Adaptive Supremum Norm Posterior Contraction: Spike-and-Slab Priors and Anisotropic Besov Spaces

William Weimin Yoo\textsuperscript{1}, Judith Rousseau\textsuperscript{2} and Vincent Rivoirard\textsuperscript{2}

\textit{Leiden University\textsuperscript{1}, Université Paris Dauphine\textsuperscript{2}}

Abstract

Supremum norm loss is intuitively more meaningful to quantify estimation error in statistics. In the context of multivariate nonparametric regression with unknown error, we propose a Bayesian procedure based on spike-and-slab prior and wavelet projections to estimate the regression function $f$ and its derivatives. We show that their posteriors contract to the truth optimally and adaptively under $L_\infty$-loss. We discovered that there is a lower limit in the range of smoothness that we can adapt to and this limit grows with dimension of the function’s domain. The master theorem through exponential error test used in Bayesian nonparametrics was not adequate to deal with this problem, and we developed a new idea by bounding posterior under the regression model with a posterior arising from some quasi-white noise model, where the latter model greatly simplifies our calculations.

Keywords: Supremum norm, Adaptive posterior contraction, Nonparametric regression, Spike-and-Slab, Wavelet tensor products, Anisotropic Besov space, Type II error.

MSC2010 classifications: Primary 62G08; secondary 62G05, 62G10, 62G20

1 Introduction

Consider the nonparametric multivariate regression model

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $Y_i$ is a response variable, $X_i$ is covariate, and $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed (i.i.d.) as $N(0, \sigma^2)$ with unknown $0 < \sigma < \infty$. Each $X_i$ takes values in some rectangular region in $\mathbb{R}^d$, which is assumed to be $[0, 1]^d$ without loss of generality. The covariates can be deterministic or are sampled from a uniform distribution on $[0, 1]^d$ independent of $\varepsilon_i$. There is some freedom in choosing the locations of the fixed covariates, as long as its empirical distribution can be approximated by a uniform distribution with an error of at most $n^{-1}$.

Suppose we observed $(Y_i, X_i), i = 1, \ldots, n$, then our main problem is to recover or estimate the unknown $f$ and its mixed partial derivatives. In the literature, recovery is performed by minimizing certain loss functions, with the $L_2$ or integrated mean square error being the most common. However, other choices of loss, especially the supremum norm or $L_\infty$, is also of interest. Unlike the $L_2$-loss, the $L_\infty$-loss is intuitively more meaningful and hence a more natural distance to use to quantify the “difference” between two functions.
Moreover, $L_\infty$-distance is used to construct simultaneous credible bands, which are visually more interpretable than $L_2$-credible sets. Also, it can be used to solve other problems such as function mode estimation discussed in [26].

Adaptive $L_2$-posterior contraction is a well-studied topic in Bayesian nonparametrics, where optimal procedures have been proposed for white noise models, inverse problems, nonparametric regression and density estimation ([1, 14, 11, 5, 18, 19, 23, 22]). Results on $L_\infty$-contraction are much more limited. In the non-adaptive case, Giné and Nickl [8] studied contraction rates in $L_r$-metric, $1 \leq r \leq \infty$, and obtained optimal rate using conjugacy for the Gaussian white noise model, and a rate for density estimation based on random wavelet series and Dirichlet process mixture, by using a testing approach. In the same context, Castillo [3] introduced techniques based on semiparametric Bernstein-von Misses theorems to obtain optimal $L_\infty$-contraction rates. Scricciolo [17] applied the techniques of [8] to obtain $L_\infty$-rates using Gaussian kernel mixtures prior for analytic true densities. Using B-splines tensor product with Gaussian coefficients and by conjugacy arguments, Yoo and Ghosal [25] established optimal $L_\infty$-posterior contraction rate for estimating multivariate regression function and its mixed partial derivatives.

To the best of our knowledge, there are only two papers on the adaptive case. In [10], the authors established optimal $L_\infty$-contraction rates for the Gaussian white noise model and gave an existential result for density estimation; while in Chapter 3 of the thesis by [21], a near optimal rate is given implicitly through a result on credible bands for regression models. For models beyond the white noise, the first mentioned paper used a complicated sieve construction to prove the existence of a procedure, which is not readily implementable in practice; while the latter paper, based on scaled Brownian motion prior, can only adapt up to smoothness of 2.

In this paper, we introduce a concrete hierarchical Bayesian method to estimate $f$ and its derivatives adaptively under the $L_\infty$-loss for nonparametric multivariate regression models. We first represent $f$ as a finite combination of tensor product wavelet basis, and endow the basis coefficients with a spike-and-slab prior. We further endow the error variance $\sigma^2$ with a continuous prior density with support on $(0, \infty)$.

Our main result shows that spike-and-slab priors with appropriate weights can estimate $f$ and its derivatives optimally and adaptively under $L_\infty$-loss, in the sense that the resulting sup-norm posterior contraction rates match with the minimax rates for this problem. The scope of our result is quite general, in that we require only the slab prior density to be bounded from above and bounded from below in some interval, and this encompasses (nonconjugate) distributions such as Gaussian, sub-Gaussian, Laplace, uniform and most $t$-distributions. The Gaussian assumption of our errors in (1.1) is simply a working model to derive expressions for the posterior, and our results will hold even if the model is misspecified and the actual data generation mechanism is sub-Gaussian.

Although research in statistics has focused on the white noise model, and translate results in this model to regression and density estimation through Le Cam’s asymptotic equivalence, it is unclear whether such results still hold in the Bayesian setting, and in view of the fact that our design points are not exactly discrete uniform, but are close to uniform up to some
error. Working in general regression models reveal a phenomenon that is non-existent in the white noise model. Namely, we discovered that there is a lower limit of adaptation that increases with dimension, such that we can only adapt to smoother functions at higher dimensions. This lower bound occurs for models beyond the white noise and for isotropic $\alpha$-smooth $f$, can be succinctly expressed as $\alpha > d/2$.

In Bayesian nonparametrics, the current state-of-the-art technique to derive contraction rates is the “master” theorem developed by [6, 20, 7]. One of the main criterion of this approach is the existence of a test for the hypotheses $H_0 : f = f_0$ against $H_1 : f \in \{f : \|f - f_0\|_\infty > M\epsilon_n\}$ that has exponentially decreasing Type I and II errors. However, we show that this is impossible to achieve for sup-norm alternatives, as its Type II error has at least polynomial rate of decrease for any test functions. For exponential error test to exists, we show that the null and the alternative hypotheses must be further separated in $L_\infty$-norm, and this increase in separation results in the contraction rate being inflated by the same factor (see also [10] for a related discussion on this sub-optimality issue).

Therefore, this motivated us to introduce a new idea for contraction rates computation, that will yield the correct sup-norm adaptive rate without additional logarithmic factors. In our approach, we bound the posterior under the regression model with a posterior arising from a quasi-white noise model, with “quasi” refers to the use of a scaling based on the wavelet basis matrix rather than the standard $n^{-1/2}$ in white noise models. In a series of steps and intersecting with appropriately chosen events, this is achieved by reducing the regression likelihood to a likelihood that resembles and retains the component-wise structure of a white noise model, where the latter likelihood structure greatly simplifies our calculations and thus giving our method a certain Bayesian “asymptotic equivalence” flavor.

The main technical challenge of this approach is the handling of discretization inherent in regression models, as many convenient wavelet properties are lost when working in the discrete domain with finite data. As an example, the wavelet basis matrix constructed from wavelets evaluated at the design points is not orthogonal and this complicates analysis. We solve this problem by approximating discrete quantities or sums by its continuous or integral versions, and thus incurring approximation errors that we propagate throughout our calculations, while at the same time keeping them under control so as not to overwhelm stochastic and truncation errors (bias) in other parts of the problem. Another generalization we considered is to allow $f_0$ to be anisotropic, i.e., different smoothness in different dimensions, and we introduce a version of the anisotropic Besov space (see Definition 3.1 below) suited for our analysis and assumes $f_0$ belongs to this space.

The paper is organized as follows. The next section introduces notations. Section 3 describes the prior and the assumptions used in this paper. The main result on adaptive $L_\infty$-posterior contraction, for $f$ and its mixed partial derivatives are presented in Section 4, and this is followed by a discussion on the lower limit of adaptation. The inadequacy of the master theorem is detailed in Section 5. Section 6 contains proofs of all main results and is further divided into 3 subsections. The proof of our $L_\infty$-contraction result is in Section 6.1, we introduce our posterior bounding technique in Section 6.2 and the rest of the proofs are in Section 6.3. The last Section 7 contains technical lemmas used throughout the proofs, where
some results such as continuous approximations on discrete objects and $L_2$-contraction for spike-and-slab priors are of independent interests.

2 Notations

Given two numerical sequences $a_n$ and $b_n$, $a_n = O(b_n)$ or $a_n \lesssim b_n$ means $a_n/b_n$ is bounded, while $a_n = o(b_n)$ means $a_n/b_n \to 0$. If $a_n \asymp b_n$, then we have $a_n = O(b_n)$ and $b_n = O(a_n)$.

For stochastic sequence $Z_n$, $Z_n = O_P(a_n)$ means $P(|Z_n| \leq C a_n) \to 1$ for some constant $C > 0$, while $Z_n = o_P(a_n)$ means $Z_n/a_n$ converges to 0 in $P$-probability. Define $\mathbb{Z}$ to be the set of integers, $\mathbb{N} = \{1, 2, \ldots \}$ to be the set of natural numbers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

Define $\|\mathbf{x}\|_p = (\sum_{k=1}^d |x_k|^p)^{1/p}$, $1 \leq p < \infty$, $\|\mathbf{x}\|_\infty = \max_{1 \leq k \leq d} |x_k|$, and write $\|\mathbf{x}\|$ for $\|\mathbf{x}\|_2$, the Euclidean norm. For $f : U \to \mathbb{R}$ on some bounded set $U \subseteq \mathbb{R}^d$ with interior points, let $\|f\|_p$ be the $L_p$-norm, and $\|f\|_\infty = \sup_{x \in U} |f(x)|$. For $\mathbf{r} = (r_1, \ldots, r_d)^T \in \mathbb{N}_0^d$, let $D^\mathbf{r}$ be the partial derivative operator $\partial^{(|\mathbf{r}|)} / \partial x_1^{r_1} \cdots \partial x_d^{r_d}$, where $|\mathbf{r}| = \sum_{k=1}^d r_k$. For a set $\mathcal{A}$, let $\mathbb{1}_\mathcal{A}$ be the indicator function on $\mathcal{A}$. For a vector $\mathbf{x}$, we write $x_{j}$ to be its $j$th component with $j$ possibly be multi-index $(j_1, \ldots, j_d)^T$, and in that case we let the entries be ordered lexicographically.

3 Wavelet series with spike-and-slab prior

Since our domain of interest is $[0,1]^d$, we will use the $\eta$-regular, boundary corrected wavelets introduced by Cohen-Daubechies-Vial (CDV) in Section 4 of [4], at each dimension $l = 1, \ldots, d$. Let $\varphi_l$ and $\psi_l$ be the CDV father and mother wavelets. The precise definition of $\eta$-regularity can be found in Definition 4.2.14 of [9] and can be thought of as encoding smoothness information about the wavelets. The CDV wavelets are compactly supported and we further assume that the derivatives $\varphi_l^{(r_l)}$, $\psi_l^{(r_l)}$ are uniformly bounded for $r_l < \eta + 1$.

Define $d$-dimensional indices $\mathbf{m} = (m_1, \ldots, m_d)^T$, $\mathbf{N} = (N_1, \ldots, N_d)^T$, $\mathbf{j} = (j_1, \ldots, j_d)^T$ and $\mathbf{k} = (k_1, \ldots, k_d)^T$. For $\mathbf{x} = (x_1, \ldots, x_d)^T$, we construct tensor products of the father wavelet as $\varphi_{\mathbf{N}, \mathbf{m}}(\mathbf{x}) = \prod_{l=1}^d 2^{N_l/2} \varphi_l(2^{N_l} x_l - m_l)$. Let $\mathcal{I}$ be the set of $2^d - 1$ sequences of the form $(i_1, \ldots, i_d)$, such that each $i_l$ can be 0 or 1, but excluding the case where $i_l = 0$ for all $l$. For $\mathbf{i} \in \mathcal{I}$, we construct mother wavelet tensor products as $\psi_{\mathbf{j}, \mathbf{k}}^{\mathbf{i}}(\mathbf{x}) = \prod_{l=1}^d 2^{h/2} \psi_l^{i_l}(2^{h} x_l - k_l)$, where $\psi_0 = \varphi_l$ and $\psi_1 = \psi_l$. For example when $d = 2$, then $\mathcal{I} = \{(0,1), (1,0), (1,1)\}$ and the mother wavelet tensor products are translated and dilated versions of $\{\varphi_1(x_1)\psi_2(x_2), \psi_1(x_1)\varphi_2(x_2), \psi_1(x_1)\psi_2(x_2)\}$.

