Uniform and Locally Adaptive Quasi-Bayesian Inference for Conditional Moment Models

Hirotaka Ito*
Universitat Pompeu Fabra

First version: January 20, 2025, This version: April 23, 2025

Abstract

This paper proposes a quasi-Bayesian method to construct confidence bands for a nonparametric function in conditional moment models. Importantly, we allow the true function to possess inhomogeneous smoothness over its domain, reflecting the locally varying complexity of the function of interest. Our confidence bands are constructed from quasi-Bayes posteriors with sieve GMM criterion functions and spike-and-slab priors. We show that the quasi-Bayes posteriors contract to the true function in the supremum norm at a rate that adapts to the unknown local smoothness. Based on this result, we derive a nonparametric Bernstein–von Mises-type result for the quasi-Bayes posteriors, in which the well-known covariance mismatch is avoided through the use of GMM criterion functions. The resulting confidence bands are shown to achieve exact asymptotic coverage, while their widths, measured in the supremum norm, shrink at a nearly locally adaptive rate.

Keywords: uniform and locally adaptive inference, quasi-Bayes, conditional moment.
JEL Classification Codes: C10, C14

*Department of Economics and Business, Universitat Pompeu Fabra, Spain. Address: Ramon Trias Fargas, 25-27, 08005 Barcelona, Spain. Email: hirotaka.ito@upf.edu.

I am deeply grateful to Geert Mesters for his support and guidance. I sincerely thank James A. Duffy, Kirill S. Evdokimov, Toru Kitagawa, Soonwoo Kwon, Andriy Norets, and participants at BSE Summer Forum 2024: Macroeconometrics and Policy Evaluation for their helpful comments and discussions. All errors are mine.

1. Introduction

Consider the conditional moment condition

$$\mathbb{E}\left[m(h_0, Z_i)|X_i\right] = 0 \ , \qquad Z_i \coloneqq (Y_i, X_i)^T \ , \tag{1}$$

where $Y_i \in \mathcal{Y} \subset \mathbb{R}$ is an endogenous variable, $X_i \in \mathcal{X} = [0,1]$ is an exogenous variable, and $m(.,.) : \mathcal{C} \times \mathbb{R}^2 \mapsto \mathbb{R}$ is a known moment function on the product space of $\mathbb{R}^2$ and a certain function class $\mathcal{C}$, which is properly defined later. Our interest lies in the structural function h_0 defined by the conditional moment condition in (1). Models like (1) have been extensively studied by Ai and Chen (2003), Chen and Pouzo (2012), and the references therein, including nonparametric regression and nonparametric quantile regression as special cases [1].

A commonly used assumption for (1) is that the smoothness of the function h_0 is deemed constant on the domain. For many economic applications, however, this assumption can be too strong and the smoothness of the function can indeed vary over the domain. In such cases, fixing the degree of smoothness to be uniform over the domain can be harmful and lead to both under- and over-coverage when standard confidence bands are used for h_0.

To overcome this problem, in this paper we relax the homogeneity assumption and develop an inference scheme that is valid under the inhomogeneity of smoothness. Specifically, the function h_0 on $[0,1]$ is assumed to be in a *point-wise* Hölder class, where the smoothness is modeled as a function on the domain (e.g. Gach, Nickl, and Spokoiny, 2013; Ročková and Rousseau, 2023). For this setting, we develop a nonparametric quasi-Bayesian framework that allows us to conduct uniform inference for h_0 while taking into account its inhomogeneity of smoothness. We show that our quasi-Bayes posterior contracts to h_0 measured in the supremum-norm by the "locally adaptive" rate, i.e., its contraction rate successfully incorporates the local smoothness of h_0. Importantly, in the case of nonparametric regression or nonparametric quantile regression, this rate corresponds to the minimax optimal rate in Stone (1982) and Chaudhuri (1991). Moreover, we construct confidence bands for h_0 from the quasi-Bayes posterior and study its asymptotic properties. Our results show that, under a certain restricted class of functions (self-similar class), our confidence bands are uniformly valid over $[0,1]$ *as well as* nearly adaptive to the "local" smoothness. Such bands provide a coherent statistical measure for recovering various properties of the targeted function h_0.

Our methodology is based on the recent nonparametric Bayesian literature, where posteriors with hierarchical priors (Spike-and-Slab priors) are shown to achieve adaptive rates of contraction in the sup-norm sense (e.g. Hoffmann, Rousseau, and Schmidt-Hieber, 2015; Yoo and Ghosal, 2016). We extend and depart from this literature by studying nonparametric *quasi-Bayesian* inference to construct confidence bands for the

[1] There are two remarks on the model (1). First, assuming that X_i is a scalar is merely for simplicity, and we can allow $X_i \in \mathbb{R}^{d_x}$ with $1 \leq d_x < \infty$ in our analysis without changing the main results. Second, we can include a finite-dimensional parameter, say $\theta \in \mathbb{R}^{d_\theta}$ with $1 \leq d_\theta < \infty$, in the model (1) and make it semiparametric one. As noted in Chen and Pouzo (2012), the main difficulty in that case is still not the presence of θ; thus, we simply focus on the nonparametric conditional moment model as it is in (1).

targeted function. This departure is necessary because the conditional moment model (1) does not define a full probabilistic likelihood. Instead, we follow Chernozhukov and Hong (2003) and use an approximate likelihood that is determined by taking the negative exponent of the J-statistic of Hansen (1982).

The quasi-Bayesian approach is particularly useful in econometric applications where (i) there could be a large number of nuisance parameters, which makes it difficult to build (correct) probabilistic models, and/or (ii) (nonparametric) parameters of our interest are characterized simply as the solutions of some loss function based on moment conditions.

We consider a hierarchical prior for the function h_0 based on an accurate sieve approximation for h_0 with Haar wavelet basis functions. The coefficients of the sieve approximation are equipped with flexible spike-and-slab priors. The crucial point is that the construction of prior does not require knowledge of the smoothness or regularity of h_0 a priori, making our inference scheme adaptive to the unknown quantity of targeted functions.

Combining the quasi-likelihood with the prior allows us to compute the quasi-Bayes posterior. For this posterior, we establish three main results. First, we show that the quasi-Bayesian posterior successfully achieves local adaptability in a sense that it contracts to the true function h_0 by the rate reflecting its local smoothness. The key idea behind this result is a carefully constructed hierarchical prior, which plays a type of regularization on sieve coefficients.

Second, we establish the distributional result on our quasi-Bayes posterior. Specifically, the quasi-Bayes posterior on the approximating Haar wavelet coefficient is shown to weakly converge to a suitable Gaussian process, centered at a certain mean function and a covariance function, whose marginals correspond to the sieve GMM estimator. This is a quasi-Bayesian version of the so-called nonparametric Bernstein-von Mises theorem in Castillo and Nickl (2013) and Castillo and Nickl (2014). A notable feature of our result is that we avoid the well-known covariance discrepancy, which arises due to the usage of misspecified likelihood (e.g. Kleijn and van der Vaart (2012)); this avoidance is achieved by formulating our loss function as properly weighted GMM criterion function.

Third, based on the weak convergence result, we construct the confidence band that satisfies the exact asymptotic coverage property and whose diameter shrinks at a nearly optimal rate. The key point is that our confidence band is constructed by intersecting two bands; one that controls low frequencies of the estimation error and the one that controls the high frequency counterpart. This approach is used in the literature of nonparametric Bayes in the case of homogeneous smoothness of h_0 (e.g. Castillo and Nickl, 2014; Ray, 2017), and we employ this method in our generalized setting.

Ideally, constructed confidence bands would satisfy both honesty and adaptivity, i.e., they would have asymptotically the specified coverage uniformly over a function class (honesty), while their length shrinks at the optimal rate reflecting the unknown smoothness or regularity of targeted function (adaptivity). As well known since Low (1997), it is generally impossible to construct such confidence bands. The literature therefore has tried to find a subclass of functions in which this construction is possible. An often used

assumption is the so-called "self-similarity" in Giné and Nickl (2010), which restricts Hölder function class to its subclass where the Hölder-norm of the functions has a suitable lower bound. Under this assumptions, we show that the length of our confidence band consists of locally adaptive rate plus some penalty term, which makes our confidence band "nearly" optimal. This penalty is unavoidable due to Armstrong (2021). Nonetheless, our approach allows us to construct confidence bands that practically work and are adaptive to local smoothness.

Literature. The construction of confidence bands for nonparametric functions has been widely studied in both the field of statistics and econometrics. (a partial list in the recent econometric literature includes Chernozhukov, Chetverikov, and Kato, 2014; Belloni et al., 2015; Chen and Christensen, 2015; Freyberger and Rai, 2018; Chen, Christensen, and Kankanala, 2025). Our results contribute to this literature by showing desirable properties of our confidence bands for functions possessing inhomogeneous smoothness.

Our approach builds on the quasi-Bayes literature (e.g. Kim, 2002; Chernozhukov and Hong, 2003; Liao and Jiang, 2011; Florens and Simoni, 2012; Kato, 2013). In particular, Liao and Jiang (2011) and Kato (2013) are closely related to ours in that they start from conditional moment restrictions where nonparametric functions are defined and asymptotic properties of quasi-Bayes posteriors are studied. The key difference is that we explicitly model the function class with inhomogeneous smoothness and closely investigate the dependence of the asymptotic properties of the quasi-Bayes posteriors on local smoothness. Note that their framework admits nonparametric instrumental variable models, which is beyond the scope of this paper, and we leave it for future research.

Organization of the paper. The remainder of this paper is organized as follows. The next section describes our modeling framework: the model, the prior, the quasi-likelihood and the posterior. Section 3 discusses the theoretical properties of the posterior. Section 4 discusses the potential extension and direction of future research.

Notation. For any $r \in [1, \infty]$, $\|.\|_r$ denotes the L^r-norm. For any vector $a \in \mathbb{R}^d$, $\|.\|$ denotes the Euclidean norm. For positive sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, let $a_n \lesssim b_n$ if $\limsup_n a_n/b_n < \infty$ and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For any set A, $|A|$ represents the cardinality of A. For a and b in $\mathbb{R}$, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For a vector $u \in \mathbb{R}^d$, let $u^{\otimes 2} = uu^T$ denote the outer product.

2. Modeling framework

In this section, we formally introduce our modeling framework. We start by defining the function class to which the structural function h_0 belongs and explain the sieve basis functions to approximate h_0. Subsequently, we discuss the hierarchical prior for the sieve coefficients and the quasi-likelihood that is based on the moment conditions.

Combining all, we obtain our quasi-posterior, which is the main object of interest in our study.

2.1. **Conditional moment model with varying smoothness.** For convenience we restate the conditional moments

$$\mathbb{E}\left[m(h_0, Z_i)|X_i\right] = 0 \ , \tag{2}$$

where the structural function h_0 is assumed to be in $L^2[0,1]$. The key feature of h_0 in this paper is that the smoothness of h_0 varies over the domain of regressors $\mathcal{X} = [0,1]$, i.e., the smoothness of h_0 is *inhomogeneous*. To formalize this notion, we follow Gach, Nickl, and Spokoiny (2013) and Ročková and Rousseau (2023), and start from the local t-Hölder class. This class is characterized, for a *specific* point $x \in [0,1]$, by

$$\mathcal{C}(t, x, M, \eta) := \left\{ h : [0,1] \to \mathbb{R};\ \|h\|_\infty \vee \sup_{0<|m|\leq\eta} \frac{|h(x+m) - h(x)|}{|m|^t} \leq M \right\} \ ,$$

where $t \leq 1$ [2]. The class $\mathcal{C}(t, x, M, \eta)$ is the set of functions h whose smoothness is governed by t, M, and η at the specific point $x \in [0,1]$. The *uniformly* locally t-Hölder class is then obtained by extending the local conditions to the entire domain $[0,1]$ as follows.

Assumption 2.1. The function h_0 belongs to the *local Hölder class*

$$h_0 \in \mathcal{C}(t, M, \eta) := \{h : [0,1] \to \mathbb{R};\ h \in \mathcal{C}(t, x, M, \eta),\ \forall x \in [0,1]\} \, ,$$

where (i) $M(\cdot)$ is bounded and $\eta(\cdot)$ is bounded away from 0 over $[0,1]$, and (ii) $t(\cdot)$ satisfies $0 < t_1 \leq \inf_{x\in[0,1]} t(x)$ for some $t_1 > 0$.

Remark 2.2. If we set all functions $t(\cdot), M(\cdot)$, and $\eta(\cdot)$ to constants, then the class $\mathcal{C}(t, M, \eta)$ reduces to the standard homogenous Hölder class (e.g. Definition 1.2 in Tsybakov (2009)).

Sieve approximation with Haar wavelets. Let $\left\{\phi_{lk} :\ l \in \mathbb{Z}_+, 0 \leq k < 2^l\right\} \subset L^2[0,1]$ be an orthonormal wavelet basis, where the indexes l and k represent the resolution level and the location respectively. Given such a basis, the function h_0 can be uniquely represented by the triangular array of wavelet coefficients $\left\{\beta^0_{lk};\ l \in \mathbb{Z}_+, 0 \leq k < 2^l\right\}$, where $\beta^0_{lk} = \langle h_0, \phi_{lk}\rangle = \int_0^1 h_0(t)\phi_{lk}(t)dt$.

In this paper, we consider the standard balanced Haar wavelet basis; $\phi_{-10}(x) = \mathbf{1}_{[0,1]}(x)$ and $\phi_{lk}(x) = 2^{l/2}\phi(2^l x - k)$, $l \in \mathbb{Z}_+, 0 \leq k < 2^l$, obtained by orthonormal dilation-translation of the function $\phi(x) = \mathbf{1}_{(0,1/2]}(x) - \mathbf{1}_{(1/2,1]}(x)$. Denote by I_{lk} the dyadic intervals of the domain corresponding to the balanced Haar wavelet basis, i.e., $I_{-10} = [0,1]$ and $I_{lk} = (k/2^l, (k+1)/2^l]$ for $l \geq 0$ and $0 \leq k < 2^l$. Note here that the

[2]Note that this condition on smoothness $t \leq 1$ comes from the usage of Haar wavelet in approximating h_0.

choice of the standard balanced Haar wavelet basis is for simplicity of illustration and is not crucial in the theoretical results of this paper.

