{0j})^2 \le \frac{\varepsilon_n^2}{2}\right\},
\end{aligned}$$

where we have used independence between different blocks in the above equality. In the spirit of (26), we have

$$\begin{aligned}&\bar{\Pi}\left\{\sum_{j=1}^{J}(\theta_j-\theta_{0j})^2\leq\frac{\varepsilon_n^2}{2}\right\}\\ &\qquad\geq\mathbb{P}(V_\alpha)\mathbb{E}\left(\bar{\Pi}_K^A\left\{\sum_{j=1}^{\infty}(\theta_j-\theta_{0j})^2\leq\frac{\varepsilon_n^2}{2}\right\}\middle|A\in V_\alpha\right).\end{aligned}\tag{31}$$

By Lemma 5.1, for each $A\in V_\alpha$,

$$\bar{\Pi}_K^A\left\{\sum_{j=1}^{\infty}(\theta_j-\theta_{0j})^2\leq\frac{\varepsilon_n^2}{2}\right\}\geq\bar{\Pi}_K^{A_\alpha}\left\{\sum_{j=1}^{\infty}(\theta_j-\theta_{0j})^2\leq\frac{\varepsilon_n^2}{2R}\right\}.$$

By Lemma 5.2, we have

$$\bar{\Pi}_K^{A_\alpha}\left\{\sum_{j=1}^{\infty}(\theta_j-\theta_{0j})^2\leq\frac{\varepsilon_n^2}{2R}\right\}\geq\exp(-C'n\varepsilon_n^2).$$

Combining what we have derived and Lemma 5.3, (6) is proved. □

PROOF OF (7) IN THEOREM 2.1. We fix the constant C in (6), and we are going to prove (7) with the same C. Remember the sieve $\mathcal{F}_n$ is defined by (5). Define the set

$$\mathcal{A}_n=\left\{A_k\leq e^{-k^2}\text{ for all }k>\frac{1}{2\alpha+1}\log(n\beta^{-1})\right\}.$$

Then

$$\bar{\Pi}(\mathcal{F}_n^c)\leq\sup_{A\in\mathcal{A}_n}\bar{\Pi}^A(\mathcal{F}_n^c)+\mathbb{P}(\mathcal{A}_n^c).$$

Condition (4) implies

$$\begin{aligned}\mathbb{P}(\mathcal{A}_n^c)&\leq\sum_{k>(2\alpha+1)^{-1}\log(n\beta^{-1})}\mathbb{P}(A_k>e^{-k^2})\\ &\leq\sum_{k>(2\alpha+1)^{-1}\log(n\beta^{-1})}\exp(-c_3e^k)\\ &\leq\exp\left(-\frac{1}{2}c_3n^{1/(2\alpha+1)}\beta^{-1/(2\alpha+1)}\right)\\ &\leq\exp(-(C+4)n\varepsilon_n^2).\end{aligned}$$

The last inequality is because $\beta \le (\frac{c_3}{2(C+4)})^{2\alpha+1}$. We bound $\bar{\Pi}^A(\mathcal{F}_n^c)$ for each $A \in \mathcal{A}_n$,

$$\begin{aligned}
\bar{\Pi}^A(\mathcal{F}_n^c) &= \bar{\Pi}^A\Biggl\{\sum_{j>(n\beta^{-1})^{1/(2\alpha+1)}} (\theta_j - \theta_{0j})^2 > \varepsilon_n^2\Biggr\} \\
&\le \bar{\Pi}^A\Biggl\{2 \sum_{j>(n\beta^{-1})^{1/(2\alpha+1)}} \theta_j^2 + 2 \sum_{j>(n\beta^{-1})^{1/(2\alpha+1)}} \theta_{0j}^2 > \varepsilon_n^2\Biggr\} \\
&\le \bar{\Pi}^A\Biggl\{\sum_{j>(n\beta^{-1})^{1/(2\alpha+1)}} \theta_j^2 \ge \frac{1}{4}\varepsilon_n^2\Biggr\} \qquad (32) \\
&\le \bar{\Pi}^A\Biggl\{\sum_{k>(2\alpha+1)^{-1}\log(n\beta^{-1})} \|\boldsymbol{\theta}_k\|^2 \ge \frac{1}{4}\varepsilon_n^2\Biggr\} \\
&\le \sum_{k>(2\alpha+1)^{-1}\log(n\beta^{-1})} \bar{\Pi}^A\{\|\boldsymbol{\theta}_k\|^2 \ge a_k\varepsilon_n^2\},
\end{aligned}$$