Since the CDV wavelets are unconditional $L_2$-bases, we can expand $f$ in the multivariate regression model of (1.1) using these bases, and this leads us to consider the following
hierarchical priors to study sup-norm contraction:

\[
f(x) = \sum_{m_1=0}^{2^{N_1}-1} \cdots \sum_{m_d=0}^{2^{N_d}-1} \varphi_{N,m}(x) + \sum_{j_1=N_1}^{J_{n_1}-1} \sum_{j_2=N_2}^{J_{n_2}-1} \cdots \sum_{j_d=N_d}^{J_{n_d}-1} \sum_{i \in I} \theta_{j,k,i} \psi_{j,k}(x),
\]

\[
\varphi \sim p(\cdot), \quad \theta_{j,k,i} \sim (1 - \omega_{j_1,\ldots,j_d,n}) \delta_0(\cdot) + \omega_{j_1,\ldots,j_d,n} p(\cdot),
\]

\[
\sigma \sim \pi_\sigma.
\]  

(3.1)

The prior on the mother coefficients are what we call a spike-and-slab, with the spike part corresponding to the point mass at 0 (\(\delta_0\) is the Dirac function) and the slab part some density \(p(\cdot)\) on \(\mathbb{R}\). With appropriate chosen weights \(\omega_{j_1,\ldots,j_d,n}\), it does a form of model selection by zeroing “unimportant” coefficients. Observe that we only assign spike-and-slab prior on the mother wavelet coefficients, this is done to prevent overly sparse models by allowing father coefficients to capture global structures of \(f\). The truncation point \(J_{n,l}\) at some fixed \(l = 1, \ldots, d\) is a sequence of positive integers increasing with \(n\), such that \(\prod_{l=1}^d 2^{J_{n,l}} = \sqrt{n/\log n}\) if the design points are fixed, and \(\prod_{l=1}^d 2^{J_{n,l}} = (n/\log n)^{1/4}\) for random design points. Here, \(N_l\) is a positive integer such that \(2^{N_l} \geq 2n_l\). Also, \(\{\varphi_m\}, \{\theta_{j,k}\}, \sigma^2\) are mutually independent with each other. For the spike-and-slab weights, we assume \(n^{-\lambda} \leq \omega_{j_1,\ldots,j_d,n} \leq \min\{2^{-\sum_{l=1}^d j_l(1+\mu_l)}, 1/2\}\), where \(\lambda > 0\) and \(\mu_{\min} > 1/2\) with \(\mu_{\min} = \min_{1 \leq l \leq d} \mu_l\). Here, \(p(\cdot)\) is such that \(p_{\max} = \sup_{x \in \mathbb{R}} p(x) < \infty\) and for some \(R_0 > 0\),

\[
\inf_{x \in [-R_0, R_0]} p(x) = p_{\min} > 0.
\]  

(3.2)

Examples of \(p(\cdot)\) include the Gaussian, sub-Gaussian, Laplace, the uniform \([-R_0, R_0]\) and most \(t\)-distributions. We let \(\pi_\sigma\) be a positive and continuous prior density with support on \((0, \infty)\), e.g., inverse gamma distribution. If the covariates \(X_i = (X_{i1}, \ldots, X_{id})^T\) for \(i = 1, \ldots, n\) are fixed, we assume they are chosen such that

\[
\sup_{x \in [0,1]^d} |G_n(x) - U(x)| = O\left(\frac{1}{n}\right),
\]  

(3.3)

where \(U(x)\) is the cumulative distribution function of a uniform on \([0,1]^d\), and \(G_n(x)\) is the empirical cumulative distribution function of \(\{X_i, i = 1, \ldots, n\}\), that is, \(G_n(x) = n^{-1} \sum_{i=1}^n 1_{\{X_i \leq x\}}\). This is particularly true if we use a discrete uniform design, that is for \(n = m^d\) for some \(m \in \mathbb{N}\), \(X_i \in \{(j-1)/(m-1) : j = 1, \ldots, m\}^d\) with \(i = 1, \ldots, n\). If the design points are random, we assume that \(X_1, \ldots, X_n \overset{i.i.d.}{\sim} U(x)\), and we have \(\sup_{x \in [0,1]^d} |G_n(x) - U(x)| = O_P(n^{-1/2})\) by Donsker’s theorem. In this paper, we will only prove results on adaptive posterior contraction rate based on fixed design points, the random case can be treated by conditioning on \(X_i, i = 1, \ldots, n\).

To study \(L_\infty\)-posterior contraction for mixed partial derivatives, we apply the differentiation operator \(D^r\) on both sides of the wavelet expansion in (3.1) to yield

\[
D^r f = \sum_m \varphi_{m} D^r \varphi_{N,m} + \sum_{j,k,i} \theta_{j,k,i} D^r \psi_{j,k},
\]  

5
where the priors on \( \vartheta_m \) and \( \theta_{j,k} \) the same as in (3.1). To study both \( f \) and its derivatives in the same framework, we adopt the convention \( D^0 f \equiv f \). We note that objects such as \( D^r \varphi_{N,m} \) and \( D^r \psi_{j,k} \) are called vaguelettes tensor products (see [2]).

To study frequentist properties and derive contraction rates of our posterior, we assume the existence of an underlying true function \( f_0 \), such that \( f_0 \) belongs to an anisotropic Besov space as defined below. Denote first \( \alpha^* \) to be the harmonic mean of \( \alpha = (\alpha_1, \ldots, \alpha_d)^T \), i.e., \( (\alpha^*)^{-1} = d^{-1} \sum_{l=1}^d \alpha_l^{-1} \).

**Definition 3.1** (Anisotropic Besov space). The anisotropic Besov function space \( \mathcal{B}^\alpha_{p,q} \) for \( \alpha = (\alpha_1, \ldots, \alpha_d)^T \) such that \( 0 < \alpha_l < \eta_l + 1, l = 1, \ldots, d \) and \( 1 \leq p, q \leq \infty \) is given as

\[
\mathcal{B}^\alpha_{p,q} \equiv \begin{cases} \{ f \in L_p([0,1]^d) : \|f\|_{\mathcal{B}^\alpha_{p,q}} < \infty \}, & 1 \leq p < \infty \\ \{ f \in C_0([0,1]^d) : \|f\|_{\mathcal{B}^\alpha_{p,q}} < \infty \}, & p = \infty \end{cases}
\]

with \( C_0([0,1]^d) \) the space of uniformly continuous functions on \([0,1]^d\), and the anisotropic Besov norm \( \|f\|_{\mathcal{B}^\alpha_{p,q}} \) is

\[
\begin{align*}
(\sum_m \|\langle f, \varphi_{N,m}\rangle\|^p)^{\frac{1}{p}} + \left( \sum_j 2^{qj} \sum_{l=1}^d \alpha_l j_l \left[ \frac{1}{2^{m_l}} - \frac{1}{2^{m_l} + 1} \right] \right) \left( \sum_k \sum_{i \in \mathcal{I}} \|\langle f, \psi_{j,k} \rangle\|^p \right)^{\frac{1}{p}}, & \quad 1 \leq q < \infty \\
(\sum_m \|\langle f, \varphi_{N,m}\rangle\|^p)^{\frac{1}{p}} + \sup_j 2^{qj} \sum_{l=1}^d \alpha_l j_l \left[ \frac{1}{2^{m_l} + 1} - \frac{1}{2^{m_l + 1}} \right] \left( \sum_k \sum_{i \in \mathcal{I}} \|\langle f, \psi_{j,k} \rangle\|^p \right)^{\frac{1}{p}}, & \quad q = \infty,
\end{align*}
\]

where we replace the \( \ell_p \)-sequence norm with the maximum norm \( \| \cdot \|_\infty \) when \( p = \infty \).

**Remark 3.1.** If we set \( \alpha_l = \alpha \) and take \( \alpha \to 0 \), then we can define the Besov spaces \( \mathcal{B}^\alpha_{p,q} \), which is the multivariate and anisotropic generalization of its univariate counterpart in Section 4.3.2 of [9]. In this case, we replace \( \sum_{l=1}^d \alpha_l j_l \left[ \frac{1}{2^{m_l}} - \frac{1}{2^{m_l} + 1} \right] \) in the exponent by \( \sum_{l=1}^d j_l/2 \).

Let

\[ K_W(x, y) = \sum_m \varphi_{W,m}(x) \varphi_{W,m}(y), \]

and define the operator \( K_W \) on \( L_p \) such that \( K_W(g)(x) = \int K_W(x, y) g(y) \, dy \) for \( g \in L_p \). Thus, observe that \( K_W(g) \) is the \( L_2 \)-projection of \( g \) onto the subspace spanned by \( \{ \varphi_{N,m} : 0 \leq m_l \leq 2^{N_l} - 1 \} \cup \{ \psi_{j,k}^l : N_l \leq j_l \leq W_l - 1, 0 \leq k_l \leq 2^{j_l} - 1, i \in \mathcal{I} \}, l = 1, \ldots, d \). That is, \( K_W(g) \) has wavelet expansion as in (3.1) but truncated at levels \( W = (W_1, \ldots, W_d)^T \). The proposition below then tells us how well these anisotropic wavelet projections approximate functions in \( \mathcal{B}^\alpha_{p,q} \).

**Proposition 3.2.** Let \( 0 \leq r_l < \eta_l + 1, l = 1, \ldots, d \). For any \( g \in \mathcal{B}^\alpha_{p,q} \), let \( K_W(g) \) be its projected version at level \( W \) as described above, then there exists constant \( C > 0 \) depending on the wavelets used such that

\[
\| D^r K_W(g) - D^r g \|_p \leq C \| D^r g \|_{\mathcal{B}^\alpha_{p,q}} \| \sum_{l=1}^d 2^{-(\alpha_l - r_l)W_l} \|_1.
\]

We are now ready to introduce the assumptions on the underlying true model for (1.1).
Assumption 1. Under the true distribution $P_0$, $Y_i = f_0(X_i) + \varepsilon_i, i = 1, \ldots, n$, where $f_0 \in B^{\alpha}_{p,q}$ and $\varepsilon_i$ are i.i.d. Gaussian with mean 0 and finite variance $\sigma_0^2 > 0$ for $i = 1, \ldots, n$. Here, $\alpha = (\alpha_1, \ldots, \alpha_d)^T \in (0, \infty)^d$ is unknown.

Remark 3.3. Inspection of the main proof shows that we can actually relax the assumption on errors so that they are sub-Gaussian, and hence allowing the model to be possibly misspecified. However, we would need to use the misspecified version of the “testing approach” (Theorem 4.1 of [12]) to prove $L_2$-contraction of the mother wavelet coefficients as part of the overall proof. We refrain from doing this because this will add extra technicalities that is a distraction for the main $L_\infty$-task at hand.

We define $\vartheta^0_m = \langle f_0, \varphi_{N,m} \rangle$ and $\theta^0_{j,k,i} = \langle f_0, \psi^i_{j,k} \rangle$ to be the true wavelet coefficients. We denote $E_0(\cdot)$ as the expectation operator taken with respect to $P_0$ and write $Y = (Y_1, \ldots, Y_n)^T$. Moreover, we write Besov ball of radius $R > 0$ as $B^{\alpha}_{p,q}(R) := \{ f : \|f\|_{B^{\alpha}_{p,q}} \leq R \}$.

4 Adaptive posterior contraction

Before establishing $L_\infty$-contraction rate for $f$, a preliminary key step is to show that the posterior distribution of $\sigma$ is consistent under the hierarchical priors of (3.1). We therefore begin with the following proposition.

Proposition 4.1. Under Assumption 1, we can conclude that for any prior on $\sigma$ with positive and continuous density, the posterior distribution of $\sigma$ is consistent, uniformly over $f_0 \in B^{\alpha}_{p,q}(R)$ for any $R > 0$ and for any $\alpha$ such that $0 < \alpha_l < \eta_l + 1$, where $\eta_l$ is the regularity of the wavelet bases at dimension $l = 1, \ldots, d$.

Using wavelet expansions such as (3.1), we can work with wavelet coefficients instead of $f$ and treat them as component-wise signals we are trying to recover. In all our calculations, the threshold $\sqrt{\log n/n}$ is of crucial importance as it serves as a cutoff to determine statistically which signal is considered “big” or “small”. The speed of which the posterior will contract to the truth in $L_\infty$-norm is then dictated by these two conditions:

1. Signal detection errors, i.e., $|\theta_{j,k}| \geq \sqrt{\log n/n}$ but the true signal is “small” $|\theta^0_{j,k}| \leq \sqrt{\log n/n}$ and vice versa are unlikely to occur under the posterior.

2. The posterior will concentrate within $\sqrt{\log n/n}$-neighborhood around large detectable signals.

Asymptotically, this implies that the spike-and-slab posterior behaves like a local wavelet thresholding operator, which does coefficient-wise thresholding with $\sqrt{\log n/n}$ as threshold. During the course of establishing these conditions, we have to deal with discrete approximation errors as encoded in (3.3), finite truncation error of Proposition 3.2 and stochastic error in our model (1.1). This requires a very delicate balancing of these opposing errors that
we propagate throughout our calculations, and we arrive at our results by ensuring that no single source of error will dominate the others.

Spike-and-slab priors were originally introduced to be used as a variable selection procedure to detect nonzero coefficients. The form of the weights $\omega_{j,n}$ in (3.1) further implies that coefficients at higher resolution levels are more likely to be negligible, and the model is sparser at those levels. The following main results show that by selecting coefficients using appropriate weights, the posteriors of $\alpha$ at each dimension beyond the regularity $\eta_l$ of the wavelets, but it will be seen that there is a lower limit of adaptation present that prevents us to adapt arbitrarily close to 0 (see Section 4.1 for a more thorough discussion).