Given the Haar wavelet basis $\{\phi_{-10}\} \cup \{\phi_{lk} : l \in \mathbb{Z}_+, 0 \leq k < 2^l\}$, the function h_0 is uniquely represented as

$$h_0 = \beta^0_{-10}\phi_{-10} + \sum_{l=0}^{\infty}\sum_{k=0}^{2^l-1} \beta^0_{lk}\phi_{lk} \ .$$

We approximate h_0 by truncating this representation. Define the linear sieve space

$$\Phi_{l_n} := \text{span}\left\{\phi_{-10}, \phi_{00}, \phi_{10}, ..., \phi_{l_n 0}, ..., \phi_{l_n 2^{l_n}-1};\ |\beta_{lk}| \leq B_n\right\} \ , \tag{3}$$

where l_n is some resolution level such that $l_n < L_{max} := \lfloor \log_2 n \rfloor$, which is a typical maximum resolution level employed in the literature in nonparametric Bayes (e.g. Hoffmann, Rousseau, and Schmidt-Hieber, 2015; Ray, 2017). B_n is any sequence such that $B_n \to \infty$ sufficiently fast, e.g., $B_n \gtrsim n$; this restriction is convenient to make each sieve space compact without affecting our results. Note here that $\dim(\Phi_{l_n}) = 2^{l_n+1}$. Let $P_{l_n} : L^2[0,1] \to \Phi_{l_n}$ be the projection operator. Then, for any $h_{l_n} = \beta^0_{-10}\phi_{-10} + \sum_{l=0}^{l_n}\sum_{k=0}^{2^l-1}\beta^0_{lk}\phi_{lk} \in \Phi_{l_n}$, we have

$$\|h_{l_n} - h_0\|_\infty \leq \|h_{l_n} - P_{l_n}h_0\|_\infty + \|P_{l_n}h_0 - h_0\|_\infty \, , \tag{4}$$

where the first term represents the error of h_{l_n} within the sieve space Φ_{l_n} and the second term is the approximation error of the space Φ_{l_n} to h_0. Intuitively, the first term of the error is affected by the prior on the sieve space while the second term is influenced by the growing speed of the resolution level l_n. Following Castillo and Ročková (2021) and Ročková and Rousseau (2023), we can derive the sufficient condition on l_n that controls the convergence of the sieve approximation error by the locally adaptive rate; see Appendix A for the discussion in detail. Regarding the error on the first term, one of our main results states that our quasi-Bayes posterior constructed on the sieve space contracts at the locally adaptive rate. These two results imply that our posterior constructed in the growing sieve space contracts to h_0 in the sup-norm sense at the locally adaptive rate. Our prior to control the first term of the error in (4) is stated next.

2.2. **Spike-and-Slab priors.** Our prior is constructed on the sieve space Φ_{l_n} in (3) using an hierarchical structure. In particular, we use a spike-and-slab prior similar to Hoffmann, Rousseau, and Schmidt-Hieber (2015) and Ročková and Rousseau (2023).

Let $\Lambda_n := \left\{(l,k) : l < L_{max}, 0 \leq k < 2^l\right\}$ be the set of all nodes (l,k) with resolution level less than L_{max}. First, a probability measure is put on the set Λ_n assigning some mass to each subset $\lambda \subset \Lambda_n$. Conditioning on the prior on λ, the next probability measure is put on the coefficients β_{lk}.

Assumption 2.3. $\Pi = (\Pi_{\Lambda_n}, \pi_{|\lambda})$ is a prior on $(\Lambda_n, \mathbb{R}^{|\Lambda_n|})$ that satisfies: (i) there exist constants $c_T, C_T > 0$ such that

$$c_T w_l \leq \frac{\Pi_{\Lambda_n}(\lambda \cup \{(l,k)\})}{\Pi_{\Lambda_n}(\lambda)} \leq C_T w_l, \quad \forall \lambda \subset \Lambda_n \text{ with } (l,k) \notin \lambda,$$

where a sequence w_l satisfies

$$n^{-K} \leq w_l \leq 2^{-l} \quad \text{for } l < L_{max},$$

for some constant $K > 0$, and (ii) $\pi_{|\lambda} = \bigotimes_{(l,k)\in\Lambda_n} \pi_{(l,k)|\lambda}$, where, conditionally on λ,

$$\beta_{lk} \sim \pi_{(l,k)|\lambda}(dx) = \mathbf{1}\{(l,k) \in \lambda\} q(x)dx + \mathbf{1}\{(l,k) \notin \lambda\}\delta_0(dx),$$

where δ_0 is the Dirac measure at 0, and q is a density on $\mathbb{R}$ such that, for some $R, c_R, C_R > 0$,

$$c_R \leq \inf_{|\beta_{lk}|\leq R} q(\beta_{lk}) \leq \sup_{\beta_{lk}\in\mathbb{R}} q(\beta_{lk}) \leq C_R.$$

Let $\Phi_{\Lambda_n} := \{\phi_{lk}\}_{(l,k)\in\Lambda_n}$, i.e., the sieve space spanned by the set of the Haar wavelet basis indexed by Λ_n. Then the above condition is interpreted that we put the hierarchical prior in the sieve space Φ_{Λ_n}. A simple example of the prior is that (1) each node (l,k) is selected with probability 2^{-l} and (2) conditional on the selection of the node (l,k), the wavelet coefficient has the standard normal prior (q is the density of the standard normal distribution).

2.3. **Quasi-Bayes framework.** The conditional moment model (2) does not characterize a complete parametric likelihood function. As such, we take the quasi-Bayesian approach, where Bayesian updating is conducted through an appropriately scaled loss function (e.g. Chernozhukov and Hong (2003))

To obtain an encompassing loss function — which takes into account all restrictions imposed by the conditional moment condition — we employ a second sieve approximation that transforms the conditional moments into unconditional moments. Specifically, as noted in Chen (2007), the conditional moment condition is equivalent to $\mathbb{E}[m(h_0; Z_i)\phi_{lk}(X_i)] = 0$, where $\{\phi_{lk}\}_{(l,k)}$ are any basis functions approximating $L^2[0,1]$. For convenience, we again use the Haar wavelet basis for the approximation. Let $\phi_{\bar{\Lambda}_n} = (\phi_{lk})_{(l,k)\in\bar{\Lambda}_n}$ be the vector of the Haar wavelet indexed on the set of nodes $\bar{\Lambda}_n := \{(l,k) : l < \bar{l}_n, 0 \leq k < 2^l\}$; the resolution level $\bar{l}_n$ affects the size of the moment-function class and we put some conditions later.

With this set-up, the sieve loss function $Q_n(h_{\Lambda_n})$ is set up on Φ_{Λ_n} so that the true function projected on Φ_{Λ_n}, denoted by $P_{\Lambda_n}h_0$, is characterized as extremum. In particular, we set up $Q_n(h_{\Lambda_n})$ in continuous updating GMM framework with the following form; defining the notation $g(h_{\Lambda_n}; Z_i) = g(h_{\Lambda_n}, \phi_{\bar{\Lambda}_n}; Z_i) := m(h_{\Lambda_n}; Z_i) \otimes \phi_{\bar{\Lambda}_n}(X_i)$, which abbreviates the dependence on $\phi_{\bar{\Lambda}_n}$ for notational simplicity, we have

$$Q_n(h_{\Lambda_n}) = \frac{1}{2} g_n(h_{\Lambda_n})^T W_n(h_{\Lambda_n}) g_n(h_{\Lambda_n}), \tag{5}$$

where $g_n(h_{\Lambda_n}) = n^{-1}\sum_{i=1}^n g(h_{\Lambda_n}; Z_i)$ is the sample analogue of the moment condition, and $W_n(h_{\Lambda_n}) = [n^{-1}\sum_{i=1}^n g(h_{\Lambda_n}; Z_i)^{\otimes 2}]^{-1}$ is a random weighting matrix.

Let Π be the hierarchical prior that satisfies the condition stated before on the space Φ_{Λ_n}. Given the data $Z^n := (Z_1', ..., Z_n')'$, the prior Π, and the sieve GMM criterion

function Q_n on Φ_{Λ_n}, we construct the following posterior,

$$\Pi_n(B|Z^n) = \frac{\int_B \exp\{-nQ_n(h_{\Lambda_n})\}\,\Pi(dh_{\Lambda_n})}{\int_{\Phi_{\Lambda_n}} \exp\{-nQ_n(h_{\Lambda_n})\}\,\Pi(dh_{\Lambda_n})}, \tag{6}$$

for any Borel set $B \subset \Phi_{\Lambda_n}$.

3. Uniform inference adaptive to local smoothness

In this section, we discuss the main theoretical properties of the posterior defined in (6). First, we derive the locally adaptive contraction rate for the quasi-posterior in the sup-norm. Second, we derive a weak Bernstein-von Mises result for the quasi-posterior. Third, we use the weak BvM result to construct locally adaptive credible bands and show its asymptotic properties.

3.1. Locally adaptive contraction rate in supremum-norm.

We establish the contraction of our posterior at the locally adaptive rate. Our main focus here is the sup-norm $\|h_{l_n} - P_{l_n}h_0\|_\infty$, where h_{l_n} and $P_{l_n}h_0$ are in the space Φ_{l_n}. Note that this norm is bounded by

$$\begin{aligned}\|h_{l_n} - P_{l_n}h_0\|_\infty &\le |\beta_{-10} - \beta^0_{-10}| + \sum_{l=0}^{l_n} \max_{0\le k<2^l} |\beta_{lk} - \beta^0_{lk}| \cdot \left\| \sum_{k=0}^{2^l-1} |\phi_{lk}| \right\|_\infty \\ &\le |\beta_{-10} - \beta^0_{-10}| + \sum_{l=0}^{l_n} 2^{l/2} \max_{0\le k<2^l} |\beta_{lk} - \beta^0_{lk}| =: \ell_\infty(h_{l_n}, P_{l_h}h_0).\end{aligned}$$

Hence we work on the ℓ_∞-norm of the projected wavelet coefficients.

We first formalise the following assumption on our data generating process. Assumption 3.1 below is standard regularity assumption in the literature using sieve approximation in nonparametric inference (e.g. Newey (1997)).

Assumption 3.1. Data generating process

(1). $\{Z_i\}_{i=1}^n = \{(Y_i, X_i)^T\}_{i=1}^n$ is an i.i.d. sequence.
(2). The density of X_i is uniformly bounded and bounded away from zero on $\mathcal{X} = [0,1]$.
(3). The smallest eigenvalue of $\mathbb{E}_{h_0}(\phi_{\Lambda_n}(X_i)^{\otimes 2})$ is bounded away from zero uniformly over $h_0 \in \mathcal{C}(t, M, \eta)$ and $\Lambda_n \subset \{(-1,0)\} \cup \{(l,k) :\ l \in \mathbb{Z}_+, 0 \le k < 2^l\}$.

Next we set assumptions on the moment function. To this end, we introduce the following notations. Let

$$\mathcal{H}_n := \{m(h_{\Lambda_n}; .);\ h_{\Lambda_n} \in \Phi_{\Lambda_n}\}$$

be a class of moment functions indexed on the sieve space Φ_{Λ_n}. Recall $\phi_{\bar{\Lambda}_n} = (\phi_{lk})_{(l,k)\in\bar{\Lambda}_n}$. Then the approximated unconditional moment conditions belongs to the class

$$\mathcal{F}_n := \{g(h_{\Lambda_n}, \phi_{\bar{\Lambda}_n}; .) = m(h_{\Lambda_n}; .) \otimes \phi_{\bar{\Lambda}_n}(.);\ m(h_{\Lambda_n}; .) \in \mathcal{H}_n\}.$$

We control the size of $\mathcal{F}_n$ through the bracketing integral

$$J_{[]}(\delta, \mathcal{F}_n, \|.\|_2) := \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{F}_n, \|.\|_2)}\, d\epsilon,$$

where $N_{[]}(\epsilon, \mathcal{F}_n, \|.\|_2)$ is the smallest number of ϵ-brackets w.r.t. $\|.\|_2$ to cover $\mathcal{F}_n$.

Assumption 3.2. Sieve GMM criterion

(1). $\mathbb{E}[m(h; Z_i)|X_i] = 0$ if and only if $h = h_0$.

(2). With $\mathcal{C} = \mathcal{C}(t, M, \eta)$ and a metric space $(\mathcal{C}, \|.\|_{\phi,2})$ where $\|.\|_{\phi_2}$ denotes the L_2-norm induced by the Haar wavelet $\{\phi_{-10}\} \cup \{\phi_{lk} :\ l \in \mathbb{Z}_+, 0 \le k < 2^l\}$, there exist a linear map $G(.;.) : \mathcal{C} \times \mathbb{R}^2 \mapsto \mathbb{R}$ that is continuous on $\mathcal{C}$ (w.r.t. $\|.\|_{\phi,2}$) and $\delta > 0$ that satisfy the following properties:
(i) given $h_0 \in \mathcal{C}$,

$$\begin{aligned}
&\sup_{h\in\mathcal{C}} \|r(h; Z_i)\|_2 \\
&\quad = \sup_{h\in\mathcal{C}} \|g(h; Z_i) - g(h_0; Z_i) - G(h_0; Z_i)(h - h_0)\|_2 \\
&\quad = o(\|h - h_0\|_{\phi,2}), \qquad \text{as } \|h - h_0\|_{\phi,2} \to 0.
\end{aligned}$$

(ii) $\sup_{h_{\Lambda_n}\in\Phi_{\Lambda_n}} \|G_n(h_{\Lambda_n}) - G_0(h_{\Lambda_n})\|_{\text{op}} \xrightarrow{p} 0$, where $G_n(h_{\Lambda_n}) := n^{-1}\sum_{i=1}^n G(h_{\Lambda_n}; Z_i)$, $G_0(h_{\Lambda_n}) := \mathbb{E}G(h_{\Lambda_n}; Z_i)$, and $\|.\|_{\text{op}}$ denotes the operator norm.