where $\sum_k a_k \le 1/4$ and we choose $a_k = ak^{-2}$. The inequality (32) is because $\theta_0 \in E_\alpha(Q)$ and $\beta \le (4Q^2)^{-(2\alpha+1)/(2\alpha)}$. Define χ_d^2 to be the chi-square random variable with degree of freedom d:

$$\begin{aligned}
&\sum_{k>(2\alpha+1)^{-1}\log(n\beta^{-1})} \bar{\Pi}^A\{\|\boldsymbol{\theta}_k\|^2 \ge a_k\varepsilon_n^2\} \\
&\quad = \sum_{k>(2\alpha+1)^{-1}\log(n\beta^{-1})} \mathbb{P}\{a_k^{-1}A_k\chi_{n_k}^2 \ge \varepsilon_n^2\} \\
&\quad = \sum_{k>(2\alpha+1)^{-1}\log(n\beta^{-1})} \mathbb{P}\{\varepsilon_n^{-2}C'e^k a_k^{-1}A_k\chi_{n_k}^2 \ge C'e^k\} \\
&\quad \le \sum_{k>(2\alpha+1)^{-1}\log(n\beta^{-1})} \exp(-C'e^k)(1 - 2\varepsilon_n^{-2}C'e^k a_k^{-1}A_k)^{-n_k/2},
\end{aligned}$$

where we can choose C' sufficiently large. On the set $\mathcal{A}_k$, for n sufficiently large,

$$A_k \le e^{-k^2} \le \frac{1}{4C'}a_k e^{-k}\varepsilon_n^2 \qquad \text{for all } k > \frac{1}{2\alpha+1}\log(n\beta^{-1}).$$

Therefore,

$$\begin{aligned}
&\sum_{k>(2\alpha+1)^{-1}\log(n\beta^{-1})} \exp(-C'e^k)(1 - 2\varepsilon_n^{-2}C'e^k a_k^{-1}A_k)^{-n_k/2} \\
&\quad \le \sum_{k>(2\alpha+1)^{-1}\log(n\beta^{-1})} \exp(-C'e^k)(\sqrt{2})^{n_k}
\end{aligned}$$

$$\begin{aligned}
&\le \sum_{k>(2\alpha+1)^{-1}\log(n\beta^{-1})} \exp\left(-\left(C'-\frac{1}{2}\log 2\right)e^k\right)\\
&\le \exp\left(-\frac{1}{2}\left(C'-\frac{1}{2}\log 2\right)\beta^{-1/(2\alpha+1)}n\varepsilon^2\right)\\
&\le \exp(-(C+4)n\varepsilon_n^2),
\end{aligned}$$

with sufficiently large C' and n. Hence,

$$\sup_{A\in\mathcal{A}_n} \bar{\Pi}^A(\mathcal{F}_n^c) \le \exp(-(C+4)n\varepsilon_n^2),$$

and we have

$$\Pi(\mathcal{F}_n^c) \le 2\exp(-(C+4)n\varepsilon_n^2).$$

Thus, the proof is complete. □

5.2. *Proof of Theorem* 2.2. Before stating the proof of Theorem 2.2, we need to establish a testing result. It will be proved in later sections.

LEMMA 5.5. *Let d be a distance satisfying the testing property* (10) *and* (11). *Suppose that there is $b>0$ such that for all $f_1, f_2 \in D$,*

$$b^{-1}d(f_1,f_2) \le \|f_1-f_2\| \le bd(f_1,f_2).$$

Then for any sufficiently large $M>0$, there exists a testing function ϕ_n, such that

$$\begin{aligned}
P_{f_0}^{(n)}\phi_n &\le 2\exp\left(-\tfrac{1}{2}LM^2n\varepsilon_n^2\right),\\
\sup_{\{f\in\mathcal{F}_n\cap\operatorname{supp}(\Pi):d(f,f_0)>M\varepsilon_n\}} P_f^{(n)}(1-\phi_n) &\le \exp(-L^2n\varepsilon_n^2).
\end{aligned}$$

The following result is Lemma 10 in Ghosal and van der Vaart (2007). It lower bounds the denominator of the posterior distribution in probability.