Therefore, this lead us to formulate our range of adaptation as

$$\mathcal{A}_r = \left\{ \alpha : \frac{2(r_l + 1)\alpha^*d}{2\alpha^* + d} < \alpha_l < \eta_l + 1, l = 1, \ldots, d \right\},$$

and if $r = 0$, we simply write $\mathcal{A}_0$ as $\mathcal{A}$.

**Theorem 4.2.** (Adaptive $L_\infty$-contraction)

(a) For the regression function:
For any $0 < R \leq R_0 - 1/2$ and some constants $\xi, M > 0$,

$$\sup_{\alpha \in \mathcal{A}} \sup_{f_0 \in \mathcal{B}_\infty(R)} E_0 \Pi \left( \|f - f_0\|_\infty > M (n/\log n)^{-\frac{\alpha^*}{2\alpha^* + d}} \right) \leq \frac{(\log n)^d}{n^{\xi}}. \tag{4.2}$$

(b) For mixed partial derivatives:
Let $1 \leq r_l < \eta_l + 1, l = 1, \ldots, d$. Then for any $0 < R \leq R_0 - 1/2$ and some large constant $M > 0$, we have uniformly over $\alpha \in \mathcal{A}_r$ and $f_0 \in \mathcal{B}_\infty(R)$ that

$$E_0 \Pi \left( \|D^r f - D^r f_0\|_\infty > M(n/\log n)^{-\frac{\alpha^*(1-\sum_{l=1}^d r_l/\alpha_l)}{2\alpha^* + d}} \right) \leq \frac{(\log n)^d}{n^{\xi}}. \tag{4.3}$$

**Remark 4.3.** For random design, $\mathcal{A}$ is defined by $\frac{4\alpha^*d}{2\alpha^* + d} < \alpha_l < \eta_l + 1, l = 1, \ldots, d$. In this case, we condition on $(Y, X_1, \ldots, X_n)$ in the posterior distribution. For mixed partial derivatives in (4.3), $\mathcal{A}_r$ is defined through $\frac{2(r_l + 2)\alpha^*d}{2\alpha^* + d} < \alpha_l < \eta_l + 1, l = 1, \ldots, d$.

**Remark 4.4.** For the isotropic case where $\alpha_l = \alpha, l = 1, \ldots, d$, $\mathcal{A}_r$ is defined through $\max\{d/2, \sum_{l=1}^d r_l\} < \alpha < \eta + 1$ for the fixed design case, and $\max\{3d/2, \sum_{l=1}^d r_l\} < \alpha < \eta + 1$ for the random case.

Theorem 4.2 has important implication in frequentist statistics. In particular, the posterior mean as an adaptive point estimator converges uniformly to $f_0$ at the same rate.

**Corollary 4.5.** Let $0 \leq r_l < \eta_l + 1, l = 1, \ldots, d$, then for any $0 < R \leq R_0 - 1/2$,

$$\sup_{\alpha \in \mathcal{A}_r} \sup_{f_0 \in \mathcal{B}_\infty(R)} E_0 \|E(D^r f|Y) - D^r f_0\|_\infty \lesssim (n/\log n)^{-\frac{\alpha^*(1-\sum_{l=1}^d r_l/\alpha_l)}{2\alpha^* + d}}.$$
4.1 The lower bound of adaptation

Theorem 4.2 in particular shows that there is a certain lower limit in the range of smoothness that we can adapt to, and this limit is increasing with $d$ the dimension of the regression function’s domain. To see this point, take $r = 0$ and rearrange the lower bound in (4.1) to $1/d + 1/(2\alpha^*) > 1/\alpha_l$ and sum both sides across $l = 1, \ldots, d$ to get $\alpha^* > d/2$. The purely technical reason for this bound is to ensure that $\|D^r f - D^r f_0\|_\infty$ decreases to 0 at the correct $L_\infty$-rate at high resolution levels.

Interestingly, this lower bound has been observed in the one-dimensional case (see [8, 10, 21, 3]) and appears to be universal when one tries to do fully Bayesian procedures on models beyond Gaussian white noise, e.g., nonparametric regression and density estimation. We argue that this is unlikely an artefact of our proofs since the papers listed employed different methods and arrive at the same lower bound, and concurrently seems to suggest that it is a fundamental (or intrinsic) feature to Bayesian estimation. This implies that at least for spike-and-slab priors, our range of adaptation shrinks and we can only adapt to smoother functions in higher dimensions. This is in contrast to frequentist procedures such as Lepski’s method or wavelet thresholding, where one can adapt the (isotropic) function smoothness $\alpha$ arbitrarily close to 0.

We would like to give some intuition to this phenomenon and to make an interesting connection to Le Cam’s asymptotic equivalence theory. It is known that in the isotropic case, nonparametric regression and white noise are asymptotically equivalent iff $\alpha > d/2$. During our theoretical investigations, we discovered that the regression likelihood in the posterior behaves like a quasi-white noise likelihood in the limit when $\alpha > d/2$ (see Section 6.2 below). Since this lower bound is linked to likelihood functions, we conclude that in order to adapt arbitrarily close to 0, we should avoid choosing tuning parameters (the number of wavelet basis used in our setting) based directly on likelihoods as in hierarchical or empirical Bayesian procedures. Instead, one can for example do Lepski’s method based on posterior mean or threshold posterior median of the wavelet coefficients. We hope to address these issues in future research and design adaptive Bayesian procedures that circumvent this lower bound restriction.

5 The master theorem of Bayesian nonparametrics

As mentioned in the introduction, we developed a new method of deriving contraction rates by comparing regression posterior with a corresponding quasi-white noise version. A statistician that has some acquaintance with modern Bayesian nonparametrics might question the necessity of this innovation, by pointing out the existence of a “master” theorem in the literature, which is the de facto state-of-the-art method in calculating posterior contraction rates (see [6, 20, 7]). As its name suggests, this master theorem is a type of “one theorem for all” and has been deployed in a myriad of models and priors ranging from the simplest Gaussian white noise to general stochastic processes and graphical models. In its most basic version adapted to our present task, the master theorem has 3 criteria ($C_1, C_2, C_3 > 0$ are
some constants):

1. The existence of a test $\phi_n$ for the hypotheses $H_0 : f = f_0$ against $H_1 : f \in \{f \in \mathcal{F}_n : \|f - f_0\|_\infty > M\epsilon_n\}$ with $\mathcal{F}_n$ some appropriately chosen sieve sets of the parameter space, such that its Type I error goes to 0 while its Type II error decreases like $e^{-C_1n\epsilon_n^2}$,

2. The prior puts at least $e^{-C_2n\epsilon_n^2}$ mass on certain Kullback-Leibler neighborhoods of radius $\epsilon_n$ around $f_0$.

3. The prior puts most of its mass in the sieve sets such that $\Pi(\mathcal{F}_n^c) \leq e^{-C_3n\epsilon_n^2}$.

Recently however, research in this area has discovered cases that do not fall in the scope of this master theorem. In our context of $L_\infty$-contraction, Giné and Nickl [8] and Hoffmann et al. [10] found that the master theorem, which corresponds to verifying the 3 conditions above, actually produces suboptimal contraction rates. In the following, we give a more precise explanation to this problem in higher dimensions and thus shed some more light on their results.

The root of the problem is found in the first testing criterion, which can be formulated more rigorously as $\sup_{f \in \mathcal{F}_n : \|f - f_0\|_\infty > M\epsilon_n} E_f(1 - \phi_n) \leq e^{-C_1n\epsilon_n^2}$ for some test function $\phi_n$. However, the proposition below shows that there exists function under $H_1$ with Type II error at least a polynomial rate of decrease. Therefore one cannot achieve $e^{-C_1n\epsilon_n^2}$, or exponential-type decrease in general if the null and alternative is separated apart by $\epsilon_n$ in sup-norm.

In our arguments below, let $\mathcal{F}_n := \{f : \|f - K_{J_n}(f)\|_\infty \lesssim \sum_{l=1}^d 2^{-a_lJ_{n,l}}\}$ be the sieve sets and we treat $\sigma = \sigma_0$ as known. The latter is justified since $\sigma$ is consistent (Proposition 4.1) and hence we can uniformize $\sigma$ over shrinking neighborhoods of $\sigma_0$ and work only with the posterior conditioned on $\sigma$.

Proposition 5.1. Under our regression model (1.1), consider the hypotheses $H_0 : f = f_0$ against $H_1 : f \in \{f : \|f - f_0\|_\infty > M\epsilon_n\}$ with $f_0 \in \mathcal{B}_{\infty,\infty}^\alpha(R)$. Let $\phi_n(X_1, \ldots, X_n; f_0) \rightarrow \{0, 1\}$ be any test function for this nonparametric testing problem such that $E_{f_0}\phi_n \rightarrow 0$, then there exists constant $Q > 0$ such that

$$\sup_{f \in \mathcal{F}_n : \|f - f_0\|_\infty > M\epsilon_n} E_f(1 - \phi_n) \gtrsim n^{-Q},$$

for $\epsilon_n = (n / \log n)^{-a^*/(2a^*+d)}$ the targeted rate.

Polynomial rates are not unique to testing problems with $L_\infty$-separation in the alternative hypothesis. In fact, posterior probabilities on shrinking $L_\infty$ or point-wise $\epsilon_n$-neighborhoods around $f_0$ tend to 0 at polynomial rates (up to a logarithmic factor), e.g., see Theorem 4.2 and Lemmas 6.1, 6.2. Exponential rates are only possible for weaker loss, with decay of the type $e^{-Cn\epsilon_n^2}$ corresponding to the $L_2$-loss and its equivalent metrics.

The previous proposition implies that for exponential error tests to exist, the null and the alternative hypotheses must be further separated in $L_\infty$-norm. To that end, let us introduce a separation factor $\rho_n > 0$ in the alternative $H_1 : f \in \{f \in \mathcal{F}_n : \|f - f_0\|_\infty > M\rho_n\epsilon_n\}$. It is then
instructive to ask how large $\rho_n$ should be so that tests with exponential Type II error start to exist. The following proposition says that $\rho_n$ must be greater than $1 + (n / \log n)^{d/(2(2\alpha^*+d))}$ up to some constant multiple, and this increase in separation results in the contraction rate being inflated by the same factor.

**Proposition 5.2.** For any $\rho_n \geq C[1 + (n / \log n)^{d/(2(2\alpha^*+d))}]$ with some constant $C > 0$, there exists a test $\phi_n$ such that for some constants $C_1, C_II > 0$,

$$E_0\phi_n \leq e^{-Cr_n\alpha^*_n^2}, \quad \sup_{f \in F_n \|f-f_0\|_\infty > M\rho_n e_n} E_f(1 - \phi_n) \leq e^{-C_{II}r_n\alpha^*_n^2}.$$  

Consequently, if we attempted to use the master theorem to prove the first assertion of Theorem 4.2, then there exists a constant $M > 0$ such that as $n \to \infty$,

$$\sup_{\alpha \in \mathcal{A}} \sup_{f_0 \in \mathcal{B}_{2,\infty}(R)} E_0\Pi(\|f - f_0\|_\infty > M\rho_n (n / \log n)^{-\alpha^*/(2\alpha^*+d)}|Y) \to 0.$$  

If we chose the lower bound $\rho_n = C[1 + (n / \log n)^{d/(2(2\alpha^*+d))}]$, then the contraction rate becomes $(n / \log n)^{-\alpha^*/(2\alpha^*+d)}$ and for comparison, Theorem 4.2 tells us that the best rate should be $(n / \log n)^{-\alpha^*/(2\alpha^*+d)}$, and we have an extra polynomial factor. That is why to obtain this optimal rate, we developed the proposed new approach of comparing regression posterior with its quasi-white noise version, and we will fully describe this approach in Section 6.2 below.

6  Proofs

Observe that many of our quantities are indexed by triple multi-indices $(j, k, i)$ such as $\theta_{j,k,i}$. However, since $i \in \mathcal{I}$ is simply an identification indices to dictate which tensor product component is a father or mother wavelet, and coupled with the fact that $\sum_{i \in \mathcal{I}} = 2^d - 1$ does not grow with $n$, we will drop this identification in our calculations below and simply work with $(j, k)$. This is to improve readability and help readers focus on the main ideas instead of the technicalities of working in multiple dimensions.

6.1  Proof of main results up to Section 4.1

**Proof of Proposition 3.2.** A tensor product polynomial has order $m = (m_1, \ldots, m_d)^T$, if it is a linear combination of $\{x_1^{i_1-1} \cdots x_d^{i_d-1}\}$ for $1 \leq i_l \leq m_l, l = 1, \ldots, d$. Recall the wavelet projection operator $K_W(f)(x) = \int K_W(x,y)f(y)dy$ with $K_W(x,y) = \sum_m \varphi_{W,m}(x)\varphi_{W,m}(y)$. We will be using two important properties of $K_W$. The first is

$$K_W(P) = P,$$  

for any tensor product polynomial $P$ with degree less than or equal to $(\eta_1, \ldots, \eta_l)^T$, the regularities of wavelet used at each dimension (see Theorem 4 of Section 2.6 [13]). The second is

$$\|K_W(f)\|_p \leq C_1\|f\|_p,$$  

(6.2)
for some constant $C_1 > 0$ and for any $f \in L_p$. This inequality follows from the calculations contained in Section 3.1.1 of [8]. As a result, $K_W$ is bounded in $L_p$ and it reproduces polynomials.