(3). uniformly over $h_{\Lambda_n} \in \Phi_{\Lambda_n}$, $W_0(h_{\Lambda_n}) = \mathbb{E}\left[g(h_{\Lambda_n}; Z_i)^{\otimes 2}\right]$ is positive definite and its largest eigenvalue is bounded above.

(4). there exists an envelop F_n for each $\mathcal{F}_n$ such that $\|F_n\|_2 < \infty$, $\mathbb{E}F_n^2\mathbf{1}\{F_n > \delta\sqrt{n}\} \to 0$ for any $\delta > 0$, and $J_{[]}(\|F_n\|_2, \mathcal{F}_n, \|.\|_2) < \infty$.

Assumption 3.2 (1) is an identification assumption. Assumption 3.2 (2)-(i) puts a Lipschitz-type smoothness restriction on the moment function $g(\beta_{\Lambda_n}; Z_i)$. Assumption 3.2 (2)-(ii) is a stochastic equicontinuity condition. Note that Assumption 3.2 (2) on continuity of $g(\beta_{\Lambda_n}; Z_i)$ in β_{Λ_n} implies that the same is true for $W_0(\beta_{\Lambda_n})$ and $W_n(\beta_{\Lambda_n})$. Assumption 3.2 (4) puts restriction on the size of function class $\mathcal{F}_n$, which in turn restricts the resolution level $\bar{l}_n$ used for unconditional moment condition. In particular, Assumption 3.2 (2)-(i) and (4) ensure that the class $\mathcal{F}_n$ is Donsker.

Now one of our main results is built upon the following Lemma 3.3.

Lemma 3.3. *Suppose that Assumptions 2.1, 2.3, 3.1, and 3.2 hold. Then, for any sequence $D_n \to \infty$, we have*

$$\sup_{h_0\in\mathcal{C}(t,M,\eta)} \mathbb{E}_{h_0}\Pi_n\left[h_{\Lambda_n} :\ \sup_{x\in[0,1]} \xi_n(x)^{-1}|h_{\Lambda_n}(x) - P_{\Lambda_n}h_0(x)| > D_n|Z^n\right] = o(1),$$

where $\xi_n(x) = (\log n/n)^{t(x)/(2t(x)+1)}$.

Proof. See Appendix B.1. □

Lemma 3.3 shows that the quasi-Bayes posterior constructed in the sieve space Φ_{Λ_n} contracts to the projected component $P_{\Lambda_n}h_0 \in \Phi_{\Lambda_n}$ of the targeted function h_0 in the sup-norm by the locally adaptive rate. In the cases of nonparametric regression or nonparametric quantile regression, this rate reflects the local smoothness in the known minimax rate of Stone (1982) or Chaudhuri (1991). Combining the result on the vanishing rate of sieve approximation error (Lemma A.2 in Appendix A), we get one of our main results as follows.

Theorem 3.4. *Suppose that the same conditions as in Theorem 3.3 hold. Then, for any sequence $D_n \to \infty$, we have*

$$\sup_{h_0 \in \mathcal{C}(t,M,\eta)} \mathbb{E}_{h_0} \Pi_n \left[h_{\Lambda_n} : \sup_{x \in [0,1]} \xi_n(x)^{-1} |h_{\Lambda_n}(x) - h_0(x)| > D_n | Z^n \right] = o(1).$$

Proof. The result is an immediate consequence of (4), (A.2) in Lemma A.2, and Lemma 3.3. □

Theorem 3.4 states that quasi-Bayes posteriors successfully achieve the minimax rate that is adaptive to the local smoothness $t(x)$. Note that the restriction to our prior in Assumption 2.3 is much weaker than the one used in Hoffmann, Rousseau, and Schmidt-Hieber (2015) while for simplicity we omit the slowly growing bound on the sequence w_l in Ročková and Rousseau (2023). More specifically, Ročková and Rousseau (2023) allows $w_l \leq n^{(1-\delta)/2} 2^{-l}$ for some constant $\delta > 0$. We can incorporate this growing upper bound in our framework and still get the same conclusion as in Theorem 3.4.

3.2. **Weak Bernstein-von Mises phenomena in quasi-Bayes framework.** Our goal is to construct credible bands that are uniformly valid as well as adaptive to local smoothness. To this end, we here establish the nonparametric BvM in the quasi-Bayes framework. Based on this result, we construct credible bands and show their desirable properties in the next subsection.

In a smooth parametric model, BvM tells us that the posterior is well approximated in total variation metric by the Gaussian distribution centered at the efficient estimator with the variance being the inverse of the Fisher information. It is well known since Freedman (1999) that BvM does not always hold in infinite dimensional settings. Here we follow the approach taken by Castillo and Nickl (2014), where BvM is established through a topology weaker than the one induced by the total variation metric.

To set up, we introduce some notations. First, our analysis below is carried out through a weak topology induced by the bounded Lipschitz metric; for any probability

measures P and Q on a metric space $(\mathcal{S}, d)$, define

$$\mathcal{D}_{\mathcal{S}}[P, Q] := \sup_{f:\|f\|_{BL}\leq 1} \left| \int_{\mathcal{S}} f(x) \left(dP(x) - dQ(x) \right) \right|,$$

$$\|f\|_{BL} = \sup_{x\in\mathcal{S}} |f(x)| + \sup_{x\neq y\in\mathcal{S}} \frac{|f(x) - f(y)|}{d(x, y)}.$$

Next, following Castillo and Nickl (2014) we consider multiscale spaces where Haar wavelet coefficients characterizing h_0 are included with appropriately scaled metric. Given a monotone increasing sequence $\omega = (\omega_l)_{l\in\mathbb{N}}$ with $\omega_l \geq 1$ for any $l \in \mathbb{N}$, define

$$\mathcal{M}(\omega) := \left\{ \beta = (\beta_{lk}) : \|\beta\|_{\mathcal{M}(\omega)} := \sup_l \frac{\max_k |\beta_{lk}|}{\omega_l} < \infty \right\},$$

and its separable closed subspace

$$\mathcal{M}_0(\omega) := \left\{ \beta \in \mathcal{M}(\omega) : \lim_{l\to\infty} \frac{\max_k |\beta_{lk}|}{\omega_l} = 0 \right\}.$$

The sequence ω is said to be admissible if $\omega_l/\sqrt{l} \to \infty$ as $l \to \infty$. The admissibility of ω helps to define a proper tight Gaussian measure on the space $\mathcal{M}_0(\omega)$, as illustrated below.

Proposition 3.5. *Let β^* be a random process on $\mathcal{M}(\omega)$ whose law takes a Gaussian form of $\mathcal{N}((\beta^*_{lk}), (V^*_{lk}))$, where (β^*_{lk}) and (V^*_{lk}) are mean and covariance functions over $\mathcal{M}(\omega)$. Let $\sup_{l,k} |\beta^*_{lk}| < \infty$ and $\sup_{l,k} |V^*_{lk}| < \infty$. Then for $\omega = (\omega_l)_{l\in\mathbb{N}} = \sqrt{l}$, we have $\mathbb{E}\|h^*\|_{\mathcal{M}(\omega)} < \infty$. For admissible sequence $\omega = (\omega_l)_{l\in\mathbb{N}}$, the process β^* determines a tight Gaussian Borel probability measure on $\mathcal{M}_0(\omega)$.*

Proof. The results follow from Proposition 2 in Castillo and Nickl (2014) combined with the assumption that mean and variance are uniformly bounded over $\mathcal{M}(\omega)$. □

Finally, we need a slightly different prior from Assumption 2.3. In particular, as in Ray (2017) we use a less sparse prior compared with the one in Assumption 2.3, in that we automatically fit a density to all nodes less than some resolution level. This point is specified as the following Assumption 3.6.

Assumption 3.6. Modified hierarchical prior
Let $\Pi = (\Pi_{\Lambda_n}, \pi_{|\lambda})$ be the same prior as in Assumption 2.3, except the following properties; conditionally on λ,

$$\beta_{lk} \overset{i.i.d.}{\sim} \pi_{(l,k)|\lambda}(dx) = \begin{cases} q(x)dx & l \leq l_0(n), \\ \mathbf{1}\{(l,k) \in \lambda\} q(x)dx + \mathbf{1}\{(l,k) \notin \lambda\}\delta_0(x) & l_0(n) < l < L_{max}, \end{cases}$$

where $l_0(n) \to \infty$.

Given these notations and assumptions, a weak BvM result in our quasi-Bayesian setting is stated as follows.

Theorem 3.7. *Quasi-Bayes weak BvM*
Suppose that Assumptions 2.1, 3.1, 3.2, and 3.6 hold. Then, for any admissible sequence $\omega = (\omega_l)_{l\in\mathbb{N}}$ *such that* $\omega_{l_0(n)} \geq C\sqrt{\log n}$ *for some* $C > 0$*, we have*

$$\sup_{h_0\in\mathcal{C}(t,M,\eta)} \mathbb{E}_{h_0}\mathcal{D}_{\mathcal{M}_0(\omega)}\left[\mathcal{L}(\Pi_n[\cdot\,|Z^n]),\ \mathcal{N}(\beta^*,(nV)^{-1})\right] = o(1),$$

where $\mathcal{L}(\Pi_n[.|Z^n])$ *represents the law of the posterior, and* $\mathcal{N}(.,.)$ *is a Gaussian probability measure on* $\mathcal{M}_0(\omega)$ *centered and scaled by* $\beta^* = (\beta^*_{lk})$ *and* $V = (V_{lk})$ *whose marginals correspond to*

$$\beta^*_{\Lambda_n} := P_{\Lambda_n}\beta^* = \arg\min_{\beta_{\Lambda_n}\in\Phi_{\Lambda_n}} Q_n(\beta_{\Lambda_n}),$$
$$V_{\Lambda_n} := P_{\Lambda_n}V = G_0(\beta^0_{\Lambda_n})^T W_0(\beta^0_{\Lambda_n})G_0(\beta^0_{\Lambda_n}).$$

Proof. See Appendix B.2. □

A few remarks follow. First, the marginals of the mean function β^* can be interpreted as the (sieve) GMM estimator while the marginals of the covariance function V corresponds to the efficient variance matrix. Hence the result in Theorem 3.7 states that the posterior process weakly converges to the specific Gaussian process defined on the multiscale space $\mathcal{M}_0(\omega)$, i.e., any marginals of the posterior law weakly converge to the Gaussian measure centered at the GMM estimator with the efficient variance matrix. Second, by formulating our criterion functions through GMM framework, our result in Theorem 3.7 does not suffer from a well-known covariance mismatch of quasi-Bayes posteriors, which arises due to the use of "misspecified" likelihood (e.g. Kleijn and van der Vaart, 2012; Müller, 2013). Thus, Theorem 3.7 guarantees a good frequentist property of the quasi-Bayes posterior and the result can be directly used for constructing credible bands later.

3.3. **Construction of credible band from posterior.** Here we construct a credible band C_n that possesses valid frequentist properties. Roughly speaking, our construction of C_n consists of two requirements. First, we rely on weak BvM in Theorem 3.7 and derive $\mathcal{M}(\omega)$-radius; Second, we pick a pivot estimator (e.g. the posterior median) and derive a local radius centering around it. C_n is then constructed by intersecting these two requirements. We below formalize these statements and prove its validity.

A well-known difficulty in constructing confidence band is that it is generally not possible to build one that is both adaptive and honest w.r.t. unknown regularity or smoothness (Low (1997)). The subsequent literature has tried to find conditions under which suitable subsets of function classes permit such confidence sets; one example is the so-called "self-similar" functions (see Giné and Nickl (2021) and references therein). This idea is applied to nonparametric Bayes framework and adaptive credible bands are constructed for homogeneous Hölderian functions (e.g. Ray, 2017; Castillo and Ročková, 2021). As in Ročková and Rousseau (2023), we extend the notion of self-similarity to locally heterogeneous Hölderian functions.

We say that $h \in \mathcal{C}(t, M, \eta)$ is *locally self-similar at* $x \in [0,1]$ if there exist some $\tau > 0$ and $j_0 \in \mathbb{N}$ such that

$$|P_{<j}h(x) - h(x)| \geq 2^{-jt(x)}\tau, \quad \text{for all } j \geq j_0,$$

where $P_{<j}h = \sum_{l<j}\sum_{k}^{2^l-1}\langle h, \phi_{lk}\rangle\phi_{lk}$ denotes the projection with the resolution level less than j. We require that the true function h_0 is locally self-similar over the domain, which is formally stated as the following assumption.

Assumption 3.8. Local Self-Similarity
The function h_0 is locally self-similar for all $x \in [0,1]$.

We denote by $\mathcal{C}_{SS}(t, M, \eta)$ the class of functions that satisfy Assumption 2.1 and 3.8.