LEMMA 5.6. *Consider $\mathcal{H}_n$ defined in* (17), *as long as*

$$\Pi\{D(P_{f_0}^{(n)},P_f^{(n)}) \le bn\varepsilon_n^2, V(P_{f_0}^{(n)},P_f^{(n)}) \le bn\varepsilon_n^2\} \ge \exp(-Cn\varepsilon_n^2),$$

we have $P_{f_0}^{(n)}(\mathcal{H}_n^c) \le \frac{1}{\bar{C}^2n\varepsilon_n^2}$ for some $\bar{C}>0$.

PROOF OF THEOREM 2.2. Notice the prior Π inherits the properties (6) and (7) from $\bar{\Pi}$. Since both $D(P_{f_0}^{(n)},P_f^{(n)})$ and $V(P_{f_0}^{(n)},P_f^{(n)})$ are upper bounded by $bn\|\theta_0-\theta\|^2$, we have

$$\begin{aligned}
&\Pi\{D(P_{f_0}^{(n)},P_f^{(n)}) \le bn\varepsilon_n^2, V(P_{f_0}^{(n)},P_f^{(n)}) \le bn\varepsilon_n^2\}\\
&\qquad \ge \Pi\left\{\sum_{j=1}^{\infty}(\theta_j-\theta_{0j})^2 \le \varepsilon_n^2\right\} \ge \exp(-Cn\varepsilon_n^2),
\end{aligned}$$

for the constant C with which Π satisfies (6) and (7). By Lemma 5.6, the K–L property of prior implies $P_{f_0}^{(n)}(\mathcal{H}_n^c) \leq \frac{1}{\bar{C}^2 n\varepsilon_n^2}$. Let $\mathcal{F}_n$ be the sieve defined in (5) and we have $\Pi(\mathcal{F}_n^c) \leq 2\exp(-(C+4)n\varepsilon_n^2)$. Letting ϕ_n be the testing function in Lemma 5.5, we have $P_{f_0}^{(n)}\Pi(d(f, f_0) > M\varepsilon_n|X^n) \leq P_{f_0}^{(n)}(\mathcal{H}_n^c) + P_{f_0}^{(n)}\phi_n + P_{f_0}^{(n)}\Pi(d(f, f_0) > M\varepsilon_n|X^n)(1-\phi_n)1_{\mathcal{H}_n}$, where the first two terms go to 0. The last term has bound

$$
\begin{aligned}
&P_{f_0}^{(n)}\Pi\big(d(f, f_0) > M\varepsilon_n|X^n\big)(1-\phi_n)1_{\mathcal{H}_n}\\
&\quad\leq \exp\big((C+2)n\varepsilon_n^2\big)P_{f_0}^{(n)}\int_{\{f\in\mathcal{F}_n:d(f,f_0)>M\varepsilon_n\}}\frac{p_f^{(n)}}{p_{f_0}^{(n)}}(X^n)(1-\phi_n)\big(X^n\big)\,d\Pi(f)\\
&\qquad+\exp\big((C+2)n\varepsilon_n^2\big)P_{f_0}^{(n)}\int_{\mathcal{F}_n^c}\frac{p_f^{(n)}}{p_{f_0}^{(n)}}(X^n)\,d\Pi(f)\\
&\quad\leq \exp\big((C+2)n\varepsilon_n^2\big)\int_{\{f\in\mathcal{F}_n:d(f,f_0)>M\varepsilon_n\}}P_{f_0}^{(n)}\frac{p_f^{(n)}}{p_{f_0}^{(n)}}(X^n)(1-\phi_n)\big(X^n\big)\,d\Pi(f)\\
&\qquad+\exp\big((C+2)n\varepsilon_n^2\big)\int_{\mathcal{F}_n^c}P_{f_0}^{(n)}\frac{p_f^{(n)}}{p_{f_0}^{(n)}}(X^n)\,d\Pi(f)\\
&\quad\leq \exp\big((C+2)n\varepsilon_n^2\big)\sup_{\{f\in\mathcal{F}_n\cap\operatorname{supp}(\Pi):d(f,f_0)>M\varepsilon_n\}}P_f^{(n)}(1-\phi_n)\\
&\qquad+\exp\big((C+2)n\varepsilon_n^2\big)\Pi\big(\mathcal{F}_n^c\big)\\
&\quad\leq \exp\big(-\big(LM^2-C-2\big)n\varepsilon_n^2\big)+2\exp\big(-2n\varepsilon_n^2\big).
\end{aligned}
$$