Define hypercubes $I_k = \prod_{i=1}^d [k_i 2^{-W_i}, (k_i + 1) 2^{-W_i}]$ for $0 \leq k_i \leq 2^{W_i} - 1$ and note that the unit cube $[0,1]^d$ is the sum of these smaller cubes over all $k$. Let $f|_{I_k}$ be the restriction of $f$ onto $I_k$. By Theorem 13.20 of [16], we know that there exists a tensor product Taylor’s polynomial $p_k$ such that

$$\|D^r(g - p_k)|_{I_k}\|_p \leq C_2 \sum_{i=1}^d 2^{-(\alpha_i - r_i)W_i}\|D^{(\alpha_i - r_i)e_i}D^r g|_{I_k}\|_p,$$

for some constant $C_2 > 0$. By linearity, observe that $D^r K_W(g) = K_W(D^r g)$. Then using (6.1), (6.2) and the triangle inequality,

$$\|D^r(K_W(g) - g)|_{I_k}\|_p \leq \|D^r(g - p_k)|_{I_k}\|_p + \|K_W(D^r g - D^r p_k)|_{I_k}\|_p \lesssim \|D^r(g - p_k)|_{I_k}\|_p \leq C \sum_{i=1}^d 2^{-(\alpha_i - r_i)W_i}\|D^{(\alpha_i - r_i)e_i}D^r g|_{I_k}\|_p,$$

for some constant $C > 0$ depending on the wavelets used. The result follows by summing both sides over $0 \leq k_i \leq 2^{W_i} - 1, l = 1, \ldots, d$ and applying Proposition 4.3.8 of [9], in view of (6.2).

\[\square\]

**Proof of Proposition 4.1.** The result is a consequence of Lemma 7.4, and it shows that the posterior for $\sigma$ contracts to $\sigma_0$ at rate $(n/\log n)^{-\alpha^*/(2\alpha^*+d)}$.

For Theorem 4.2, we will only prove the mixed partial derivatives case of (4.3), as (4.2) is a special case by setting $r = 0$ and interpreting $D^\theta f \equiv f$.

**Proof of Theorem 4.2.** As a consequence of CDV wavelets being compactly supported and their derivatives are uniformly bounded, it follows that for $r = (r_1, \ldots, r_d)^T$ with $0 \leq r_i < n_i + 1, l = 1, \ldots, d$,

$$\left\|\sum_{m_1 \in \mathbb{Z}} \cdots \sum_{m_d \in \mathbb{Z}} D^r \varphi_{N,m_1} \right\|_\infty = O(1), \quad \left\|\sum_{k \in \mathbb{Z}^d} D^r \psi_{j,k} \right\|_\infty = O\left(\prod_{l=1}^d 2^{(1/2 + r_l)j_l}\right). \quad (6.3)$$

If $f_0 \in B_{\infty,\infty}^\alpha(R)$, then by using the wavelet characterization of (3.5), it follows that for $\alpha_l < n_l + 1, l = 1, \ldots, d$,

$$\|\vartheta_0\|_\infty \leq R, \quad \|\vartheta^0_0\|_\infty \leq R2^{-\sum_{l=1}^d \alpha_l j_l} \left(\frac{1}{2^{\frac{1}{1/2 + r_l}}\gamma}\right), \quad (6.4)$$

for any $j$. Note that (3.1) implicitly implies that $\vartheta_{j,k} = 0$ when $j_l > J_{n,l}$ for some $l = 1, \ldots, d$. Denote $P = \{(j,k) : \vartheta_{j,k} \neq 0\}$ the set of nonzero wavelet coefficients. In view of (6.4) above, we define for some constant $\gamma > 0$,

$$J_n(\gamma) = \left\{(j,k) : |\vartheta^0_{j,k}| > \prod_{l=1}^d \min \left\{2^{-\alpha_l j_l} \left(\frac{1}{2^{\frac{1}{1/2 + r_l}}}\gamma\right), \gamma \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha_l - 1}}\right\}\right\}, \quad (6.5)$$

\[12\]
and for some constants $0 < \gamma < \overline{\gamma} < \infty$, the events

$$
\mathcal{A} := \left\{ \sup_{(j,k) \in J_n(\gamma)} |\theta_{j,k} - \theta^0_{j,k}| \leq \overline{\gamma} \sqrt{\frac{\log n}{n}} \right\},
$$

$$
\mathcal{B} := \left\{ P \cap J_n(\gamma)^c = \emptyset \right\} = \bigcap_{(j,k) \in J_n(\gamma)^c} \left\{ \theta_{j,k} = 0 \right\},
$$

$$
\mathcal{C} := \left\{ P^c \cap J_n(\overline{\gamma}) = \emptyset \right\} = \bigcap_{(j,k) \in J_n(\overline{\gamma})} \left\{ \theta_{j,k} \neq 0 \right\}.
$$

(6.6)

As discussed in the two criteria in Section 4, getting the correct sup-norm rate involves showing that $\mathcal{A}$ occurs with posterior probability tending to 1, and we do not make any signal detection errors as represented by events $\mathcal{B}$ and $\mathcal{C}$.

Let $\mathcal{U}_n$ be a shrinking neighborhood of $\sigma_0$ such that $E_0 \Pi(\sigma \in \mathcal{U}_n| Y) \to 1$. Observe that for $\epsilon_{n,r} := (n/\log n)^{-\alpha^*(1-\Sigma_{l=1}^d (\rho_l/\alpha_l))/(2\alpha^* + d)}$ and some large enough constant $M > 0$ to be specified below, $E_0 \Pi(||D^r f - D^r f_0|| \rightarrow_M > M \epsilon_{n,r}| Y)$ is bounded above by

$$
E_0 \sup_{\sigma \in \mathcal{U}_n} \Pi(||D^r f - D^r f_0|| \rightarrow_M > M \epsilon_{n,r}| Y, \sigma) + E_0 \Pi(\sigma \notin \mathcal{U}_n| Y)
$$

$$
\leq E_0 \sup_{\sigma \in \mathcal{U}_n} \Pi(||D^r f - D^r f_0|| \rightarrow_M > M \epsilon_{n,r}) \cap A \cap B| Y, \sigma) + E_0 \Pi(\sigma \notin \mathcal{U}_n| Y)
$$

$$
+ E_0 \sup_{\sigma \in \mathcal{U}_n} \Pi(B^c| Y, \sigma) + E_0 \sup_{\sigma \in \mathcal{U}_n} \Pi(C^c| Y, \sigma) + E_0 \sup_{\sigma \in \mathcal{U}_n} \Pi(\mathcal{A}^c \cap C| Y, \sigma).
$$

(6.7)

By Proposition 4.1, the second term tends to 0. By Lemmas 6.1 and 6.2 in Section 6.2 below, the last three terms tend to 0. We then proceed by showing that the first term on the right hand side of (6.7) approaches 0 as $n \to \infty$. We have for any $x \in [0, 1]^d$,

$$
|D^r f(x) - D^r f_0(x)| \leq \sum_{m_0 = 0}^{2^{N_l - 1}} \sum_{m_d = 0}^{2^{N_d - 1}} |\vartheta_m - \vartheta^0_m||D^r \varphi_{N,m}(x)|
$$

$$
+ \sum_{j_1 = N_1}^{2^{N_1 - 1}} \sum_{k_1 = 0}^{2^{N_d}} \cdots \sum_{j_d = N_d}^{2^{N_d - 1}} \sum_{k_d = 0}^{2^{N_d - 1}} |\theta_{j,k} - \theta^0_{j,k}||D^r \psi_{j,k}(x)|.
$$

(6.8)

Writing $\vartheta = \{\vartheta_m : 0 \leq m_l \leq 2^{N_l} - 1, 1 \leq l \leq d\}$ and using the fact that $||x||_\infty \leq ||x||$ for any $x \in \mathbb{R}^n$, the first sum above is bounded by

$$
||\vartheta - \vartheta^0||_\infty \left| \sum_{m_0 = 0}^{2^{N_l - 1}} \sum_{m_d = 0}^{2^{N_d - 1}} |D^r \varphi_{N,m}| \right|_\infty \lesssim ||\vartheta - \vartheta^0||_\infty \lesssim \epsilon_{n,r},
$$

(6.9)

where the last inequality follows from (6.3) and Corollary 7.5.

To bound the second sum, let $J_{n,l}(\alpha)$ be such that $2^{J_{n,l}(\alpha)} = (n/\log n)^\alpha^*/(\alpha_l(2\alpha^* + d))$. By (6.4) with $j = (J_{n,1}(\alpha), \ldots, J_{n,d}(\alpha))^T$, we have $|\theta^0_j| \leq |\theta^0_j|_\infty \leq C(\log n/n)^{1/2}$ for some
constant $C > 0$. Therefore, if we choose $\gamma$ small enough, we will have $|\theta_{j,k}^0| > \gamma (\log n/n)^{1/2}$ for $j_l \leq J_{n,l}(\alpha) - 1, l = 1, \ldots, d$. In other words,

$$J_n(\gamma) \subset I_n(\alpha) := \{(j,k) : N_l \leq j_l < J_{n,l}(\alpha), 0 \leq k_l < 2^h, l = 1, \ldots, d\}, \quad (6.10)$$

for sufficiently small $\gamma$. Let the sum $\sum_{(j,k)}$ be an abbreviation of $\sum_{j_1} \sum_{k_1} \cdots \sum_{j_d} \sum_{k_d}$. Using $I_n(\alpha)$ and its complement, the second sum on the right hand side of (6.8) is

$$\sum_{(j,k) \in I_n(\alpha)} \left| \theta_{j,k} - \theta_{j,k}^0 \right| |D^n \psi_{j,k}(x)| + \sum_{(j,k) \in I_n(\alpha)^c} \left| \theta_{j,k} - \theta_{j,k}^0 \right| |D^n \psi_{j,k}(x)|. \quad (6.11)$$

We first bound the second term with summation index in $I_n(\alpha)^c$, which can be further decomposed as

$$\sum_{(j,k) \in I_n(\alpha)^c} \left| \theta_{j,k}^0 \right| |D^n \psi_{j,k}(x)| + \sum_{(j,k) \in I_n(\alpha)^c \cap P} \left| \theta_{j,k} - \theta_{j,k}^0 \right| |D^n \psi_{j,k}(x)|.$$ 

Taking complements on both sides of (6.10), we have $I_n(\alpha)^c \subset J_n(\gamma)^c$. Thus, the second sum above is bounded by $\sum_{J_n(\gamma)^c \cap P} \left| \theta_{j,k} - \theta_{j,k}^0 \right| |D^n \psi_{j,k}(x)|$. This is zero with posterior probability tending to 1 under event $B$. In view of (6.5), the first sum with summation indices in $(j,k) \in I_n(\alpha) \cap P^c \subset J_n(\gamma)^c$ is bounded above by

$$\max\{R, \gamma\} \sum_{(j,k) \in I_n(\alpha)^c} \prod_{l=1}^d \min \left\{ 2^{-\alpha_l j_l (\frac{1}{2} + \frac{1}{2\alpha_l})}, \left( \frac{\log n}{n} \right)^{1/(2d)} \right\} |D^n \psi_{j,k}(x)|. \quad (6.12)$$