Now we construct C_n. First, let

$$C_n^1 := \left\{h : \|h - h^*\|_{\mathcal{M}(\omega)} \leq R_n(\alpha, Z^n)/\sqrt{n}\right\}, \tag{7}$$

where $R_n(\alpha, Z^n)$ is chosen such that $\Pi_n[C_n^1|Z^n] = 1 - \alpha$ for some $\alpha \in (0,1)$. Next, define the *local radius* as

$$\sigma_n(x) = \nu_n\sqrt{\frac{\log n}{n}}\sum_{l=0}^{L_{max}-1}\mathbf{1}\left\{(l, k_l(x)) \in \Lambda_n^*\right\}\left|\phi_{lk_l(x)}(x)\right|,$$

where (i) ν_n is any diverging sequence, (ii) $k_l(x) \in \{0, 1, ..., 2^l - 1\}$ is such that $x \in I_{lk_l(x)}$ for any $x \in [0,1]$, and (iii) Λ_n^* is the set of nodes defined as

$$\Lambda_n^* := \{(l,k) \in \Lambda_n : \Pi_n[(l,k) \in \lambda|Z^n] \geq 1/2\}, \tag{8}$$

i.e., all nodes $(l,k) \in \Lambda_n^*$ has its posterior mass (defined coordinate wise) greater than $1/2$. This naturally leads to the posterior median estimator

$$\hat{h}_{med}(x) = \sum_{(l,k)\in\Lambda_n^*}\beta_{lk}^*\phi_{lk}(x).$$

Using $\hat{h}_{med}(x)$ and the local radius $\sigma_n(x)$, we define

$$C_n^2 := \left\{h : \sup_{x\in[0,1]}\left[\frac{1}{\sigma_n(x)}|h(x) - \hat{h}_{med}(x)|\right] \leq 1\right\}. \tag{9}$$

Our C_n is then constructed by intersecting C_n^1 and C_n^2, i.e.,

$$C_n = \left\{h : h \in C_n^1 \cap C_n^2\right\}.$$

Given the above setting, our results on C_n are formally stated in the following Theorem 3.9.

Theorem 3.9. *Suppose that Assumptions 2.1, 3.1, 3.2, 3.6, and 3.8 hold. Then, for any admissible sequence* $\omega = (\omega)_{l\in\mathbb{N}}$ *such that* $\omega_{j_0(n)} \geq C\sqrt{\log n}$ *for some* $C > 0$, *we have*

$$\sup_{h_0\in\mathcal{C}_{SS}(t,M,\eta)}\mathbb{P}_{h_0}(h_0 \in C_n) = 1 - \alpha + o(1), \quad \text{as } n \to \infty, \tag{10}$$

and the posterior satisfies

$$\sup_{h_0 \in \mathcal{C}_{SS}(t,M,\eta)} \mathbb{E}_{h_0} \Pi_n \left[C_n | Z^n \right] = 1 - \alpha + o(1), \quad \text{as } n \to \infty. \tag{11}$$

Moreover, the L^∞-diameter of C_n satisfies

$$\sup_{f,g \in C_n} \left[\sup_{x \in [0,1]} \xi_n(x)^{-1} |f(x) - g(x)| \right] = O_p(\nu_n). \tag{12}$$

Proof. See Appendix B.3. □

Remark 3.10. Although the self-similar assumption or its variant has been used in the literature since Giné and Nickl (2010) to achieve both honesty and adaptivity, Armstrong (2021) finds it impossible to do so in a "fully" adaptive manner; there appears a term that contains a constant used to define the self-similarity of the function class. As discussed in Armstrong (2021), however, it is still possible to construct "practical" credible bands by explicitly specifying the penalty term; our results in Theorem 3.9 are in the same spirit.

4. Discussion

This paper has proposed the quasi-Bayesian inference scheme for nonparametric functions in conditional moment models. Importantly, we allow the targeted function to possess inhomogeneous smoothness and have shown the locally adaptive properties of the quasi-Bayes posteriors. Given these results, we have constructed confidence bands and established their asymptotic properties.

One important direction of research is in the computational aspect of our quasi-Bayes posteriors. There could be a difficulty in a typical application of Markov Chain Monte Carlo (MCMC) in our context, especially because (1) we have infinite dimensional parameter as our target and (2) we allow the targeted function to be defined by non-smooth criterion function (e.g. nonparametric quantile functions). Hamiltonian Monte Carlo (HMC) is a natural candidate to deal with the problem (1) instead of a standard MCMC, and there are some papers (e.g. Chaari et al. (2016)), where HMC is tailored to handle non-smoothness. However, their properties need further investigation both from theoretical and empirical perspectives.

Appendices

A. Controlling sieve approximation error

Here we briefly state the result about convergence of the sieve approximation error $\|P_{l_n}h_0 - h_0\|_\infty$. The key observation that affects this convergence rate is the decaying speed of the wavelet coefficients $\beta^0_{l_nk} \to 0$ as $n \to \infty$, which, as shown in Lemma 1 in Ročková and Rousseau (2023), is governed by the local smoothness of h_0. For the sake of completeness, we state the result of the decaying speed of the wavelet coefficients $\beta^0_{l_nk}$ in the following Lemma A.1.

Lemma A.1. *Suppose that Assumption 2.1 holds. For any $l > 0$ and $x \in [0,1]$, let (i) $k_l(x) \in \{0, 1, ..., 2^l - 1\}$ be such that $x \in I_{lk_l(x)}$, (ii) $M_{lk} = \max_{x \in I_{lk}} M(x)$, and (iii) $\eta_{lk} = \min_{x \in I_{lk}} \eta(x)$. Then $l \geq \log_2 \left[1/\left(2\eta_{lk_l(x)}\right)\right]$ implies $\left|\beta^0_{lk_l(x)}\right| \leq 2M_{lk_l(x)}2^{-l[t(x)+1/2]}$.*

Proof. See Lemma 1 in Ročková and Rousseau (2023). □

Note here that Ročková and Rousseau (2023) proceed their analysis in the fixed design framework, i.e., the input x is non-random. In this paper, we do not necessarily take this framework; hence Lemma A.1 is interpreted as the result conditioning on realized values of random variables $X_i = x \in [0,1]$.

Lemma A.2. *Suppose that Assumption 2.1 holds. Let $k_l(x), M_{lk_l(x)}$, and $\eta_{lk_l(x)}$ be as in Lemma A.1, and define*

$$l_n(x)^* := \max\left\{\log_2\left[1/\left(2\eta_{lk_l(x)}\right)\right],\ d_l(x,\nu)\right\}, \tag{A.1}$$

where, for some constant $\nu > 0$,

$$d_l(x,\nu) = \left\lfloor \log_2 \left[C_l(x,\nu)\left(\frac{n}{\log n}\right)^{\frac{1}{2t(x)+1}}\right]\right\rfloor, \quad C_l(x,\nu) = \left(\frac{2M_{lk_l(x)}}{\nu}\right)^{\frac{1}{t(x)+1/2}}.$$

Then for any $l_n(x)$ with $l_n(x)^* \leq l_n(x)$, we have

$$\sup_{x\in[0,1]} \xi_n(x)^{-1}\left|P_{l_n}h_0(x) - h_0(x)\right| = O(1), \tag{A.2}$$

where $\xi_n(x) = (\log n/n)^{t(x)/(2t(x)+1)}$, and

$$\left|\beta^0_{l_nk_{l_n}(x)}\right| \leq \nu\sqrt{\log n/n} \tag{A.3}$$

Proof. Let $l_n(x)$ be such that $l_n(x) \geq l_n(x)^*$ for all $x \in [0,1]$. Then, by the definition of the projection operator P_{l_n}, we have

$$h_0(x) - P_{l_n}h_0(x) = \sum_{l > l_n(x)} \sum_{k=0}^{2^l-1} \beta^0_{lk}\phi_{lk}(x).$$

The result about sieve approximation rate in (A.2) immediately follows from Lemma B.2 in Ročková and Rousseau (2023). The second result in (A.3) follows from the construction of $l_n(x)^*$ and from Lemma A.1. □

Lemma A.2 provides the sufficient condition for the sieve approximation error to contract uniformly over $[0,1]$ at the locally adaptive minimax rate $\xi_n(x)$. For example, the assumptions of $\eta(.)$ and $M(.)$ show that $\lfloor \log_2 (n/\log n) \rfloor \leq l_n < L_{max}$ satisfies the condition in Lemma A.2, leading to the adaptive minimax rate. This result is combined with our main result of posterior contraction.

B. Theoretical background

In this section, we show the proofs of the theoretical results in the main text.

B.1. **Proofs of the results in Section 3.1.** Here we prove Theorem 3.4. Following Hoffmann, Rousseau, and Schmidt-Hieber (2015), we first define

$$\mathcal{L}_n(\gamma) := \left\{(l,k);\ \left|\beta_{lk}^0\right| > \gamma\sqrt{\log n/n}\right\} \quad \text{with } \gamma > 0. \tag{B.1}$$

We below use the following Lemma to establish Theorem 3.4.

Lemma B.1. *Suppose that the conditions in Theorem 3.4 hold. Then there exist $\gamma^* \geq \gamma_* > 0$ such that*

$$\sup_{h_0 \in \mathcal{C}(t,M,\eta)} \mathbb{E}_{h_0} \Pi_n \left[\lambda^c \cap \mathcal{L}_n(\gamma^*) \neq \emptyset | Z^n\right] = o(1), \tag{B.2}$$

and

$$\sup_{h_0 \in \mathcal{C}(t,M,\eta)} \mathbb{E}_{h_0} \Pi_n \left[\lambda \cap \mathcal{L}_n^c(\gamma_*) \neq \emptyset | Z^n\right] = o(1). \tag{B.3}$$

Moreover, the bound γ^* can be chosen large enough such that

$$\sup_{h_0 \in \mathcal{C}(t,M,\eta)} \mathbb{E}_{h_0} \Pi_n \left[\max_{(l,k)\in\mathcal{L}_n(\gamma_*)} \left|\beta_{lk} - \beta_{lk}^0\right| > \gamma^*\sqrt{\log n/n} | Z^n\right] = o(1). \tag{B.4}$$

Proof. (B.2)
First note that

$$\Pi_n \left[\lambda^c \cap \mathcal{L}_n(\gamma^*) \neq \emptyset | Z^n\right] \leq \sum_{(l,k)\in\mathcal{L}_n(\gamma^*)} \Pi_n \left[(l,k) \notin \lambda | Z^n\right]. \tag{B.5}$$

For $\lambda \subset \Lambda_n$ with $(l,k) \notin \lambda$, let $\lambda^+ = \lambda \cup \{(l,k)\}$. Given an arbitrary λ, let

$$m_n(\lambda) = \int_{\mathbb{R}^{|\Lambda_n|}} \exp\left\{-nQ_n(\beta_{\Lambda_n})\right\} \pi_{|\lambda}(d\beta_{\Lambda_n}),$$

where $\beta_{\Lambda_n} = (\beta_{lk})_{(l,k)\in\Lambda_n}$. Note that the expressions $Q_n(\beta_{\Lambda_n})$ and $\pi_{|\lambda}(d\beta_{\Lambda_n})$ are understood as criterion function and prior evaluated at the function $h_{\Lambda_n} \in \Phi_{\Lambda_n}$ whose sieve coefficients are β_{Λ_n}; we use this intersection in the proofs. Now, each of the summand in (B.5) is expressed as

$$\Pi_n \left[(l,k) \notin \lambda | Z^n\right] = \frac{\sum_{\lambda\subset\Lambda_n} \mathbf{1}\{(l,k) \notin \lambda\} D(\lambda,\lambda^+) \Pi_{\Lambda_n}(\lambda^+) m_n(\lambda^+)}{\sum_{\lambda\subset\Lambda_n} \Pi_{\Lambda_n}(\lambda) m_n(\lambda)},$$

where

$$D(\lambda,\lambda^+) = \frac{\Pi_{\Lambda_n}(\lambda) m_n(\lambda)}{\Pi_{\Lambda_n}(\lambda^+) m_n(\lambda^+)}.$$

By Assumption 2.3, we have $\Pi_{\Lambda_n}(\lambda)/\Pi_{\Lambda_n}(\lambda^+) \leq 1/(c_T w_l) \leq n^K/c_T$. Regarding $m_n(\lambda)/m_n(\lambda^+)$, we employ the quadratic expansion of the sieve GMM criterion. To this

end, note that Assumption 3.2, together with the Markov inequality, implies $g_n(\beta_{\Lambda_n}) = g_n(\beta^0_{\Lambda_n}) + G_n(\beta^0_{\Lambda_n})(\beta_{\Lambda_n} - \beta^0_{\Lambda_n}) + o_p(\|\beta_{\Lambda_n} - \beta^0_{\Lambda_n}\|)$. Using this decomposition, we have

$$\begin{aligned}
&Q_n(\beta_{\Lambda_n}) - Q_n(\beta^0_{\Lambda_n}) \\
&\quad = g_n(\beta^0_{\Lambda_n})^T W_n(\beta_{\Lambda_n}) G_n(\beta^0_{\Lambda_n})(\beta_{\Lambda_n} - \beta^0_{\Lambda_n}) \\
&\qquad + \frac{1}{2}(\beta_{\Lambda_n} - \beta^0_{\Lambda_n})^T G_n(\beta^0_{\Lambda_n})^T W_n(\beta_{\Lambda_n}) G_n(\beta^0_{\Lambda_n})(\beta_{\Lambda_n} - \beta^0_{\Lambda_n}) \\
&\qquad + \frac{1}{2} g_n(\beta^0_{\Lambda_n})^T \left[W_n(\beta_{\Lambda_n}) - W_n(\beta^0_{\Lambda_n})\right] g_n(\beta^0_{\Lambda_n}) + o_p(\|\beta_{\Lambda_n} - \beta^0_{\Lambda_n}\|^2) \\
&\quad = \underbrace{g_n(\beta^0_{\Lambda_n})^T W_n(\beta^0_{\Lambda_n}) G_n(\beta^0_{\Lambda_n})}_{=:\Delta_n(\beta^0_{\Lambda_n})^T}(\beta_{\Lambda_n} - \beta^0_{\Lambda_n}) \\
&\qquad + \frac{1}{2}(\beta_{\Lambda_n} - \beta^0_{\Lambda_n})^T \underbrace{G_n(\beta^0_{\Lambda_n})^T W_n(\beta^0_{\Lambda_n}) G_n(\beta^0_{\Lambda_n})}_{=:J_n(\beta^0_{\Lambda_n})}(\beta_{\Lambda_n} - \beta^0_{\Lambda_n}) \\
&\qquad + R_n(\beta_{\Lambda_n}) + o_p(\|\beta_{\Lambda_n} - \beta^0_{\Lambda_n}\|^2),
\end{aligned}$$

where

$$\begin{aligned}
&R_n(\beta_{\Lambda_n}) \\
&\quad = g_n(\beta^0_{\Lambda_n})^T \left[W_n(\beta_{\Lambda_n}) - W_n(\beta^0_{\Lambda_n})\right] G_n(\beta^0_{\Lambda_n})(\beta_{\Lambda_n} - \beta^0_{\Lambda_n}) \\
&\quad + \frac{1}{2}(\beta_{\Lambda_n} - \beta^0_{\Lambda_n})^T G_n(\beta^0_{\Lambda_n})^T \left[W_n(\beta_{\Lambda_n}) - W_n(\beta^0_{\Lambda_n})\right] G_n(\beta^0_{\Lambda_n})(\beta_{\Lambda_n} - \beta^0_{\Lambda_n}) \\
&\quad + \frac{1}{2} g_n(\beta^0_{\Lambda_n})^T \left[W_n(\beta_{\Lambda_n}) - W_n(\beta^0_{\Lambda_n})\right] g_n(\beta^0_{\Lambda_n}).
\end{aligned}$$