We pick M satisfying $M > \sqrt{L^{-1}(C+2)}$, and then every term goes to 0. The proof is complete. □

Acknowledgments. We want to thank Andrew Barron for the helpful discussion on this topic, and wish to thank the Associate Editor and the referees for their constructive comments and suggestions that lead to the improvement of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Rate exact Bayesian adaptation with modified block priors" (DOI: 10.1214/15-AOS1368SUPP; .pdf). The supplementary material [Gao and Zhou (2015)] contains the remaining proofs and numerical studies of the block prior.

REFERENCES

BARRON, A. R. (1988). *The Exponential Convergence of Posterior Probabilities with Implications for Bayes Estimators of Density Functions*. Univ. of Illinois, Champaign.

BARRON, A. R. (1989). Uniformly powerful goodness of fit tests. *Ann. Statist.* **17** 107–124. MR0981439

BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718

BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398. MR1425958

BROWN, L. D., CAI, T. T., LOW, M. G. and ZHANG, C.-H. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.* **30** 688–707. MR1922538

CASTILLO, I., KERKYACHARIAN, G. and PICARD, D. (2014). Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields* **158** 665–710. MR3176362

DE JONGE, R. and VAN ZANTEN, J. H. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Ann. Statist.* **38** 3300–3320. MR2766853

GAO, C. and ZHOU, H. H. (2015). Supplement to "Rate exact Bayesian adaptation with modified block priors." DOI:10.1214/15-AOS1368SUPP.

GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

GHOSAL, S., LEMBER, J. and VAN DER VAART, A. (2008). Nonparametric Bayesian model selection and averaging. *Electron. J. Stat.* **2** 63–89. MR2386086

GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-I.I.d. observations. *Ann. Statist.* **35** 192–223. MR2332274

GOLUBEV, G. K., NUSSBAUM, M. and ZHOU, H. H. (2010). Asymptotic equivalence of spectral density estimation and Gaussian white noise. *Ann. Statist.* **38** 181–214. MR2589320

HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **43** 2259–2295. MR3396985

KRUIJER, W., ROUSSEAU, J. and VAN DER VAART, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.* **4** 1225–1257. MR2735885

KRUIJER, W. and VAN DER VAART, A. (2008). Posterior convergence rates for Dirichlet mixtures of beta densities. *J. Statist. Plann. Inference* **138** 1981–1992. MR2406419

LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. MR0334381

NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430. MR1425959

RIVOIRARD, V. and ROUSSEAU, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Anal.* **7** 311–333. MR2934953

ROUSSEAU, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.* **38** 146–180. MR2589319

SCHWARTZ, L. (1965). On Bayes procedures. *Probab. Theory Related Fields* **4** 10–26.

SCRICCIOLO, C. (2006). Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *Ann. Statist.* **34** 2897–2920. MR2329472

SHEN, W., TOKDAR, S. T. and GHOSAL, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. MR3094441

SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337

VAN DER VAART, A. and VAN ZANTEN, H. (2007). Bayesian inference with rescaled Gaussian process priors. *Electron. J. Stat.* **1** 433–448. MR2357712

VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442

DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06511
USA
E-MAIL: chao.gao@yale.edu
huibin.zhou@yale.edu