If we define sets $Q_l, l = 1, \ldots, d$, where $Q_l$ can be $\{N_l \leq j_l \leq J_{n,l}(\alpha) - 1\}$ or $\{j_l \geq J_{n,l}(\alpha)\}$, but with the constraint that not all $Q_l$’s are $\{N_l \leq j_l \leq J_{n,l}(\alpha) - 1\}$. Then the summation in (6.12) is over $(j,k)$ such that $j$ takes on all $2^d - 1$ possible combinations of the $Q_l$’s, and each combination has the form

$$\sum_{j_1 \in Q_1} \sum_{k_1=0}^{2^h-1} \cdots \sum_{j_d \in Q_d} \sum_{k_d=0}^{2^h-1} \prod_{l=1}^d \min \left\{ 2^{-\alpha_l j_l (\frac{1}{2} + \frac{1}{2\alpha_l})}, \left( \frac{\log n}{n} \right)^{1/(2d)} \right\} |D^n \psi_{j,k}(x)|$$

$$\leq \prod_{l=1}^d \sum_{j_l \in Q_l} \min \left\{ 2^{-\alpha_l j_l (\frac{1}{2} + \frac{1}{2\alpha_l})}, \left( \frac{\log n}{n} \right)^{1/(2d)} \right\} \sup_{x \in [0,1]^d} \sum_{k_1=0}^{2^h-1} \cdots \sum_{k_d=0}^{2^h-1} |D^n \psi_{j,k}(x)|$$

$$\leq \prod_{l=1}^d \sum_{j_l \in Q_l} 2^{(r_l+1/2)j_l} \min \left\{ 2^{-\alpha_l j_l (\frac{1}{2} + \frac{1}{2\alpha_l})}, \left( \frac{\log n}{n} \right)^{1/(2d)} \right\} \quad (6.13)$$

where the last inequality follows from (6.3). The two expressions inside the minimum function will have the same order if $j_l = J_{n,l}(\alpha)$ and $2^{-\alpha_l j_l (\frac{1}{2} + \frac{1}{2\alpha_l})}$ will have a larger order when $j_l < J_{n,l}(\alpha)$, while $(\log n/n)^{1/(2d)}$ will dominate if $j_l \geq J_{n,l}(\alpha)$. Therefore under the regime
\( Q_l = \{ N_l \leq j_l \leq J_{n,l}(\alpha) - 1 \} \), we have that
\[
\sum_{N_l \leq j_l < J_{n,l}(\alpha)} 2^{(r_l+1/2)j_l} \min \left\{ 2^{-\alpha_l j_l \left( \frac{1}{d} + \frac{1}{2n} \right)}, \left( \frac{\log n}{n} \right)^{1/(2d)} \right\} \lesssim 2^{(r_l+1/2)J_{n,l}(\alpha)} \left( \frac{\log n}{n} \right)^{1/(2d)}
\]
\[
\lesssim \left( n/\log n \right)^{-\frac{\alpha^*}{2n^2 + d}} \left( \frac{1}{d} + \frac{1}{2n} - \frac{r_l}{4} \right);
\]
while under the regime \( Q_l = \{ j_l \geq J_{n,l}(\alpha) \} \), we will have
\[
\sum_{j_l \geq J_{n,l}(\alpha)} 2^{(r_l+1/2)j_l} \min \left\{ 2^{-\alpha_l j_l \left( \frac{1}{d} + \frac{1}{2n} \right)}, \left( \frac{\log n}{n} \right)^{1/(2d)} \right\} \lesssim 2^{-\alpha_l \left( \frac{1}{d} + \frac{1}{2n} \right) - r_l - 1/2} J_{n,l}(\alpha)
\]
\[
\lesssim \left( n/\log n \right)^{-\frac{\alpha^*}{2n^2 + d}} \left( \frac{1}{d} + \frac{1}{2n} - \frac{r_l}{4} \right),
\]
where the first inequality above is justified since \( \alpha_l \left( \frac{1}{d} + \frac{1}{2n} \right) - r_l - 1/2 > 0 \) for \( l = 1, \ldots, d \) on \( \alpha \in \mathbb{A}_r \). Putting this bound back to (6.13) and using the fact that there are only \( 2^d - 1 \) combinations of the \( Q_l \)'s, it then follows that the right hand side of (6.12) is
\[
O \left( \left( n/\log n \right)^{-\frac{\alpha^*}{2n^2 + d} \sum_{\gamma=1}^d \left( \frac{1}{d} + \frac{1}{2n} - \frac{r_l}{4} \right)} \right) = O(\epsilon_{n,r}).
\]

Using a similar decomposition as before, the first sum with summation indices in \( I_n(\alpha) = \{ (j,k) : N_l \leq j_l \leq J_{n,l}(\alpha) - 1, 0 \leq k_l \leq 2^h - 1 \} \) of (6.11) can be decomposed into
\[
\sum_{(j,k) \in I_n(\alpha) \cap J_n(\gamma)} |\theta_{j,k} - \theta_{j,k}^0| |D^r \psi_{j,k}(x)| + \sum_{(j,k) \in I_n(\alpha) \cap J_n(\gamma)^c \cap \mathcal{P}} |\theta_{j,k} - \theta_{j,k}^0| |D^r \psi_{j,k}(x)|
\]
\[
+ \sum_{(j,k) \in I_n(\alpha) \cap J_n(\gamma)^c \cap \mathcal{P}^c} |\theta_{j,k} - \theta_{j,k}^0| |D^r \psi_{j,k}(x)|.
\]

Recalling \( J_n(\gamma) \subset I_n(\alpha) \) for the first sum above, and the second sum in the decomposition vanishes under event \( \mathcal{B} \), the above is
\[
\sum_{(j,k) \in J_n(\gamma)} |\theta_{j,k} - \theta_{j,k}^0| |D^r \psi_{j,k}(x)| + \sum_{(j,k) \in I_n(\alpha) \cap J_n(\gamma)^c \cap \mathcal{P}^c} |\theta_{j,k}^0| |D^r \psi_{j,k}(x)|.
\]
By intersecting with event \( \mathcal{A} \), the right hand side above is further bounded by
\[
\max_{(j,k) \in J_n(\gamma)} |\theta_{j,k} - \theta_{j,k}^0| \sum_{j_1=N_1}^{J_{n,1}(\alpha)-1} \cdots \sum_{j_d=N_d}^{J_{n,d}(\alpha)-1} \sum_{k_1=0}^{2^d-1} \cdots \sum_{k_d=0}^{2^d-1} |D^r \psi_{j,k}(x)|
\]
\[
+ \sum_{(j,k) \in I_n(\alpha) \cap J_n(\gamma)^c} \gamma \sqrt{\frac{\log n}{n}} |D^r \psi_{j,k}(x)|
\]
\[
\lesssim \prod_{l=1}^d \sum_{N_l \leq j_l < J_{n,l}(\alpha)} 2^{(1/2+r_l)j_l} \sqrt{\frac{\log n}{n}} \lesssim \epsilon_{n,r},
\] (6.14)

15
Lemmas 6.1 and 6.2, it follows that the right hand side of (6.7) approaches 0 as
sufficiently large constant $M > 0$ involving only individual coefficient
$	heta$. Bounding of posteriors

This is detailed in Section 6.2 below.

Now, combining the bounds established in (6.9), (6.12) and (6.14) into (6.8), we conclude that $\|D^r f - D^r f_0\|_1 \leq ME_{n,r}$ for some sufficiently large constant $M > 0$ under the posterior distribution. Using this fact with
Lemmas 6.1 and 6.2, it follows that the right hand side of (6.7) approaches 0 as $n \to \infty$.

It now remains to show that the last three terms in (6.7) approach 0 asymptotically, and
this is detailed in Section 6.2 below.

6.2 Bounding of posteriors

In view of (6.6) above, it is clear that we need to bound posterior probabilities of events
involving only individual coefficient $\theta_{j,k}$. To accomplish this, we bound posterior of $\theta_{j,k}$
under the regression model by posterior of $\theta_{j,k}$ arising from some quasi-white noise model,
where the latter model greatly simplifies calculations through its component-wise structure.

We first define notations. If the rows and columns of a matrix are each indexed by
d-dimensional multi-indices, we assume that these multi-indices are arranged in the
lexicographic order. Let $\mathbf{i} = (i_1, \ldots, i_d)^T$ and $\mathbf{j} = (j_1, \ldots, j_d)^T$. For a matrix $A$
indexed by 2d-dimensional indices, we write $a_{\mathbf{i},\mathbf{j}}$ or $A_{\mathbf{i},\mathbf{j}}$ to be the $(\mathbf{i}, \mathbf{j})$th entry, $A_{\mathbf{j}}$,
to be the $\mathbf{j}$th row of $A$, $A_{\mathbf{j},-\mathbf{j}}$ to be the $\mathbf{j}$th row of $A$ such that the $\mathbf{j}$th entry of that row is excluded,
and $A_{-\mathbf{j},-\mathbf{j}}$ to be a matrix created as a result of deleting the $\mathbf{j}$th row and $\mathbf{j}$th column of $A$.
For a vector $x$, we write $x_{\mathbf{j}}$ to be its $\mathbf{j}$th component, and $x_{-\mathbf{j}}$ to the $x$ such that its $\mathbf{j}$th
component is excluded.

Given observations $\{X_1, \ldots, X_n\}$, we construct the father wavelet matrix $B$
such that its $(h,m)$th entry is $\varphi_{N,m}(X_h)$, for $1 \leq h \leq n$ and $0 \leq m \leq 2^N - 1, l = 1, \ldots, d$. In
addition, we define the mother wavelet matrix $\Psi_j$ such that its $(h,k)$th entry is $\psi_{j,k}(X_h)$,
for $1 \leq h \leq n$ and $0 \leq k \leq 2^k - 1, l = 1, \ldots, d$.

Observe that for $d \times 1$ vectors $a, b, c, e, \Psi^T B \Psi b$ is a matrix indexed by 2d-dimensional
indices, such that $(\Psi^T B \Psi b)_{\mathbf{i},\mathbf{e}} = \sum_{\mathbf{j}} \Psi_{\mathbf{i},\mathbf{e}}(X) \Psi_{\mathbf{b},\mathbf{e}}(X)$. Similarly, we also have $(\Psi^T B \Psi b)_{\mathbf{a},\mathbf{m}} = \Psi_{\mathbf{a},\mathbf{m}}(X) \varphi_{N,m}(X)$. Recall that $\vartheta = \{\vartheta_m : 0 \leq m \leq 2^N - 1, l = 1, \ldots, d\}$ and define
$\mathbf{\theta}_{\mathbf{j}} = \{\theta_{j,k} : 0 \leq k \leq 2^k - 1, l = 1, \ldots, d\}$ for fixed $\mathbf{j}$. Define $\mathbf{\theta} = \{\mathbf{\theta} : \mathbf{\theta}_{j,k}$ is excluded\}$,
where $\mathbf{\theta} = \{\theta_{j,k} : N_l \leq j_i \leq J_{n,l} - 1, l = 1, \ldots, d\}$. We write $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d)^T$ and
the truncation error $\xi = F_0 - \sum_{j=1}^{d} \Psi_{j} \mathbf{\theta}^0_j$, where $F_0 = (f_0(X_1), \ldots, f_0(X_n))^T$.

Write (1.1) as $Y = B \theta_0 + \sum_{j=1}^{d} \sum_{j=1}^{J_{n,d}} \Psi_{j} \mathbf{\theta}^0_j + \xi + \varepsilon$ under the true distribution
$P_0$. Then $\|Y - B \theta - \sum_{j=1}^{d} \sum_{j=1}^{J_{n,d}} \Psi_{j} \mathbf{\theta}^0_j\|^2$ is

$$
(\theta_{j,k} - \theta_{0,j,k}^0)^2 (\Psi^T_j \Psi_j)_{k,k} + 2(\theta_{j,k} - \theta_{0,j,k}^0) \beta_n(\tilde{\theta}) + H_n(\tilde{\theta}) + \|\xi + \varepsilon\|^2,
$$

where we have separated the $(j,k)$th component out from the rest such that

$$
\beta_n(\tilde{\theta}) := (\Psi^T_j \Psi_j)_{k,-k}(\theta_j - \theta_{0,j}^0)_{-k} + \sum_{\alpha \neq j} (\Psi^T_j \Psi_{\alpha})_{k,-k} (\theta_{\alpha} - \theta_{0,\alpha}^0)
$$

$$
+ (\Psi^T_j B)_{k,-k}(\theta - \theta_0) - (\xi^T \Psi_j)_{k} - (\varepsilon^T \Psi_j)_{k}.
$$
for $\tilde{\Theta} := (\vartheta, \theta_{-(j,k)})$ and $H_n(\tilde{\Theta})$ is

$$
\begin{align*}
(\theta_j - \theta_j^0)^T_k (\Psi_j^T \Psi_j)_k - k, k (\theta_j - \theta_j^0)_k + 2(\vartheta - \tilde{\Theta}_0)^T (B^T \Psi_j)_k (\theta_j - \theta_j^0)_k \\
+ 2 \sum_{a \neq j} (\theta_j - \theta_j^0)^T_k (\Psi_j^T \Psi_a)_k (\theta_a - \theta_a^0) + (\vartheta - \tilde{\Theta}_0)^T B^T B (\vartheta - \theta_0) \\
+ \sum_{a \neq j} \sum_{b \neq j} (\theta_a - \theta_a^0)^T \Psi_a^T \psi_b (\theta_b - \theta_b^0) - 2 \xi^T B (\vartheta - \vartheta_0) \\
+ 2 \sum_{a \neq j} (\vartheta - \tilde{\Theta}_0)^T B^T \Psi_a (\theta_a - \theta_a^0) - 2 \varepsilon^T B (\vartheta - \theta_0) \\
- 2(\xi^T \Psi_j)_k (\theta_j - \theta_j^0)_k - 2 \sum_{a \neq j} \xi^T \Psi_a (\theta_a - \theta_a^0) \\
- 2(\varepsilon^T \Psi_j)_k (\theta_j - \theta_j^0)_k - 2 \sum_{a \neq j} \varepsilon^T \Psi_a (\theta_a - \theta_a^0).
\end{align*}
$$

The prior density of $\tilde{\Theta}$ is $d\Pi(\tilde{\Theta}) = d\Pi(\vartheta)d\Pi(\theta_{-(j,k)})$, where $d\Pi(\vartheta) = \prod_m p(\vartheta_m)$ and

$$
d\Pi(\theta_{-(j,k)}) = \prod_{(x,y) \neq (j,k)} [(1 - \omega_{x,y})d\delta_0(\theta_{x,y}) + \omega_{x,y} p(\theta_{x,y})d\theta_{x,y}].
$$