As immediate consequences of Assumption 3.2, it follows that (i) $g_n(\beta^0_{\Lambda_n}) = O_p(1/\sqrt{n})$, (ii) $\left\|W_n(\beta_{\Lambda_n}) - W_n(\beta^0_{\Lambda_n})\right\|_{\text{HS}} = O_p(\left\|\beta_{\Lambda_n} - \beta^0_{\Lambda_n}\right\|)$ as $\beta_{\Lambda_n} \to \beta^0_{\Lambda_n}$, where $\|.\|_{\text{HS}}$ represents the Hilbert-Schmidt norm, and (iii) $G_n(\beta^0_{\Lambda_n}) = O_p(1)$. Thus, as $\beta_{\Lambda_n} \to \beta^0_{\Lambda_n}$ and $n \to \infty$, we have

$$\begin{aligned}
R_n(\beta_{\Lambda_n}) &= O_p\left(\left(\left\|\beta_{\Lambda_n} - \beta^0_{\Lambda_n}\right\|^2 / \sqrt{n}\right) \vee \left(\left\|\beta_{\Lambda_n} - \beta^0_{\Lambda_n}\right\|^3\right) \vee \left(\left\|\beta_{\Lambda_n} - \beta^0_{\Lambda_n}\right\| / n\right)\right) \\
&\lesssim O_p\left(\left(\left\|\beta_{\Lambda_n} - \beta^0_{\Lambda_n}\right\| \vee 1/\sqrt{n}\right)^3\right).
\end{aligned}$$

Our analysis above indicates that $R_n(\beta_{\Lambda_n})$ becomes negligible as $\beta_{\Lambda_n} \to \beta^0_{\Lambda_n}$, leaving only $Q_n(\beta_{\Lambda_n}) - Q_n(\beta^0_{\Lambda_n}) = \Delta_n(\beta^0_{\Lambda_n})^T(\beta_{\Lambda_n} - \beta^0_{\Lambda_n}) + (\beta_{\Lambda_n} - \beta^0_{\Lambda_n})^T J_n(\beta^0_{\Lambda_n})(\beta_{\Lambda_n} - \beta^0_{\Lambda_n})$ relevant to determine the contraction rate. Since these terms are of the additive form in each element of β_{Λ_n}, the independence of priors $\pi_{|\lambda^+}$ and $\pi_{|\lambda}$ implies

$$\begin{aligned}
\frac{m_n(\lambda)}{m_n(\lambda^+)} &= \frac{\int_{\mathbb{R}^{|\Lambda_n|}} \exp\left\{-n\left[Q_n(\beta_{\Lambda_n}) - Q_n(\beta^0_{\Lambda_n})\right]\right\} \pi_{|\lambda}(d\beta_{\Lambda_n})}{\int_{\mathbb{R}^{|\Lambda_n|}} \exp\left\{-n\left[Q_n(\beta_{\Lambda_n}) - Q_n(\beta^0_{\Lambda_n})\right]\right\} \pi_{|\lambda^+}(d\beta_{\Lambda_n})} \\
&\lesssim \frac{\exp\left\{-n\tilde{Q}_n(0)\right\}}{\int_{\mathbb{R}} \exp\left\{-n\tilde{Q}_n(\beta_{lk})\right\} q(\beta_{lk}) d\beta_{lk}},
\end{aligned}$$

where,

$$\tilde{Q}_n(\beta_{lk}) = \left[\Delta_n(\beta^0_{\Lambda_n})\right]_{lk}(\beta_{lk} - \beta^0_{lk}) + \left[J_n(\beta^0_{\Lambda_n})\right]_{lk,lk}(\beta_{lk} - \beta^0_{lk})^2, \tag{B.6}$$

and $\left[\Delta_n(\beta^0_{\Lambda_n})\right]_{lk}$ and $\left[J_n(\beta^0_{\Lambda_n})\right]_{lk,lk}$ represent the lk-element and (lk, lk)-element of the vector and the matrix respectively.

For the upper bound of the numerator of $m_n(\lambda)/m_n(\lambda^+)$, we have

$$\tilde{Q}_n(0) = -\left[\Delta_n(\beta^0_{\Lambda_n})\right]_{lk}\beta^0_{lk} + \left[J_n(\beta^0_{\Lambda_n})\right]_{lk,lk}(\beta^0_{lk})^2.$$

Recall that $|\beta^0_{lk}| > \gamma^*\sqrt{\log n/n}$ at the node $(l,k) \in \mathcal{L}_n(\gamma^*)$. Moreover, it follows from Assumption 3.2 that $\left[J_n(\beta^0_{\Lambda_n})\right]_{lk,lk} = O_p(1)$ and that $\left[\Delta_n(\beta^0_{\Lambda_n})\right]_{lk} = O_p(1/\sqrt{n})$. Combining these arguments, we have $(\gamma^*)^2(\log n/n) \lesssim \tilde{Q}_n(0)$ stochastically, and hence $\exp\{-n\tilde{Q}_n(0)\} \lesssim \exp\left\{-(\gamma^*)^2\log n\right\} = 1/n^{(\gamma^*)^2}$ in probability.

For the denominator, let $N_n := \{\beta_{lk} : \ |\beta_{lk} - \beta^0_{lk}| \le \sqrt{\log n/n}\}$. Then $N_n \subset [-R, R]$ for suitably large R, and

$$\begin{aligned}\int_{\mathbb{R}} \exp\left\{-n\tilde{Q}_n(\beta_{lk})\right\} q(\beta_{lk})d\beta_{lk} &\ge c_R \int_{-R}^{R} \exp\left\{-n\tilde{Q}_n(\beta_{lk})\right\} d\beta_{lk} \\ &\ge c_R \int_{N_n} \exp\left\{-n\tilde{Q}_n(\beta_{lk})\right\} d\beta_{lk}.\end{aligned}$$

The same reasoning implies that we have $\tilde{Q}_n(\beta_{lk}) \lesssim \log n/n$ on the region N_n. Hence, we get $\exp\{-\log n\} = 1/n \lesssim \int_{\mathbb{R}} \exp\left\{-n\tilde{Q}_n(\beta_{lk})\right\} q(\beta_{lk})d\beta_{lk}$ in probability. Combining the results on these bounds, we have $n^{-\left(1-(\gamma^*)^2\right)} \lesssim m_n(\lambda^+)/m_n(\lambda)$, i.e., $m_n(\lambda)/m_n(\lambda^+) \lesssim n^{1-(\gamma^*)^2}$.

Finally we get

$$\begin{aligned}\mathbb{E}_{h_0}\Pi_n\left[\lambda^c \cap \mathcal{L}_n(\gamma^*) \ne \emptyset | Z^n\right] &\le \sum_{(l,k)\in\mathcal{L}_n(\gamma^*)} \mathbb{E}_{h_0}\Pi_n\left[(l,k) \notin \lambda | Z^n\right] \\ &\le |\mathcal{L}_n(\gamma^*)|\, \mathbb{E}_{h_0} D(\lambda, \lambda^+) \\ &\lesssim n^{1+K+1-(\gamma^*)^2} = o(1),\end{aligned}$$

where the last equality follows for a large enough γ^* such that $(\gamma^*)^2 > 2 + K$.

Proof. (B.3)
Similarly to the proof for (B.2), we have

$$\Pi_n\left[\lambda \cup \mathcal{L}_n^c(\gamma_*) \ne \emptyset | Z^n\right] \le \sum_{(l,k)\notin\mathcal{L}_n(\gamma_*)} \Pi_n\left[(l,k) \in \lambda | Z^n\right], \tag{B.7}$$

where each of the summand is expressed as

$$\Pi_n\left[(l,k) \in \lambda | Z^n\right] = \frac{\sum_{\lambda\subset\Lambda_n} \mathbf{1}\{(l,k) \in \lambda\} D(\lambda, \lambda^-)\Pi_{\Lambda_n}(\lambda^-)m_n(\lambda^-)}{\sum_{\lambda\subset\Lambda_n} \Pi_{\Lambda_n}(\lambda)m_n(\lambda)},$$

with $\lambda^- = \lambda \backslash \{(l.k)\}$ and

$$D(\lambda, \lambda^-) = \frac{\Pi_{\Lambda_n}(\lambda) m_n(\lambda)}{\Pi_{\Lambda_n}(\lambda^-) m_n(\lambda^-)}.$$

By Assumption 2.3, $\Pi_{\Lambda_n}(\lambda)/\Pi_{\Lambda_n}(\lambda^-) \leq C_T w_l \leq C_T 2^{-l}$ for $l < L_{max}$. Regarding to $m_n(\lambda)/m_n(\lambda^-)$, the same reasoning as in the proof in (B.2) shows

$$\frac{m_n(\lambda)}{m_n(\lambda^-)} = \frac{\int_{\mathbb{R}} \exp\left\{-n\tilde{Q}_n(\beta_{lk})\right\} q(\beta_{lk}) d\beta_{lk}}{\exp\left\{-n\tilde{Q}_n(0)\right\}}, \tag{B.8}$$

where $\tilde{Q}_n$ is as in (B.6). Recall that $|\beta^0_{lk}| \leq \gamma_* \sqrt{\log n/n}$ at the node $(l,k) \notin \mathcal{L}_n(\gamma_*)$. Thus, $\tilde{Q}_n(0) \lesssim (\gamma_*)^2 (\log n/n)$ stochastically and hence $\exp\left\{-(\gamma_*)^2 \log n\right\} = 1/n^{(\gamma_*)^2} \lesssim \exp\{-n\tilde{Q}_n(0)\}$ in probability. For the numerator, note that $N_n \subset \{\beta_{lk} : |\beta_{lk}| \leq (\gamma_*+1)\sqrt{\log n/n}\}$ whose Lebesgue measure is $2(\gamma_*+1)\sqrt{\log n/n}$. Hence,

$$\begin{aligned}
&\int_{\mathbb{R}} \exp\left\{-n\tilde{Q}_n(\beta_{lk})\right\} q(\beta_{lk}) d\beta_{lk} \\
&= \int_{N_n} \exp\left\{-n\tilde{Q}_n(\beta_{lk})\right\} q(\beta_{lk}) d\beta_{lk} + \int_{N_n^c} \exp\left\{-n\tilde{Q}_n(\beta_{lk})\right\} q(\beta_{lk}) d\beta_{lk} \\
&\lesssim \int_{N_n} \exp\left\{-n\tilde{Q}_n(\beta_{lk})\right\} q(\beta_{lk}) d\beta_{lk} + n^{-1} \\
&\lesssim C_R \times \exp\left\{-n\tilde{Q}_n(\beta^0_{lk})\right\} \times 2(\gamma_*+1)\sqrt{\log n/n} + n^{-1} \asymp n^{-1}.
\end{aligned}$$

Combining the bounds, we have $m_n(\lambda)/m_n(\lambda^-) \lesssim n^{(\gamma_*)^2-1}$.

Finally we get

$$\begin{aligned}
\mathbb{E}_{h_0}\left[\lambda \cap \mathcal{L}_n^c(\gamma_*) \neq \emptyset | Z^n\right] &\leq \sum_{(l,k) \notin \mathcal{L}_n(\gamma_*)} \mathbb{E}_{h_0} \Pi_n\left((l,k) \in \lambda | Z^n\right) \\
&\leq \sum_{l \leq L_{max}} 2^l \times \mathbb{E}_{h_0} D(\lambda, \lambda^-) \\
&\leq C_T \times L_{max} \times n^{(\gamma_*)^2-1} \\
&\lesssim \log_2 n \times n^{(\gamma_*)^2-1} = o(1),
\end{aligned}$$

where the last equality follows given that γ_* is sufficiently small.

Proof. (B.4)
First note that it suffices by (B.2) to consider the support λ with $\lambda^c \cap \mathcal{L}_n(\gamma^*) = \emptyset$. Now we have

$$\begin{aligned}
&\Pi_n\left[\max_{(l,k) \in \mathcal{L}_n(\gamma_*)} \left|\beta_{lk} - \beta^0_{lk}\right| > \gamma^* \sqrt{\log n/n} \text{ and } \beta_{lk} \neq 0 | Z^n\right] \\
&\lesssim n \times \max_{(l,k) \in \mathcal{L}_n(\gamma_*)} \Pi_n\left[\left|\beta_{lk} - \beta^0_{lk}\right| > \gamma^* \sqrt{\log n/n} \text{ and } \beta_{lk} \neq 0 | Z^n\right],
\end{aligned}$$

where the inequality follows from a union bound. For each $(l,k) \in \mathcal{L}_n(\gamma_*)$,

$$\begin{aligned}
&\Pi_n \left[\left|\beta_{lk} - \beta_{lk}^0\right| > \gamma^* \sqrt{\log n/n} \text{ and } \beta_{lk} \neq 0 | Z^n \right] \\
&\leq \frac{\int_{\mathbb{R}} \mathbf{1}\left\{ |\beta_{lk} - \beta_{lk}^0| > \gamma^* \sqrt{\log n/n} \right\} \exp\left\{ -n\tilde{Q}_n(\beta_{lk}) \right\} q(\beta_{lk}) d\beta_{lk}}{\int_{\mathbb{R}} \exp\left\{ -n\tilde{Q}_n(\beta_{lk}) \right\} q(\beta_{lk}) d\beta_{lk}} \lesssim n^{1-(\gamma^*)^2},
\end{aligned}$$

where the last result on bound is due to the fact that the numerator is bounded above by $C_R \times \exp\{-(\gamma^*)^2 \log n\} \asymp n^{-(\gamma^*)^2}$ in probability, while the denominator is bounded below by n^{-1} in probability, by the same argument as in the proof for (B.2). Combining the arguments,

$$\begin{aligned}
&\mathbb{E}_{h_0} \Pi_n \left[\max_{(l,k) \in \mathcal{L}_n(\gamma_*)} \left|\beta_{lk} - \beta_{lk}^0\right| > \gamma^* \sqrt{\log n/n} \text{ and } \beta_{lk} \neq 0 | Z^n \right] \\
&\lesssim n^{2-(\gamma^*)^2} = o(1),
\end{aligned}$$

provided that $\gamma^* > \sqrt{2}$. □

Proof. Lemma 3.3
We finally establish Lemma 3.3. Let $\bar{\Lambda}_n \subset \Lambda_n$ be a set of λ such that $\mathcal{L}_n(\gamma^*) \subset \lambda \subset \mathcal{L}_n(\gamma_*)$. Note here that

$$\left\{ (l,k) : \; l \leq \min_{x \in I_{lk}} d_l(x, \gamma^*) \right\} \subset \mathcal{L}_n(\gamma_*) \subset \left\{ (l,k) : \; l \leq \min_{x \in I_{lk}} d_l(x, \gamma_*) \right\}, \tag{B.9}$$

where $d_l(x,.)$ is as in Lemma A.2.

Recall $\xi_n(x)^{-1} = (n/\log n)^{t(x)/(2t(x)+1)}$. Now we bound the quantity of interest as follows,

$$\begin{aligned}
&\xi_n(x)^{-1} \left| h_{\Lambda_n}(x) - P_{\Lambda_n} h_0(x) \right| \\
&\quad \leq \xi_n(x)^{-1} \sum_{(l,k) \in \lambda \subset \bar{\Lambda}_n} \mathbf{1}\{x \in I_{lk}\} 2^{l/2} \left|\beta_{lk} - \beta_{lk}^0\right| \\
&\qquad + \xi_n(x)^{-1} \sum_{(l,k) \notin \lambda \subset \bar{\Lambda}_n} \mathbf{1}\{x \in I_{lk}\} 2^{l/2} \left|\beta_{lk}^0\right|.
\end{aligned}$$

For the first term in the right hand side above, it follows from (B.4) and (B.9) that

$$\begin{aligned}
&\sup_{x \in [0,1]} \xi_n(x)^{-1} \sum_{(l,k) \in \lambda \subset \bar{\Lambda}_n} \mathbf{1}\{x \in I_{lk}\} 2^{l/2} \left|\beta_{lk} - \beta_{lk}^0\right| \\
&\leq \gamma^* \sqrt{\log n/n} \times \sup_{x \in [0,1]} \xi_n(x)^{-1} 2^{d_l(x,\gamma_*)/2} \\
&\lesssim (\log n/n)^{1/2 - t(x)/(2t(x)+1) - 1/(4t(x)+2)} = O(1).
\end{aligned}$$

For the second term in the right hand side, note that $(l,k) \notin \lambda \subset \bar{\Lambda}_n$ and (B.9) imply $|\beta_{lk}^0| \leq \gamma^* \sqrt{\log n/n}$ and $l \leq d_l(x, \gamma_*)$. Hence,

$$\begin{aligned}
&\sup_{x\in[0,1]} \xi_n(x)^{-1} \sum_{(l,k)\notin\lambda\subset\bar{\Lambda}_n} \mathbf{1}\{x \in I_{lk}\} 2^{l/2} \left|\beta_{lk}^0\right| \\
&\leq \gamma^* \sqrt{\log n/n} \times \sup_{x\in[0,1]} \xi_n(x)^{-1} 2^{d_l(x,\gamma_*)/2} \\
&\lesssim (\log n/n)^{1/2 - t(x)/(2t(x)+1) - 1/(4t(x)+2)} = O(1).
\end{aligned}$$

Lemma 3.3 follows by combining the two bounds derived above.

B.2. **Proofs of the results in Section 3.2.** We first prove the weak BvM phenomena in our quasi-Bayesian setting. Our proof extends Castillo and Nickl (2014) and Ray (2017) to our quasi-Bayesian framework. To this end, we use the following Lemmas.

Lemma B.2. *Let $\mathbb{E}^{\Pi_n}(.|Z^n)$ denote the expectation w.r.t. the posterior Π_n. Then, under the same conditions as in Theorem 3.7, for any $(l,k) \in \Lambda_n$ with $l \leq l_0(n)$ and for any $t > 0$,*

$$\mathbb{E}_{h_0}\mathbb{E}^{\Pi_n}\left[e^{t\sqrt{n}(\beta_{lk}-\tilde{\beta}_{lk})}|Z^n\right] \lesssim e^{t^2}, \tag{B.10}$$

where $\tilde{\beta}_{lk}$ denotes a random variable whose law is equivalent to the marginal of $\mathcal{N}(\beta^*, (nV)^{-1})$, i.e., $\tilde{\beta}_{lk} \sim N(\beta_{lk}^*,\ (nV_{lk})^{-1})$.

Proof. Fix any $(l,k) \in \Lambda_n$ with $l \leq l_0(n)$. Note that the random variable $\tilde{\beta}_{lk} \sim N(\beta_{lk}^*, (nV_{lk})^{-1})$ can be expressed as $\tilde{\beta}_{lk} = \beta_{lk}^0 + r_{lk}/\sqrt{n}$ with $r_{lk} \sim N(\beta_{lk}^* - \beta_{lk}^0,\ (V_{lk})^{-1})$. By the Bayes' formula combined with the independence structure of our prior, the expectation w.r.t. the posterior satisfies

$$\begin{aligned}
\mathbb{E}^{\Pi_n}\left[e^{t\sqrt{n}(\beta_{lk}-\tilde{\beta}_{lk})}|Z^n\right] &= \frac{\int_{\mathbb{R}} e^{t\sqrt{n}(\beta_{lk}-\tilde{\beta}_{lk})} e^{-n\tilde{Q}_n(\beta_{lk})} q(\beta_{lk}) d\beta_{lk}}{\int_{\mathbb{R}} e^{-n\tilde{Q}_n(\beta_{lk})} q(\beta_{lk}) d\beta_{lk}} \\
&= e^{-tr_{lk}} \frac{\int_{\mathbb{R}} e^{-t\sqrt{n}(\beta_{lk}^0-\beta_{lk})} e^{-n\tilde{Q}_n(\beta_{lk})} q(\beta_{lk}) d\beta_{lk}}{\int_{\mathbb{R}} e^{-n\tilde{Q}_n(\beta_{lk})} q(\beta_{lk}) d\beta_{lk}}
\end{aligned}$$

where $\tilde{Q}_n(.)$ is as in the proof in Lemma B.1. Since r_{lk} is Gaussian, if we bound almost surely the fraction of the integrals above by some constant $C > 0$, then the result follows from the standard sub-Gaussian characterization. To this end, first notice that, letting $E_n := \{\beta_{lk} : |\beta_{lk} - \beta_{lk}^0| \leq \sqrt{D/n}\}$ with an arbitrary constant $D > 0$, the denominator is bounded below as

$$\int_{\mathbb{R}} e^{-n\tilde{Q}_n(\beta_{lk})} q(\beta_{lk}) d\beta_{lk} \geq \int_{E_n} e^{-n\tilde{Q}_n(\beta_{lk})} q(\beta_{lk}) \geq c_R \times e^{-D},$$

where, by the same reasoning as in the proof of (B.2), we used $\tilde{Q}_n(\beta_{lk}) \lesssim D/n$ on E_n in addition to the property of the prior $q(.)$ for the second inequality. For the numerator, let $v = \sqrt{n}(\beta_{lk}^0 - \beta_{lk})$ and note that the components $\Delta_n(\beta_{\Lambda_n}^0)$ and $J_n(\beta_{\Lambda_n}^0)$ are bounded

in probability by Assumption 3.2; thus, combined with the property of the prior $q(.)$, it follows that

$$\int_{\mathbb{R}} e^{-t\sqrt{n}(\beta_{lk}^0-\beta_{lk})}e^{-n\tilde{Q}_n(\beta_{lk})}q(\beta_{lk})d\beta_{lk} \lesssim \int_{\mathbb{R}} e^{-tv}e^{-(v^2+v)}q(\beta_{lk}^0 - v/\sqrt{n})dv$$
$$\leq C_R \times e^{-(t+1)/2} \leq C_R \times e^{-1/2},$$

for any $t > 0$. Combining the two bounds, we can take $C = (C_R/c_R) \times e^{D-1/2}$ and complete the proof.

□

Lemma B.3. *Under the same conditions as in Theorem 3.7, for any sequence $D_n \to \infty$ we have*

$$\sup_{h_0 \in \mathcal{C}(t,M,\eta)} \mathbb{E}_{h_0}\Pi_n \left[h : \|h - h_0\|_{\mathcal{M}(\omega)} \geq n^{-1/2}D_n | Z^n\right] = o(1), \tag{B.11}$$

where $\omega = (\omega_l)_{l\in\mathbb{N}}$ is any admissible sequence that satisfies $\omega_{l_0(n)} \geq c\sqrt{\log n}$ for some $c > 0$.

Proof. Following Ray (2017), first define

$$A_n := \{\lambda^c \cap \mathcal{L}_n(\Lambda_n^*) = \emptyset\} \cap \{\lambda \cap \mathcal{L}_n^c(\lambda_*) = \emptyset\}$$
$$\cap \left\{ \max_{(l,k)\in\mathcal{L}_n(\gamma_*)} \left|\beta_{lk} - \beta_{lk}^0\right| \leq \gamma^* \sqrt{\log n/n} \right\}.$$

Note that Lemma B.1 shows

$$\sup_{h_0 \in \mathcal{C}(t,M,\eta)} \mathbb{E}_{h_0}\Pi_n(A_n^c | Z^n) = o(1),$$

for some $0 < \gamma_* \leq \gamma^* < \infty$.

Next we bound the quantity of interest as follows,

$$\begin{aligned}
&\mathbb{E}_{h_0}\Pi_n\left[\{\|h - h_0\|_{\mathcal{M}(\omega)} > n^{-1/2}D_n\}|Z^n\right] \\
&\leq \mathbb{E}_{h_0}\Pi_n\left[\{\|h - h_0\|_{\mathcal{M}(\omega)} > n^{-1/2}D_n\} \cap \{\|P_{l_0(n)}(h - h_0)\|_{\mathcal{M}(\omega)} \leq n^{-1/2}D_n\} \cap A_n|Z^n\right] \\
&\quad + \mathbb{E}_{h_0}\Pi_n\left[\{\|h - h_0\|_{\mathcal{M}(\omega)} > n^{-1/2}D_n\} \cap \{\|P_{l_0(n)}(h - h_0)\|_{\mathcal{M}(\omega)} > n^{-1/2}D_n\} \cap A_n|Z^n\right] \\
&\quad + \mathbb{E}_{h_0}\Pi_n(A_n^c|Z^n) \\
&\leq \mathbb{E}_{h_0}\Pi_n\left[\{\|P_{>l_0(n)}(h - h_0)\|_{\mathcal{M}(\omega)} > n^{-1/2}D_n\} \cap A_n|Z^n\right] \\
&\quad + \mathbb{E}_{h_0}\Pi_n\left[\{\|P_{l_0(n)}(h - h_0)\|_{\mathcal{M}(\omega)} > n^{-1/2}D_n\} \cap A_n|Z^n\right] \\
&\quad + o(1),
\end{aligned}$$

where $P_{>l_0(n)}$ is the projection onto the space spanned by $\{\phi_{lk}; l > l_0(n), 0 \leq k < 2^l\}$.