Define $K_n(\tilde{\Theta}) := \exp \left\{ \frac{\beta_n(\tilde{\Theta})^2}{2\sigma^2(\Psi_j^T \Psi_j)_k} - H_n(\tilde{\Theta})/(2\sigma^2) \right\}$. Let $\mathcal{A}$ and $\mathcal{W}$ be two measurable sets on the parameter space of $(\vartheta, \theta)$, and $\Omega$ be an event on $\varepsilon$ or equivalently on $Y$. Then in view of (6.15) and by completing the squares,

$$
\begin{align*}
\Pi(\theta_{j,k} \in \mathcal{A}, (\vartheta, \theta_{-(j,k)}) \in \mathcal{W}|Y, \sigma) &= \frac{\int_{\mathcal{W}} \int_{\mathcal{A}} e^{-\sigma^{-1}(X - B\vartheta - \sum \Psi_j \theta_j)^T (X - B\vartheta - \sum \Psi_j \theta_j)} d\Pi(\vartheta, \theta)\int_{-\infty}^{\infty} e^{-\sigma^{-1}(x - B\vartheta - \sum \Psi_j \theta_j)^T (x - B\vartheta - \sum \Psi_j \theta_j)} d\Pi(\vartheta, \theta)}{\int_{\mathcal{W}} \int_{\mathcal{A}} e^{-\sigma^{-1}(X - B\vartheta - \sum \Psi_j \theta_j)^T (X - B\vartheta - \sum \Psi_j \theta_j)} d\Pi(\vartheta, \theta) d\Pi(\vartheta, \theta)}
\leq \frac{\int_{\mathcal{W}} \int_{\mathcal{A}} \exp \left\{ -\frac{1}{2\sigma^2} \left( \Psi_j^T \Psi_j \right)_k \left[ \theta_{j,k} - \theta_{j,k}^0 + \frac{\beta_n(\tilde{\Theta})}{\left( \Psi_j^T \Psi_j \right)_k} \right]^2 \right\} d\pi(\theta_{j,k}) K_n(\tilde{\Theta}) d\Pi(\tilde{\Theta})}{\int_{\mathcal{W}} \int_{\mathcal{A}} \exp \left\{ -\frac{1}{2\sigma^2} \left( \Psi_j^T \Psi_j \right)_k \left[ \theta_{j,k} - \theta_{j,k}^0 + \frac{\beta_n(\tilde{\Theta})}{\left( \Psi_j^T \Psi_j \right)_k} \right]^2 \right\} d\pi(\theta_{j,k}) K_n(\tilde{\Theta}) d\Pi(\tilde{\Theta})}
=: \frac{I_{\mathcal{W}}}{I_{\mathcal{W}}} n(A, \tilde{\Theta}) K_n(\tilde{\Theta}) d\Pi(\tilde{\Theta}) d\Pi(\tilde{\Theta})
\end{align*}
$$

Therefore, if $\mathcal{W}$ and $\Omega$ are both chosen such that $|\beta_n(\tilde{\Theta})|$ has sharp upper bound of the correct order uniformly over $\tilde{\Theta}$, then we can untangle the exponential factor with $K_n(\tilde{\Theta})$, and the ratio will look like a posterior from a sequence white noise model, with $\int_{\mathcal{W}} K_n(\tilde{\Theta}) d\Pi(\tilde{\Theta})$ cancelling out each other. Hence, we reduce our posterior to one where we only have to compare the $k$th component at the top and bottom, just like in the case when we have quasi-white noise model of the form $Y_{j,k} = \theta_{j,k} + \sigma (\Psi_j^T \Psi_j)_k^{-1/2} \varepsilon_{j,k}$.

The optimal choices of $\mathcal{W}$ and $\Omega$ depend on the statistical problem at hand, and also implicitly depend on the choice of the basis used through the entries of $\Psi_j^T \Psi_j$. For orthonormal
basis such as the wavelets used in this paper, the diagonal entries are typically of the order $n$ under some conditions on the truncation level $J_{n,l}, l = 1, \ldots, d$ (see Lemma 7.2 below). Let us denote $\beta_n(\Theta) = \beta_n(\tilde{\Theta}) + (e^T \Psi_j)_k$. As we will show below, the appropriate $W$ for our wavelet regression model is

$$W_n = \{ \Theta : |\beta_n(\tilde{\Theta})| \leq \tau_n \sqrt{n \log n} \}, \quad (6.17)$$

with $\tau_n \to 0$ given in (7.8) of Lemma 7.6; while $\Omega$ has the form

$$\Omega_n(c) := \bigcap_{N_l \leq J_l \leq J_{n,l}} \left\{ \left| \frac{\beta_n(\tilde{\Theta}) - \beta_n(\Theta)}{\sqrt{\sigma_0^2(\Psi_j^T \Psi_j)_{k,k}}} \right| \leq \left( 2 \log \prod_{l=1}^d 2^k + c \log n \right)^{1/2} \right\}. \quad (6.18)$$

It turns out that for $W_n$ to hold true with high probability under the posterior, we need $\alpha^* > d/2$ (see Lemma 7.6 below), and this is ensured by the lower bound imposed on $\alpha_l, l = 1, \ldots, d$ in (4.1). Using the technique discussed, we then show that the last three terms in (6.7) are negligible under the posterior.

**Lemma 6.1.** For small enough $\gamma$ and large enough $\tau$, there exist constants $P_1, P_2 > 0$ such that uniformly in $f_0 \in \mathcal{B}_\infty^\alpha(R)$ with $\alpha \in \mathbb{A}$ and any $0 < R \leq R_0 - 1/2$,

$$E_0 \sup_{\sigma \in U_n} \Pi(\mathcal{B}^c|Y, \sigma) = E_0 \sup_{\sigma \in U_n} \Pi(\mathcal{P} \cap \mathcal{J}_n(\gamma)^c \neq \emptyset|Y, \sigma) \leq \frac{(\log n)^d}{n^{P_1}}, \quad (6.19)$$

$$E_0 \sup_{\sigma \in U_n} \Pi(\mathcal{C}^c|Y, \sigma) = E_0 \sup_{\sigma \in U_n} \Pi(\mathcal{P}^c \cap \mathcal{J}_n(\gamma) \neq \emptyset|Y, \sigma) \leq \frac{(\log n)^d}{n^{P_2}}. \quad (6.20)$$

**Proof of Lemma 6.1.** We first prove (6.19). By (6.6), we can write $[\mathcal{P} \cap \mathcal{J}_n(\gamma)^c \neq \emptyset] = \cup_{(j,k) \in \mathcal{J}_n(\gamma)^c} \{ \theta_{j,k} \neq 0 \}$. Recall that $U_n$ is a shrinking neighborhood of $\sigma_0$, and $\sigma \in U_n$ implies that $\sigma^2 = \sigma_0^2 + o(1)$. Using the fact that the posterior probability is bounded by 1, we have $E_0 \sup_{\sigma \in U_n} \Pi(\mathcal{P} \cap \mathcal{J}_n(\gamma)^c \neq \emptyset|Y, \sigma)$ is bounded above by

$$E_0 \sup_{\sigma \in U_n} \sum_{(j,k) \in \mathcal{J}_n(\gamma)^c} \Pi(\theta_{j,k} \neq 0|Y, \sigma) I_{\Omega_n(\gamma)} + P_0[I_{\Omega_n(\gamma)^c}]. \quad (6.21)$$

To bound the last term, observe that $\tilde{\beta}_n(\Theta) = \beta_n(\tilde{\Theta}) = (e^T \Psi_j)_k$ is Gaussian under $P_0$, with mean 0 and variance $\sigma_0^2(\Psi_j^T \Psi_j)_{k,k}$. Thus, using the fact that $P(\sum_i \varepsilon_i a_i > z) \leq 2e^{-z^2/(2\sigma^2||a||^2)}$ for $z > 0$ and constants $a = (a_1, \ldots, a_n)^T$, if $\varepsilon_i, i = 1, \ldots, n$ are Gaussian with mean 0 and variance $\sigma^2$, we have $P_0[I_{\Omega_n(\gamma)^c}]$ is bounded above by

$$\sum_{J_{n,1} \leq n_{j_1} \leq J_{n,1}} \cdots \sum_{J_{n,d} \leq n_{j_d} \leq J_{n,d}} \sum_{k_1=0}^{J_{n,d}-1} \sum_{k_d=0}^{J_{n,d}-1} P_0 \left( |(e^T \Psi_j)_k| > (2 \log \prod_{l=1}^d 2^l + \gamma \log n)^{1/2} \sigma_0(\Psi_j^T \Psi_j)_{k,k}^{1/2} \right)$$

$$\leq 2(2^d - 1)n^{-2^d} \prod_{l=1}^d (J_{n,l} - N_l) \lesssim n^{-2^d}(\log n)^d. \quad (6.22)$$

18
The right hand side above approaches 0 as \( n \to \infty \) for any \( \gamma > 0 \). Recall that in (6.16) above, we defined

\[
I_n(A, \tilde{\Theta}) := \int_A \exp \left\{ -\frac{(\Psi_j^T \Psi_j)_{k,k}}{2\sigma^2} \left[ \theta_{j,k} - \theta_{j,k}^0 + \frac{\beta_n(\tilde{\Theta})}{(\Psi_j^T \Psi_j)_{k,k}} \right] \right\} d\pi(\theta_{j,k})
\]

Therefore, to bound the first term, observe that for \((j, k) \in J_n(\gamma)^c\),

\[
\Pi(\theta_{j,k} \neq 0|Y, \sigma) I_{\Omega_n(\gamma)} \leq \int_{W_n} I_n(\theta_{j,k} \neq 0, \tilde{\Theta}) K_n(\tilde{\Theta}) d\Pi(\tilde{\Theta}) \leq \int_{W_n} I_n(\theta_{j,k} = 0, \tilde{\Theta}) K_n(\tilde{\Theta}) d\Pi(\tilde{\Theta}) + O_{\rho_0}(n^{-B}), \tag{6.23}
\]

where \( O_{\rho_0}(n^{-B}) \) for some constant \( B > 0 \) follows from Lemma 7.6. When \( \theta_{j,k} \neq 0 \), then \( d\pi(\theta_{j,k}) = \omega_{j,n} p(\theta_{j,k}) d\theta_{j,k} \). Since \( p_{\text{max}} = \sup_{x \in \mathbb{R}} p(x) < \infty \) by assumption, we can upper bound \( I_n(\theta_{j,k} \neq 0, \tilde{\Theta}) \) by

\[
\omega_{j,n} p_{\text{max}} \int_{-\infty}^{\infty} \exp \left\{ -\frac{(\Psi_j^T \Psi_j)_{k,k}}{2\sigma^2} \left[ \frac{\beta_n(\tilde{\Theta})}{(\Psi_j^T \Psi_j)_{k,k}} \right]^2 \right\} dx = \omega_{j,n} p_{\text{max}} \frac{\sqrt{2\pi\sigma^2}}{(\Psi_j^T \Psi_j)_{k,k}}.
\]

Therefore, in view of the lower bound in Lemma 7.2,

\[
\int_{W_n} I_n(\theta_{j,k} \neq 0, \tilde{\Theta}) K_n(\tilde{\Theta}) d\Pi(\tilde{\Theta}) \leq \omega_{j,n} n^{-1/2} \int_{W_n} K_n(\tilde{\Theta}) d\Pi(\tilde{\Theta}). \tag{6.24}
\]

Now, \( \int_{W_n} I_n(\theta_{j,k} = 0, \tilde{\Theta}) K_n(\tilde{\Theta}) d\Pi(\tilde{\Theta}) \) is

\[
(1 - \omega_{j,n}) \int_{W_n} \exp \left\{ -\frac{(\Psi_j^T \Psi_j)_{k,k}}{2\sigma^2} \left[ \theta_{j,k}^0 - \frac{\beta_n(\tilde{\Theta})}{(\Psi_j^T \Psi_j)_{k,k}} \right]^2 \right\} K_n(\tilde{\Theta}) d\Pi(\tilde{\Theta}). \tag{6.25}
\]

To lower bound the expression above, we proceed by lower bounding the first exponential factor. By definition, we have \( |\theta_{j,k}^0| \leq \gamma \sqrt{\log n} / n \) for \((j, k) \in J_n(\gamma)^c\). Under \( W_n \) and \( \Omega_n(\gamma) \), we have by the triangle inequality that

\[
\sqrt{(\Psi_j^T \Psi_j)_{k,k}} |\theta_{j,k}^0 - \frac{\beta_n(\tilde{\Theta})}{(\Psi_j^T \Psi_j)_{k,k}}| \leq \sqrt{(\Psi_j^T \Psi_j)_{k,k}} |\theta_{j,k}^0| + \left| \frac{\beta_n(\tilde{\Theta})}{(\Psi_j^T \Psi_j)_{k,k}} \right| \leq \sqrt{C_\gamma} \sqrt{\log n} + \sigma_0 \sqrt{2 \log 2 \Sigma_{j=1}^d \gamma_j + \gamma \log n} \leq 2 \sqrt{C_\gamma} \sqrt{\log n} + \sigma_0 (2 \log 2 \Sigma_{j=1}^d \gamma_j + \gamma \log n)^{1/2},
\]