For the first term, on the resolution level $l \in (l_0(n), L_{max})$ and on A_n we have

$$\left\|P_{l_0(n)<l<L_{max}}(h - h_0)\right\|_{\mathcal{M}(\omega)} = \sup_{l_0(n)<l<L_{max}} \left(\frac{\max_k |\beta_{lk} - \beta_{lk}^0|}{\omega_l}\right)$$
$$\leq \frac{\gamma^*\sqrt{\log n/n}}{\omega_{l_0(n)}} = O(n^{-1/2}),$$

given the assumption $\omega_{l_0(n)} \geq c\sqrt{\log n}$ for some $c > 0$. On the resolution level $l > L_{max}$, note that $L_{max} > l_n^*(x)$ for any $x \in [0,1]$ in Lemma A.2. Thus, $|\beta_{lk}^0| \lesssim \sqrt{\log n/n}$ and

$$\begin{aligned}\|P_{>L_{max}}(h-h_0)\|_{\mathcal{M}(\omega)} = \|P_{>L_{max}}h_0\|_{\mathcal{M}(\omega)} &= \sup_{L_{max}<l}\left(\frac{\max_k|\beta_{lk}^0|}{\omega_l}\right) \\ &\lesssim \frac{\sqrt{\log n/n}}{\omega_{l_0(n)}} = O(n^{-1/2}),\end{aligned}$$

given the assumptions that ω_l is monotonically increasing and that $\omega_{l_0(n)} \geq c\sqrt{\log n}$ for some $c > 0$. These two results imply

$$\mathbb{E}_{h_0}\Pi_n\left[\{\|P_{>l_0(n)}(h-h_0)\|_{\mathcal{M}(\omega)} > n^{-1/2}D_n\} \cap A_n | Z^n\right] = o(1).$$

For the second term, recall that $\mathbb{E}^{\Pi_n}(.|Z^n)$ denotes the expectation w.r.t the posterior distribution. We show that $\mathbb{E}_{h_0}\mathbb{E}^{\Pi_n}\left[\sqrt{n}\left\|P_{l_0(n)}(h-h_0)\right\|_{\mathcal{M}(\omega)}|Z^n\right] < \infty$, i.e., the expected value of multiscale norm multiplied by $\sqrt{n}$ is bounded; this, combined with an application of Markov's inequality, shows that the second term is $o(1)$. First we bound the expectation as follows,

$$\begin{aligned}&\mathbb{E}_{h_0}\mathbb{E}^{\Pi_n}\left[\sqrt{n}\left\|P_{l_0(n)}(h-h_0)\right\|_{\mathcal{M}(\omega)}|Z^n\right] \\ &\leq \underbrace{\mathbb{E}_{h_0}\mathbb{E}^{\Pi_n}\left[\sqrt{n}\left\|P_{l_0(n)}(h-\tilde{h})\right\|_{\mathcal{M}(\omega)}|Z^n\right]}_{=:(A)} + \underbrace{\mathbb{E}_{h_0}\left[\sqrt{n}\left\|P_{l_0(n)}(\tilde{h}-h_0)\right\|_{\mathcal{M}(\omega)}\right]}_{=:(B)},\end{aligned}$$

where $\tilde{h}$ is a random process on $\mathcal{M}(\omega)$ whose law is Gaussian $\mathcal{N}(\beta^*, (nV)^{-1})$. Regarding the term (B), note that $\sqrt{n}(\tilde{h}-h_0) \sim \mathcal{N}(\beta^*-\beta^0, V^{-1})$, where the mean and the covariance are bounded uniformly over Λ_n. It then follows that

$$\mathbb{E}_{h_0}\left[\sqrt{n}\left\|P_{l_0(n)}(\tilde{h}-h_0)\right\|_{\mathcal{M}(\omega)}\right] \leq \mathbb{E}_{h_0}\left[\sqrt{n}\left\|\tilde{h}-h_0\right\|_{\mathcal{M}(\omega)}\right] < \infty,$$

where the last bound is due to Proposition 3.5 and the assumption that $\omega = (\omega_l)_{l\in\mathbb{N}}$ is admissible. For the term (A), we first bound it with a fixed constant $M > 0$ as follows,

$$\begin{aligned}(A) &= \mathbb{E}_{h_0}\mathbb{E}^{\Pi_n}\left[\sup_{l\leq l_0(n)}\left(\frac{\max_k\sqrt{n}\,|\beta_{lk}-\beta_{lk}^*|}{\omega_l}\right)|Z^n\right] \\ &\leq M + \int_M^\infty \mathbb{P}\left(\sup_{l\leq l_0(n)}\left(\frac{\max_k\sqrt{n}\,|\beta_{lk}-\beta_{lk}^*|}{\sqrt{l}}\right) > t\right)dt,\end{aligned}$$

where $\mathbb{P}$ denotes the low with the expectation $\mathbb{E}_{h_0}\mathbb{E}^{\Pi_n}[.|Z^n]$. Now the sub-Gaussian bound derived in Lemma B.2 imply that, for any $t > 0$,

$$\mathbb{P}\left(\sqrt{n}\,|\beta_{lk}-\beta_{lk}^*| > t\right) \lesssim e^{-Kt^2},$$

where $K > 0$ is some universal constant. Using this sub-Gaussian concentration inequality, we can further bound the tail integral as follows,

$$\begin{aligned}
&\int_M^\infty \mathbb{P}\left(\sup_{l\le l_0(n)}\left(\frac{\max_k \sqrt{n}\,|\beta_{lk}-\beta^*_{lk}|}{\sqrt{l}}\right) > t\right) dt \\
&\le \sum_{(l,k):l\le l_0(n)} \int_M^\infty \mathbb{P}\left(\sqrt{n}\,|\beta_{lk}-\beta^*_{lk}| > \sqrt{l}t\right) dt \\
&\lesssim \sum_{l\le l_0(n)} 2^l \int_M^\infty e^{-Klt^2} dt \lesssim \sum_{l\le l_0(n)} 2^l e^{-KlM^2} < \infty,
\end{aligned}$$

for sufficiently large M. This shows that the second term is $o(1)$ and completes the proof. □

Lemma B.4. *Suppose that Assumptions 2.1, 3.1, 3.2, and 3.6 hold. Then, for any fixed $l^* \in \mathbb{N}$,*

$$\mathcal{D}_{TV}\left[\mathcal{L}\left(\Pi_n[.|Z^n]\circ P_{l^*}^{-1}\right),\ \mathcal{N}(\beta^*,(nV)^{-1})\circ P_{l^*}^{-1}\right] = o_p(1),$$

where $\mathcal{D}_{TV}[P,Q]$ denotes the total variation metric between probability measures P and Q.

Proof. Let $\beta_{l^*} = (\beta_{lk})_{l\le l^*, 0\le k<2^l} \in \mathbb{R}^{l^*+1}$ be the vector consisting of all Haar wavelet coefficients up to the resolution level l^*. Then the statement in Lemma B.4 is equivalent to

$$\mathcal{D}_{\mathrm{TV}}\left[\mathcal{L}\left(\Pi_n[\sqrt{n}(\beta_{l^*}-\beta^0_{l^*})|Z^n]\right), N(\beta^*_{l^*}-\beta^0_{l^*}, V_{l^*}^{-1})\right] = o_p(1).$$

This parametric BvM statement would be deduced as a special case of Theorem 1 in Chernozhukov and Hong (2003)[3]. To this end, we verify the conditions of their Proposition 1, which are sufficient for the conditions in Theorem 1 in Chernozhukov and Hong (2003).

First, condition (i) and (ii) of Proposition 1 in Chernozhukov and Hong (2003) are simply deduced from our definition of the criterion function and our Assumption 3.2[4]. Second, regarding their condition (iii), recall that our objective function $Q_n(\beta_{l^*})$ (subtracted by $Q_n(\beta^0_{l^*})$) admits the quadratic expansion

$$\begin{aligned}
&Q_n(\beta_{l^*}) - Q_n(\beta^0_{l^*}) \\
&\quad = \Delta_n(\beta^0_{l^*})^T(\beta_{l^*}-\beta^0_{l^*}) + \frac{1}{2}(\beta_{l^*}-\beta^0_{l^*})^T J_n(\beta^0_{l^*})(\beta_{l^*}-\beta^0_{l^*}) + o_p(\left\|\beta_{l^*}-\beta^0_{l^*}\right\|^2),
\end{aligned}$$

where $\Delta_n(\beta^0_{l^*}) = g_n(\beta^0_{l^*})^T W_n(\beta^0_{l^*}) G_n(\beta^0_{l^*})$ and $J_n(\beta^0_{l^*}) = G_n(\beta^0_{l^*})^T W_n(\beta^0_{l^*}) G_n(\beta^0_{l^*})$. Our Assumption 3.2 implies $\Delta_n(\beta^0_{l^*}) = \Delta_0(\beta^0_{l^*}) + o_p(1/n)$ and $J_n(\beta^0_{l^*}) = J_0(\beta^0_{l^*}) + o_p(1/n)$, where $\Delta_0(\beta^0_{l^*}) := g_n(\beta^0_{l^*})^T W_0(\beta^0_{l^*}) G_0(\beta^0_{l^*})$ and $J_0(\beta^0_{l^*}) := G_0(\beta^0_{l^*})^t W_0(\beta^0_{l^*}) G_0(\beta^0_{l^*})$; thus Assumption 3.2 on asymptotic normality of $g_n(\beta^0_{l^*})$ induces condition iii. Finally, we

[3] $\alpha = 0$ in their notation corresponds to the finite dimensional BvM (w.r.t. the total variation metric) in quasi-Bayes framework.

[4] Instead of (4.1) in Chernozhukov and Hong (2003), we have $\beta^0_{l^*}$ characterized as a unique minimizer by Assumption 3.2; this suffices in our analysis.

verify condition iv. Now it follows by Assumption 3.2 that, as $\|\beta_{l^*} - \beta_{l^*}^0\| \leq \delta_n \to 0$, we have

$$\frac{\sqrt{n}\,\|[g_n(\beta_{l^*}) - g_n(\beta_{l^*}^0)] - [g_0(\beta_{l^*}) - g_0(\beta_{l^*}^0)]\|_2}{1 + \sqrt{n}\,\|\beta_{l^*} - \beta_{l^*}^0\|}$$
$$\leq \frac{\sqrt{n}\,\|[G_n(\beta_{l^*}^0) - G_0(\beta_{l^*}^0)]\,(\beta_{l^*} - \beta_{l^*}^0)\|_2 + [n^{-1}\sum_{i=1}^n r(\beta_{l^*}; Z_i) - \mathbb{E}r(\beta_{l^*}; Z_i)]\,\|\beta_{l^*} - \beta_{l^*}^0\|}{1 + \sqrt{n}\,\|\beta_{l^*} - \beta_{l^*}^0\|}$$
$$= o_p(1).$$

Thus our Assumptions satisfies conditions in Proposition 1 in Chernozhukov and Hong (2003), establishing the finite dimensional BvM phenomena of the quasi-Bayes posterior w.r.t. the total variation metric.

□

Proof. Theorem 3.7
By the triangle inequality, for any fixed $l^* \in \mathbb{N}$ we have

$$\begin{aligned}
&\mathcal{D}_{\mathcal{M}_0(\omega)}\left[\mathcal{L}\left(\Pi_n[.|Z^n]\right),\, \mathcal{N}(\beta^*, (nV)^{-1})\right] \\
&\quad\leq \mathcal{D}_{\mathcal{M}_0(\omega)}\left[\mathcal{L}\left(\Pi_n[.|Z^n]\right),\, \mathcal{L}\left(\Pi_n[.|Z^n] \circ P_{l^*}^{-1}\right)\right] \\
&\qquad+ \mathcal{D}_{\mathcal{M}_0(\omega)}\left[\mathcal{L}\left(\Pi_n[.|Z^n] \circ P_{l^*}^{-1}\right),\, \mathcal{N}(\beta^*, (nV)^{-1}) \circ P_{l^*}^{-1}\right] \\
&\qquad+ \mathcal{D}_{\mathcal{M}_0(\omega)}\left[\mathcal{N}(\beta^*, (nV)^{-1}) \circ P_{l^*}^{-1},\, \mathcal{N}(\beta^*, (nV)^{-1})\right].
\end{aligned}$$

First, the parametric BvM phenomena in the quasi-Bayes posteriors in Lemma B.4 shows that the expected value of the second term is $o(1)$ (note that the total variation metric dominates the bounded Lipschitz metric).

For the first term, the $\sqrt{n}-$convergence of the posterior to h_0 in terms of multiscale norm derived in Lemma B.3 implies that we can take large enough $l^* < \infty$ such that its bounded Lipschitz metric can be arbitrarily small asymptotically. Since this holds almost surely, the expected value of the first term can be arbitrary small. Regarding the third term, an application of Proposition 3.5 indicates that the corresponding Gaussian measure is tight; hence taking large enough $l^* < \infty$ makes the third term arbitrary small. Combining all the arguments above completes the proof.

B.3. **Proofs of the results in Section 3.3.** We use the following Lemma, which is an extension of Lemma B.3 in Ročková and Rousseau (2023) into our spike-and-slab prior.

Lemma B.5. *Let Λ_n^* be defined as in (8). Then, under the same conditions as in Theorem 3.9, there exists an event $\mathcal{A}_n^*$ such that $\mathbb{P}(\mathcal{A}_n^*) = 1 + o(1)$ as $n \to \infty$ on which Λ_n^* satisfies the following properties;*

(1). *For some $\gamma^* > 0$ large enough with $\mathcal{L}_n(\gamma^*)$ defined as in (B.1),*

$$\Lambda_n^* \supseteq \mathcal{L}_n(\gamma^*).$$

(2). *For some $\gamma_* > 0$ small enough with $d_l(x, \gamma_*)$ defined as in Lemma A.2,*

$$l \leq \inf_{x \in I_{lk}} d_l(x, \gamma_*),$$

for any $(l, k) \in \Lambda_n^*$.

Proof. Let

$$L_1 = \{\lambda : \lambda \supseteq \mathcal{L}_n(\gamma^*)\} \quad \text{and} \quad L_2 = \{\lambda : l \leq \inf_{x \in I_{lk}} d_l(x, \gamma_*),\ \forall (l, k) \in \lambda\},$$

and define $\mathcal{A}_n^* := \{Z^n : \Pi_n[L_1 \cap L_2 | Z^n] \geq 3/4\}$. Note that $\Pi_n[L_1^c | Z^n] = o_p(1)$ by Lemma B.1. Moreover we have $\Pi_n[L_2^c | Z^n] = o_p(1)$. To see this, suppose that the statement is not true. Then there exists λ and some node $(l, k_l(x)) \in \lambda$ such that $l > \inf_{x \in I_{lk}} d_l(x, \gamma_*)$ and that the posterior puts strictly positive mass on λ asymptotically. It follows by Lemma A.1 that $|\beta^0_{lk_l(x)}| \leq \gamma_* \sqrt{\log n / n}$ for some $x \in [0, 1]$, i.e., $(l, k_l(x)) \notin \mathcal{L}_n(\gamma_*)$. But this contradicts Lemma B.1 (B.3); this establishes the claim. Hence, these observations combined with an union bound lead to $\mathbb{P}(\mathcal{A}_n^*) = 1 + o(1)$.