19
for large enough $n$ because $\tau_n \to 0$ as $n \to \infty$. Note also that by Lemma 7.2, $\sqrt{C_1n} \leq \sqrt{C_2n}$ for some constants $C_1, C_2 > 0$, because $2\Sigma_{l=1}^d j_l = \sqrt{n/\log n} = o(n)$ by assumption. By squaring both sides, we have

$$
(\Psi_j \Psi_j)_{k,k} \left[ \theta_{j,0} - \frac{\beta_{n}(\Theta)}{(\Psi_j \Psi_j)_{k,k}} \right]^2 \leq 4C_2\gamma^2 \log n + \sigma_0^2(2 \log 2 \sum_{l=1}^d j_l + \gamma \log n) + 4\sigma_0 \sqrt{C_2\gamma} \sqrt{\log n (2 \log 2 \sum_{l=1}^d j_l + \gamma \log n)}^{1/2} \leq \kappa(\gamma) \log n + 2\sigma_0^2 \log 2 \sum_{l=1}^d j_l,
$$

where $\kappa(\gamma) = 4C_2\gamma^2 + 4\sigma_0 \sqrt{C_2\gamma}^{3/2} + (4\sigma_0 \sqrt{2C_2} + \sigma_0^2)\gamma$, and the last inequality follows from using $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$. Thus, $\int_{\mathcal{W}_n} I_n([\theta_{j,k} = 0], \Theta) K_n(\Theta) d\Pi(\Theta)$ is bounded below by

$$
(1 - \omega_{j,n}) \exp \left( -\kappa(\gamma) \log n 2\sigma^2 - \frac{\sigma_0^2}{\sigma^2} \log 2 \sum_{l=1}^d j_l \right) \int_{\mathcal{W}_n} K_n(\Theta) d\Pi(\Theta).
$$

By assumption, we have $\omega_{j,n} \leq \min\{2^{-\sum_{l=1}^d j_l(1+\mu_l)} 1/2\}$ for $\mu_{min} > 1/2$. Using the fact that $x/(1 - x) \leq 2x$ for $0 \leq x \leq 0.5$, the upper and lower bounds in (6.24) and (6.26), the first term on the right hand side of (6.21) can be bounded above up to some constant by

$$
n_{2^{\kappa(\gamma)-1/2} \sigma} \sum_{j_1 = N_1}^{J_{1,1}} \sum_{k_1 = 0}^{2^{j_1-1}} \cdots \sum_{j_d = N_d}^{J_{d,d}} \sum_{k_d = 0}^{2^{j_d-1}} 2^{\sigma_0^2 \sum_{l=1}^d j_l} 2\omega_{j,n} \leq n_{2^{\kappa(\gamma)-1/2} \sigma} \prod_{l=1}^d \sum_{j_l = N_l}^{J_{l,1}} 2^{j_l (\sigma_0^2/\sigma^2 - \mu_l)}.
$$

Now if $\sigma_0^2/\sigma^2 > \mu_l$, we have $\sum_{j_l = N_l}^{J_{l,1}} 2^{j_l (\sigma_0^2/\sigma^2 - \mu_l)} \leq 2^{J_{l,1} (\sigma_0^2/\sigma^2 - \mu_l)}$, while for $\sigma_0^2/\sigma^2 \leq \mu_l$ for all $l = 1, \ldots, d$, the right hand side of (6.27) is $O(n^{\kappa(\gamma)/[2\sigma_0^2 + o(1)] - 1/2})$ after uniformizing over $\sigma \in \mathcal{U}_n$, and it will tend to 0 if $\gamma$ is small enough. On the other hand, if $\sigma_0^2/\sigma^2 > \mu_l$ for at least one $l = 1, \ldots, d$, then the right hand side of (6.27) is bounded above up to some constant by

$$
n_{2^{\kappa(\gamma)-1/2} \sigma} \prod_{l=1}^d 2^{J_{l,1} (\sigma_0^2/\sigma^2 - \mu_l)} \leq n_{-\mu_{min} - 1/2 + \sigma_0^2/\sigma^2 + \frac{\sigma_0^2}{\sigma^2} \kappa(\gamma)} \sigma.
$$

Uniformizing across $\sigma \in \mathcal{U}_n$, the right hand side above is $(\sigma_0 + o(1)) n^{1/2 - \mu_{min} + o(1) + \kappa(\gamma)/[2\sigma_0^2 + o(1)]}$, which will approach 0 as $n \to \infty$, since $\mu_{min} > 1/2$ by prior assumption and if $\gamma$ is chosen small enough.

We now prove the second assertion (6.20). By (6.10), $(j, k) \in \mathcal{J}_n(\gamma) \subset \mathcal{J}_n(\gamma) \subset \mathcal{I}_n(\alpha)$, then $2^{-\sigma_0 j_l [d^{-1} + (2\sigma^2)^{-1}]}$ dominates $\overline{\tau}(\log n/\log n)^{1/2d}$ inside the minimum function of (6.5) since $j_l \leq J_{n,l}(\alpha) - 1, l = 1, \ldots, d$. Therefore, the definition of event $C$ in (6.6) can be reduced to

$$
C := \bigcap_{\{\theta_{j,k} \neq 0\}} \{\theta_{j,k} \neq 0\},
$$

(6.29)
for small enough $\gamma > 0$. Taking complements and using the same decomposition as in (6.21), $E_0 \sup_{\sigma \in \Omega_n} \Pi(\mathcal{P}^c \cap \mathcal{J}_n(\gamma) \neq \emptyset | \mathbf{Y}, \sigma)$ is bounded above by

$$E_0 \sup_{\sigma \in \Omega_n} \sum_{\{ |\theta^0_{j,k}| > \bar{\gamma} \sqrt{\log n/n} \}} \Pi(\theta_{j,k} = 0 | \mathbf{Y}, \sigma) \mathbb{I}_{\Omega_n(1)} + P_0[\Omega_n(1)^c].$$

(6.30)

Using the same argument leading to (6.22) by substituting $\gamma$ by 1,

$$P_0[\Omega_n(1)^c] \lesssim n^{-1/2}(\log n)^d \to 0$$

(6.31)

as $n \to \infty$. To bound the first term, note that in the present case,

$$\Pi(\theta_{j,k} = 0 | \mathbf{Y}, \sigma) \mathbb{I}_{\Omega_n(1)} \leq \int_{\mathcal{W}_n} I_n(\{ |\theta_{j,k} = 0 \}, \bar{\Theta}) K_n(\bar{\Theta}) d\Pi(\bar{\Theta}) \int_{\mathcal{W}_n} I_n(\{ |\theta_{j,k} \neq 0 \}, \bar{\Theta}) K_n(\bar{\Theta}) d\Pi(\bar{\Theta}) \mathbb{I}_{\Omega_n(1)} + \Pi(\mathcal{W}^c \cap \mathbf{Y} | \mathbf{Y}, \sigma),$$

and the second term on the right hand side above is $O_{p_0}(n^{-B})$ for some constant $B > 0$ by Lemma 7.6. To upper bound $\int_{\mathcal{W}_n} I_n(\{ |\theta_{j,k} = 0 \}, \bar{\Theta}) K_n(\bar{\Theta}) d\Pi(\bar{\Theta})$, we need to upper bound the first exponential factor in (6.25). Now, $(\Psi_T \Psi_j)^{1/2} \geq C_1 \sqrt{n}$ by Lemma 7.2 since $2^{\sum_{l=1}^d \tau_l} = o(n)$ for $j_l < J_{n,l}(\alpha), l = 1, \ldots, d$. Applying the reverse triangular inequality twice and under $\mathcal{W}_n$ and $\Omega_n(1)$, we have for any $|\theta^0_{j,k}| > \bar{\gamma} \sqrt{\log n/n},$

$$\sqrt{(\Psi_T \Psi_j)_{k,k}} |\theta^0_{j,k} - \beta_n(\bar{\Theta})| \leq \frac{\beta_n(\bar{\Theta})}{(\Psi_T \Psi_j)_{k,k}} \left( |\beta_n(\bar{\Theta})| - |\beta_n(\bar{\Theta}) - \beta_n(\bar{\Theta})| \right) \leq \sqrt{C_1 \bar{\gamma} \sqrt{\log n} - (\tau_n / \sqrt{C_1}) \sqrt{\log n} - \sigma_0 \left( 2 \log 2^{\sum_{l=1}^d \tau_l} + \log n \right)^{1/2} > 0.5 \sqrt{C_1 \bar{\gamma} \sqrt{\log n}},$$

if $\bar{\gamma}$ is chosen large enough since $\tau_n \to 0$ as $n \to \infty$. Therefore,

$$\int_{\mathcal{W}_n} I_n(\{ |\theta_{j,k} = 0 \}, \bar{\Theta}) K_n(\bar{\Theta}) d\Pi(\bar{\Theta}) \leq (1 - \omega_{j,n}) \exp \left( \frac{C_1 \bar{\gamma}^2}{8 \sigma^2} \log n \right) \int_{\mathcal{W}_n} K_n(\bar{\Theta}) d\Pi(\bar{\Theta}).$$

By (3.2), $p(x) \geq p_{\min} > 0$ for $|x| \leq R_0$. Thus, we bound $I_n(\{ |\theta_{j,k} \neq 0 \}, \bar{\Theta})$ from below by

$$= p_{\min} \int_{-R_0 - \theta^0_{j,k}}^{R_0 - \theta^0_{j,k}} \exp \left\{ - \frac{1}{2 \sigma^2} \left( \frac{(\Psi_T \Psi_j)_{k,k}}{\beta_n(\bar{\Theta})} \left( x + \frac{\beta_n(\bar{\Theta})}{(\Psi_T \Psi_j)_{k,k}} \right)^2 \right) \right\} dx \geq p_{\min} \frac{2 \sqrt{2 \pi \sigma^2}}{\sqrt{(\Psi_T \Psi_j)_{k,k}}} \left[ 2 \Phi \left( \frac{\sqrt{(\Psi_T \Psi_j)_{k,k}}}{\sigma} \left( R_0 - \theta^0_{j,k} - \frac{\beta_n(\bar{\Theta})}{(\Psi_T \Psi_j)_{k,k}} \right) \right) - 1 \right],$$

21
where $\Phi$ is the cumulative distribution function of a standard normal. We proceed by lower bounding the expression inside $\Phi$. By the triangle inequality,

$$
|\theta^0_{j,k} - \frac{\beta_n(\bar{\Theta})}{(\Psi_j^T \Psi_j)_{k,k}}| \leq |\theta^0_{j,k}| + \frac{|\beta_n(\bar{\Theta}) - \beta_n(\bar{\Theta})|}{(\Psi_j^T \Psi_j)_{k,k}} + \frac{|\beta_n(\bar{\Theta})|}{(\Psi_j^T \Psi_j)_{k,k}}.
$$

(6.32)

By (6.4), we have $|\theta^0_{j,k}| \leq R$. On $\Omega_n(1)$, the second term above is $O_P(\sqrt{\log n/n}) = o_P(1)$ since $(\Psi_j^T \Psi_j)_{k,k} \geq C_1 n$ for any $(j,k) \in J_n(\gamma)$ by Lemma 7.2. Under $\mathcal{W}_n$ and applying Lemma 7.2 again, the third term above is $o(\sqrt{\log n/n}) = o(1)$. Then under the assumption $R \leq R_0 - 1/2$, we have for large $n$ that the right hand side of (6.32) is bounded by $R + \frac{1}{4} \leq R_0 - \frac{1}{4}$.

Hence, another application of Lemma 7.2 yields

$$
I_n([\theta_{j,k} \neq 0], \tilde{\Theta}) \geq p_{\min} \sqrt{\frac{2\pi\sigma^2}{(\Psi_j^T \Psi_j)_{k,k}}} \left[ 2\Phi \left( \frac{\sqrt{C_1 n}}{4\sigma} \right) - 1 \right].
$$

Using the fact that $P(|Z| > z) \leq 2e^{-z^2/2}$ for $z > 0$ and $Z \sim \mathcal{N}(0,1)$, we will obtain for any $\sigma \in \mathcal{U}_n$ and for $n$ large enough,

$$
2\Phi \left( \frac{\sqrt{C_1 n}}{4\sigma} \right) - 1 = 1 - P \left( |Z| > \frac{\sqrt{C_1 n}}{4\sigma} \right) \geq 1 - 2e^{-C_1 n/(32\sigma^2)} \geq 1/\sqrt{2}.
$$

Consequently, in view of Lemma 7.2, we have for large enough $n$,

$$
\int_{\mathcal{W}_n} I_n([\theta_{j,k} \neq 0], \tilde{\Theta})K_n(\tilde{\Theta})d\Pi(\tilde{\Theta}) \gtrsim \omega_{j,n} n^{-1/2} \sigma \int_{\mathcal{W}_n} K_n(\tilde{\Theta})d\Pi(\tilde{\Theta}).
$$

(6.33)

By assumption, $\omega_{j,n} \geq n^{-\lambda}$ with $\lambda > 0$. Thus, the first term on the right hand side of (6.30) is bounded above up to a constant by

$$
\sup_{\sigma \in \mathcal{U}_n} \sum_{\{\theta^0_{j,k} \geq \gamma \sqrt{\log n/n}\}} \frac{1 - \omega_{j,n}}{\sigma \omega_{j,n}} \frac{1 - \omega_{j,n}}{\sigma \omega_{j,n}} \lesssim \frac{1}{\sigma_0 + o(1)} n^{-\frac{C_1}{\pi(\sigma_0 + o(1))} \sigma^2 - \frac{3}{2}}.
$$

where we bounded $\sum_{\{\theta^0_{j,k} \geq \gamma \sqrt{\log n/n}\}}$ by $\sum_{j_1 = N_1}^{J_n - 1} \sum_{j_d = N_d}^{J_n - 1} \sum_{k_1 = 0}^{2^{j_1} - 1} \cdots \sum_{k_d = 0}^{2^{j_d} - 1}$. The right hand side will approach 0 if $\gamma$ is chosen large enough as $n \to \infty$. \hfill \Box