Now on the event $\mathcal{A}_n^*$, for any $(l, k) \in \mathcal{L}_n(\gamma^*)$, we have

$$\Pi_n[(l, k) \in \lambda | Z^n] \geq \Pi_n[L_1 | Z^n] \geq 3/4 > 1/2.$$

Hence we have $\mathcal{L}_n \subseteq \Lambda_n^*$ and get the first property. For the second one, note that, on the event $\mathcal{A}_n^*$, for any (l, k) such that $l > \inf_{x \in I_{lk}} d_l(x, \gamma_*)$, we have $\Pi_n[(l, k) \in \lambda | Z^n] < 1/4 < 1/2$. This implies $(l, k) \notin \Lambda_n^*$ and the second property follows. □

Proof. Theorem 3.9

(i). Proof of (12): diameter of C_n

We first establish the result on the diameter of C_n. For any $f, g \in C_n$, we have

$$\begin{aligned}
&\sup_{x \in [0,1]} \left[(\nu_n \xi_n(x))^{-1} |f(x) - g(x)| \right] \\
&\leq \sup_{x \in [0,1]} \left[(\nu_n \xi_n(x))^{-1} \left(|f(x) - \hat{h}_{med}(x)| + |\hat{h}_{med}(x) - g(x)| \right) \right] \\
&\leq 2 \sup_{x \in [0,1]} \left[\frac{\sigma_n(x)}{\nu_n \xi_n(x)} \right].
\end{aligned}$$

On the event $\mathcal{A}_n^*$ defined in Lemma B.5, for any $x \in [0, 1]$ we have

$$\begin{aligned}
\frac{\sigma_n(x)}{\nu_n \xi_n(x)} &\leq \nu_n^{-1} \left(\frac{n}{\log n} \right)^{\frac{t(x)}{2t(x)+1} - \frac{1}{2}} \sum_{l=0}^{L_{max}-1} 2^{l/2} \mathbf{1}\{(l, k_l(x)) \in \Lambda_n^*\} \\
&\lesssim \nu_n^{-1} \times 2^{d_l(x, \gamma_*)/2} \left(\frac{n}{\log n} \right)^{\frac{t(x)}{2t(x)+1} - \frac{1}{2}} = O(1),
\end{aligned}$$

where the last line follows from $2^{d_l(x, \gamma_*)} \lesssim (n / \log n)^{1/(2t(x)+1)}$, which is an immediate consequence from Lemma A.2.

(ii). Proof of (10): coverage of C_n

We here prove the coverage property of C_n in (10). To this end, given the construction

of C_n, it suffices to show (1) $\mathbb{P}(h_0 \in C_n^1) = 1 - \alpha + o(1), \alpha \in (0,1)$, and (2) $\mathbb{P}(h_0 \in C_n^2) = 1 + o(1)$.

For (1), this is a consequence from Theorem 5 in Castillo and Nickl (2014) combined with Proposition 3.5.

For (2), we use the following two Lemmas.

Lemma B.6. *Let*

$$\mathcal{B}_n^* = \mathcal{A}_n^* \cap \left\{ g(\beta_{\Lambda_n}^0; Z_i) \lesssim \sqrt{\log n / n} \right\}.$$

Then for some $\gamma^* \geq \gamma_* > 0$, we have $\mathbb{P}(\mathcal{B}_n^*) \to 1$ as $n \to \infty$.

Proof. This immediately follows from Lemma B.5 and $g(\beta_{\Lambda_n}^0; .) = O_p(1/\sqrt{n})$, i.e., it is in a Donsker class. □

Lemma B.7. *Let Assumption 3.8 hold and define a sequence $s_n(x)$ such that $2^{s_n(x)} \asymp (n/\log n)^{1/(2t(x)+1)}$. Then for any $x \in [0,1]$ and $l \geq s_n(x) \geq j_0$, we have*

$$|\beta_{lk_l(x)}^0| \gtrsim 2^{-s_n(x)[t(x)+1/2]}.$$

Proof. The argument here is similar to Proposition 3 in Hoffmann, Rousseau, and Schmidt-Hieber (2015). By Assumption 3.8, there exists some $\tau > 0$ and $j_0 \in \mathbb{N}$ such that, for $j \geq j_0$,

$$2^{-jt(x)}\tau \leq |P_{<j}h_0(x) - h_0(x)| \leq \sum_{l \geq j} 2^{l/2} |\beta_{lk_l(x)}^0|.$$

For all $N \in \mathbb{N}$, there exists $j \geq s_n(x) \geq j_0$ such that

$$\begin{aligned}
|\beta_{jk_j(x)}^0| &\geq \frac{1}{N} \sum_{l=s_n(x)}^{s_n(x)+N-1} |\beta_{lk_l(x)}^0| \\
&\geq \frac{2^{-(s_n(x)+N)/2}}{N} \left(\sum_{l=s_n(x)}^{\infty} 2^{l/2} |\beta_{lk_l(x)}^0| - \sum_{l=s_n(x)+N}^{\infty} 2^{l/2} |\beta_{lk_l(x)}^0| \right) \\
&\geq \frac{2^{-(s_n(x)+N)/2}}{N} \left(2^{-s_n(x)t(x)}\tau - 2^{-(s_n(x)+N)t(x)}\tau \right) \\
&\geq 2^{-s_n(x)[t(x)+1/2]} \times \frac{2^{-N/2+1}\tau}{N}.
\end{aligned}$$

□

An implication of Lemma B.7 is that, for each $x \in [0,1]$, there exists $l \geq s_n(x) \geq j_0$ such that $|\beta_{lk_l(x)}^0| \gtrsim \sqrt{\log n / n}$. Hence we have a lower bound for $\sigma_n(x)$ as follows

$$\sigma_n(x) \geq \nu_n \sqrt{\frac{\log n}{n}} |\phi_{lk_l(x)}| \geq \nu_n \sqrt{\frac{\log n}{n}} 2^{s_n(x)/2} \geq \nu_n \left(\frac{\log n}{n} \right)^{\frac{t(x)}{2t(x)+1}}.$$

Moreover the similar arguments as in the proof of Lemma 3.3 and Theorem 3.4 show that, on the event $\mathcal{B}_n^*$, we have

$$\sup_{x\in[0,1]} \xi_n(x)^{-1}|h_0(x) - \hat{h}_{med}(x)| \leq D,$$

for some large $D > 0$. Combining the lower bound for $\sigma_n(x)$ yields

$$\mathcal{B}_n^* \subseteq \left\{ \sup_{x\in[0,1]} \left[\frac{1}{\sigma_n(x)} |h_0(x) - \hat{h}_{med}(x)| \right] \leq 1 \right\}.$$

Since $\mathbb{P}(\mathcal{B}_n^*) \to 1$ as shown in Lemma B.6, we have shown (2) $\mathbb{P}(h_0 \in C_n^2) = 1 + o(1)$ and the coverage property in 10 is established.

(iii). Proof of (11): the posterior contraction to C_n
By construction of C_n, it suffices to show $\Pi_n[C_n^2|Z^n] = 1 + o_p(1)$. Note that

$$\begin{aligned}
&\mathbb{E}_{h_0}\Pi_n \left[h : \sup_{x\in[0,1]} \frac{1}{\sigma_n(x)} |h(x) - \hat{h}_{med}(x)| \leq 1 | Z^n \right] \\
&\geq \mathbb{E}_{h_0}\Pi_n \left[h : \sup_{x\in[0,1]} \frac{1}{\sigma_n(x)} |h(x) - h_0(x)| \leq 1 | Z^n \right] \times \mathbf{1}_{\mathcal{B}_n^*} + o(1) \\
&= 1 + o(1),
\end{aligned}$$

where $\mathbb{P}(\mathcal{B}_n^{*c}) = o(1)$ and Theorem 3.4 are used.

References

Ai, Chunrong and Xiaohong Chen (2003). "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions". *Econometrica* 71(6), pp. 1795–1843.

Armstrong, Timothy B. (2021). "Adaptation Bounds for Confidence Bands under Self-Similarity". *Bernoulli* 27(2), pp. 1348–1370.

Belloni, Alexandre et al. (2015). "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results". *Journal of Econometrics* 186(2), pp. 345–366.

Castillo, Ismaël and Richard Nickl (2013). "Nonparametric Bernstein–von Mises Theorems in Gaussian White Noise". *The Annals of Statistics* 41(4), pp. 1999–2028.

– (2014). "On the Bernstein–von Mises Phenomenon for Nonparametric Bayes Procedures". *The Annals of Statistics* 42(5), pp. 1941–1969.

Castillo, Ismaël and Veronika Ročková (2021). "Uncertainty Quantification for Bayesian CART". *The Annals of Statistics* 49(6), pp. 3482–3509.

Chaari, Lotfi et al. (2016). "A Hamiltonian Monte Carlo Method for Non-Smooth Energy Sampling". *IEEE Transactions on Signal Processing* 64(21), pp. 5585–5594.

Chaudhuri, Probal (1991). "Nonparametric Estimates of Regression Quantiles and Their Local Bahadur Representation". *The Annals of Statistics* 19(2), pp. 760–777.

Chen, Xiaohong (2007). "Large Sample Sieve Estimation of Semi-Nonparametric Models". *Handbook of Econometrics* 6, pp. 5549–5632.

Chen, Xiaohong, Timothy Christensen, and Sid Kankanala (2025). "Adaptive Estimation and Uniform Confidence Bands for Nonparametric Structural Functions and Elasticities". *The Review of Economic Studies* 92(1), pp. 162–196.

Chen, Xiaohong and Timothy M. Christensen (2015). "Optimal Uniform Convergence Rates and Asymptotic Normality for Series Estimators under Weak Dependence and Weak Conditions". *Journal of Econometrics* 188(2), pp. 447–465.

Chen, Xiaohong and Demian Pouzo (2012). "Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals". *Econometrica* 80(1), pp. 277–321.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato (2014). "Anti-Concentration and Honest, Adaptive Confidence Bands". *The Annals of Statistics* 42(5), pp. 1787–1818.

Chernozhukov, Victor and Han Hong (2003). "An MCMC Approach to Classical Estimation". *Journal of Econometrics* 115(2), pp. 293–346.

Florens, Jean-Pierre and Anna Simoni (2012). "Nonparametric Estimation of an Instrumental Regression: A Quasi-Bayesian Approach Based on Regularized Posterior". *Journal of Econometrics* 170(2), pp. 458–475.

Freedman, David (1999). "Wald Lecture: On the Bernstein-von Mises Theorem with Infinite-Dimensional Parameters". *The Annals of Statistics* 27(4), pp. 1119–1141.

Freyberger, Joachim and Yoshiyasu Rai (2018). "Uniform Confidence Bands: Characterization and Optimality". *Journal of Econometrics* 204(1), pp. 119–130.

Gach, Florian, Richard Nickl, and Vladimir Spokoiny (2013). "Spatially Adaptive Density Estimation by Localised Haar Projections". *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 49(3), pp. 900–914.

Giné, Evarist and Richard Nickl (2010). "Confidence Bands in Density Estimation". *The Annals of Statistics* 38(2), pp. 1122–1170.

– (2021). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge university press.

Hansen, Lars Peter (1982). "Large Sample Properties of Generalized Method of Moments Estimators". *Econometrica* 50(4), pp. 1029–1054.

Hoffmann, Marc, Judith Rousseau, and Johannes Schmidt-Hieber (2015). "On Adaptive Posterior Concentration Rates". *The Annals of Statistics* 43(5), pp. 2259–2295.

Kato, Kengo (2013). "Quasi-Bayesian Analysis of Nonparametric Instrumental Variables Models". *The Annals of Statistics* 41(5), pp. 2359–2390.

Kim, Jae-Young (2002). "Limited Information Likelihood and Bayesian Analysis". *Journal of Econometrics* 107(1-2), pp. 175–193.

Kleijn, B.J.K. and A. W. van der Vaart (2012). "The Bernstein-von-Mises Theorem under Misspecification". *Electronic Journal of Statistics* 6, pp. 354–381.

Liao, Yuan and Wenxin Jiang (2011). "Posterior Consistency of Nonparametric Conditional Moment Restricted Models". *The Annals of Statistics* 39(6), pp. 3003–3031.

Low, Mark G. (1997). "On Nonparametric Confidence Intervals". *The Annals of Statistics* 25(6), pp. 2547–2554.

Müller, Ulrich K. (2013). "Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix". *Econometrica* 81(5), pp. 1805–1849.

Newey, Whitney K. (1997). "Convergence Rates and Asymptotic Normality for Series Estimators". *Journal of Econometrics* 79(1), pp. 147–168.

Ray, Kolyan (2017). "Adaptive Bernstein–von Mises Theorems in Gaussian White Noise". *The Annals of Statistics* 45(6), pp. 2511–2536.

Ročková, Veronika and Judith Rousseau (2023). "Ideal Bayesian Spatial Adaptation". *Journal of the American Statistical Association* 119(547), pp. 2078–2091.

Stone, Charles J. (1982). "Optimal Global Rates of Convergence for Nonparametric Regression". *The Annals of Statistics* 10(4), pp. 1040–1053.

Tsybakov, Alexandre B. (2009). *Introduction to Nonparametric Estimation*. Springer New York.

Yoo, William Weimin and Subhashis Ghosal (2016). "Supremum Norm Posterior Contraction and Credible Sets for Nonparametric Multivariate Regression". *The Annals of Statistics* 44(3), pp. 1069–1102.