**Lemma 6.2.** For small enough $\gamma$ and large enough $\gamma$, there exists constant $P_3 > 0$ such that uniformly in $f_0 \in \mathcal{B}_{\infty,\infty}^\alpha(R)$ with $\alpha \in \mathbb{A}$ and any $0 < R \leq R_0 - 1/2$,

$$
E_0 \sup_{\sigma \in \mathcal{U}_n} \Pi(\mathcal{A}^c \cap C[Y, \sigma]) \leq \frac{(\log n)^d}{n^{P_3}}.
$$

**Proof of Lemma 6.2.** By (6.6), $\mathcal{A}^c = \cup_{\{\theta^0_{j,k} > \gamma \sqrt{\log n/n}\}}[\{\theta_{j,k} - \theta^0_{j,k} > \gamma \sqrt{\log n/n}\}]$. Split the union such that $\mathcal{A}^c = \mathcal{A}_1 \cup \mathcal{A}_2$ where $\mathcal{A}_1$ is union over $\{\gamma \sqrt{\log n/n} < |\theta^0_{j,k}| \leq \gamma \sqrt{\log n/n}\}$
and \( A_2 \) is over the complement \( \{ |\theta_{j,k}^0 | > \sqrt{T} \log n/n \} \). Then \( A^c \cap C = (A_1 \cap C) \cup (A_2 \cap C) \). Define \( Z_{j,k} := \{ |\theta_{j,k} - \theta_{j,k}^0 | > \sqrt{T} \log n/n \} \cap \{ \theta_{j,k} \neq 0 \} \). In view of (6.29), observe that
\[ A_1 \cap C = \bigcup \{ [T \log n/n < |\theta_{j,k}^0 | \leq \sqrt{T} \log n/n] \} Z_{j,k}; \]
while \( A_2 \cap C = \bigcup \{ |\theta_{j,k}^0 | > \sqrt{T} \log n/n \} \) \( Z_{j,k} \). Therefore by using a union bound, we can bound \( E_0 \sup_{\sigma \in I_{\gamma}} \Pi(A^c \cap C|Y, \sigma) \) from above by

\[
E_0 \sup_{\sigma \in I_{\gamma}} \sum_{(j,k) \in J_n(\gamma)} \Pi(\theta_{j,k} \in Z_{j,k}|Y, \sigma) \mathds{1}_{\Omega_{n}(1)} + P_{0}(\Omega_{n}(1)^c). \tag{6.34}
\]

In view of (6.31), the second term is bounded above by \( n^{-1/2}(\log n)^d \), which goes to 0 as \( n \to \infty \). Using the same decomposition as in (6.21), we find that

\[
\Pi(\theta_{j,k} \in Z_{j,k}|Y, \sigma) \mathds{1}_{\Omega_{n}(1)} \leq \frac{\int_{W_n} I_n(Z_{j,k}, \Theta)K_n(\Theta) d\Pi(\Theta)}{\int_{W_n} I_n(\theta_{j,k} \neq 0, \Theta)K_n(\Theta) d\Pi(\Theta)} \mathds{1}_{\Omega_{n}(1)} + \Pi(W_n^c|Y, \sigma),
\]

where \( \Pi(W_n^c|Y, \sigma) = O_{P_0}(n^{-B}) \) for some constant \( B > 0 \) follows from Lemma 7.6. Recall that when \( \theta_{j,k} \neq 0 \), then \( d\Pi(\theta_{j,k}) \leq C_{\theta} \sigma_{\max} \sqrt{2\log 2} \). It follows that \( I_n(Z_{j,k}, \Theta) \) is bounded above by

\[
p_{\max, \omega, \gamma} \int_{|x| > \gamma \sqrt{\log n/n}} \exp \left\{ -\frac{\langle \Psi_j^T \Psi_j \rangle_{k,k}}{2\sigma^2} \left( x + \frac{\beta_n(\Theta)}{(\Psi_j^T \Psi_j)_{k,k}} \right)^2 \right\} dx.
\]

Since \( (j, k) \in J_n(\gamma) \subseteq J_n(\alpha) \) by (6.10), we have \( 2\sum_{i=1}^d J_i = o(n) \) and \( (\Psi_j^T \Psi_j)_{k,k} \geq \sqrt{C_1 n} \) by Lemma 7.2. Then if \( |x| > \gamma \sqrt{\log n/n} \) and under \( W_n \) and \( \Omega_{n}(1) \), we have by twice application of the reverse triangular inequality that

\[
\left| x + \frac{\beta_n(\tilde{\Theta})}{(\Psi_j^T \Psi_j)_{k,k}} \right| > |x| - \left| \frac{\beta_n(\tilde{\Theta})}{(\Psi_j^T \Psi_j)_{k,k}} \right| + \left| \frac{\beta_n(\Theta)}{(\Psi_j^T \Psi_j)_{k,k}} - \beta_n(\tilde{\Theta}) \right| \geq \sqrt{\frac{\log n}{n}} - \frac{\tau_n}{C_1} \sqrt{\frac{\log n}{n}} - \frac{\sigma_0}{\sqrt{C_1 n}} \sqrt{2\log 2\sum_{i=1}^d J_i} + \log n > \frac{\gamma}{2} \sqrt{\frac{\log n}{n}},
\]

if \( \gamma \) is chosen large enough since \( \tau_n \to 0 \) as \( n \to \infty \). We conclude that

\[
\left\{ |x| \geq \frac{\sqrt{\log n}}{n} \right\} \subset \left\{ x + \frac{\beta_n(\tilde{\Theta})}{(\Psi_j^T \Psi_j)_{k,k}} > \frac{1}{2} \gamma \sqrt{\frac{\log n}{n}} \right\}.
\]

Therefore, \( I_n(Z_{j,k}, \tilde{\Theta}) \) is further bounded above by

\[
p_{\max, \omega, \gamma} \exp \left\{ -\frac{\langle \Psi_j^T \Psi_j \rangle_{k,k} \log n}{16n\sigma^2} \right\} \int_{\mathbb{R}} \exp \left\{ -\frac{\langle \Psi_j^T \Psi_j \rangle_{k,k}}{4\sigma^2} \left( x + \frac{\beta_n(\tilde{\Theta})}{(\Psi_j^T \Psi_j)_{k,k}} \right)^2 \right\} dx \leq n^{-1/2} \sigma_{\omega, n} \exp \left\{ -\frac{C_1 \gamma^2 \log n}{16\sigma^2} \right\},
\]

23
again utilizing the bounds in Lemma 7.2. Using the upper bound established above with the lower bound for \( \int_{ \mathcal{W} } I_n(\theta, k \neq 0, \Theta) K_n(\Theta) d\Pi(\Theta) \) derived (6.33), the first term in (6.34) is bounded above up to some constant by

\[
\sup_{\sigma \in \mathbb{H}} \sum_{j_1 = N_1}^{J_{n,1} - 1} \cdots \sum_{j_d = N_d}^{J_{n,d} - 1} \sum_{k_1 = 0}^{2^{j_1} - 1} \cdots \sum_{k_d = 0}^{2^{j_d} - 1} n^{-\gamma(C_1 + 2(16\sigma^2))} \sigma \lesssim (\sigma_0 + o(1)) n^{-\frac{C_1}{16(16\sigma^2 + 1)/2 - 1}},
\]

and will approach 0 if \( \gamma \) is large enough as \( n \to \infty \). \( \square \)

**Proof of Corollary 4.5.** First note that the loss \( f \mapsto \| f - f_0 \|_\infty \) is unbounded but convex for any \( f_0 \). For \( u \in \mathbb{N} \), define slices \( \mathcal{F}_u := \{ M \epsilon_{n,r} u < \| D^r f - D^r f_0 \|_\infty \leq M \epsilon_{n,r} (u + 1) \} \) and sieve \( \mathcal{F} := \{ \| D^r f - D^r f_0 \|_\infty > M \epsilon_{n,r} \} \).

Now, if we introduced an extra factor in 3 places: the right hand sides of the definitions of \( \mathcal{F}_u(\gamma) \) in (6.5), \( A \) in (6.6) and in (6.18) by replacing \( c \log n \) with \( cu^2 \log n \), we see that by slightly modifying the proof of Theorem 4.2,

\[
E_0(\mathcal{F}_u|Y) \leq (\log n)^d \exp \{-C \log (n) u^2\},
\]

for some universal constant \( C > 0 \). By observing that \( \mathcal{F} = \bigcup_{u=1}^{\infty} \mathcal{F}_u \), we can write

\[
E_0(\| D^r f - D^r f_0 \|_\infty | Y) \leq M \epsilon_{n,r} + \sum_{u=1}^{\infty} E_0(1_{\mathcal{F}_u} \| D^r f - D^r f_0 \|_\infty | Y)
\leq M \epsilon_{n,r} \left( 1 + 2(\log n)^d \sum_{u=1}^{\infty} u e^{-C \log (n) u^2} \right) \lesssim \epsilon_{n,r},
\]

where the sum \( (\log n)^d \sum_{u=1}^{\infty} u e^{-C \log (n) u^2} = (\log n)^d (n^{-C} + 2n^{-4C} + 3n^{-9C} + \cdots) \) converges when \( n \) is large enough. By Jensen’s inequality, \( \| E(D^r f | Y) - D^r f_0 \|_\infty \leq E(\| D^r f - D^r f_0 \|_\infty | Y) \) and the result follows by taking \( E_0 \) on both sides. \( \square \)

### 6.3 Proof of results in Section 5

Let \( \| \cdot \|_n \) be the \( L_2 \)-norm with respect to the empirical measure of \( \{ X_1, \ldots, X_d \} \) and \( 1_d \) be a \( d \)-dimensional vector of ones. For the proofs in this subsection, there is no qualitative difference in distinguishing between father and mother wavelet coefficients, and for notational simplicity, we combine the father and mother parts of the wavelet expansion in (3.1) into a single sum. To that end, let us write the wavelet projection of \( f \) at resolution levels \( \mathbf{J}_n = (J_{n,1}, \ldots, J_{n,d})^T \) as \( K_{\mathbf{J}_n}(f)(\mathbf{x}) = \sum_{j_1 = N_1 - 1}^{J_{n,1} - 1} \cdots \sum_{j_d = N_d - 1}^{J_{n,d} - 1} \sum_{k_1 = 0}^{2^{j_1} - 1} \cdots \sum_{k_d = 0}^{2^{j_d} - 1} \theta_{j,k} \phi_{j,k}(\mathbf{x}) \), by delegating the father wavelets and their coefficients to the level \( j_l = N_l - 1, l = 1, \ldots, d \).

**Proof of Proposition 5.1.** Let \( g \in \mathcal{F}_n \) be such that \( \| g - f_0 \|_\infty > M \epsilon_n \) and \( \| g - f_0 \|_n \leq C \inf_{f \in \mathcal{F}_n} \| f - f_0 \|_\infty > M \epsilon_n \) for some constant \( C > 0 \). Denote \( \phi_{LR} \) to be the likelihood ratio test for the simple hypotheses \( H_0 : f = f_0 \) versus \( H_1 : f = g \). Furthermore, let \( \phi_n \) be
any test for the same hypotheses such that its Type I error is bounded by $0 < \delta < 1$. By a change of Gaussian measure and using the Cauchy-Schwartz inequality, we can write

$$E_{f_0}(1 - \phi_{LR}) \leq \sqrt{E_g(1 - \phi_{LR})} \sqrt{\int \left( \frac{dP_n}{dP_g} \right)^2 dP_g} \leq \sqrt{E_g(1 - \phi_{LR})} e^n \|f_0 - g\|^2_n/(2\sigma^2). \tag{6.35}$$

Since $\phi_{LR}$ is the uniformly most powerful test, it follows that $E_{f_0}\phi_{LR} \leq E_{f_0}\phi_{n} \leq \delta$ and

$$E_g(1 - \phi_{n}) \geq E_g(1 - \phi_{LR}) \geq \left[ E_{f_0}(1 - \phi_{LR}) \right]^2 e^{-n} \|f_0 - g\|^2_n/\sigma^2 \geq (1 - \delta)^2 \exp \left\{ -\frac{C^2 n}{\sigma^2} \inf_{f \in F: \|f - f_0\| > M\epsilon_n} \|f - f_0\|_n^2 \right\}, \tag{6.36}$$

by virtue of (6.35) above and the assumption on $g$. Take $2J_{n,t} = n^{1/(2\alpha)}$, and since $2\Sigma_{l=1}^d J_{n,t} = o(n)$ because of $\alpha^* > d/2$, we have by intersecting with $F_n$ and the triangle inequality that $\|f - f_0\|_n$ is bounded above by

$$\|K_{J_n}(f) - K_{J_n}(f_0)\|_n + \|f - K_{J_n}(f)\|_n + \|f_0 - K_{J_n}(f_0)\|_n \leq \|\theta - \theta_0\| + \sum_{l=1}^d 2^{-\alpha_l J_{n,t}},$$

where we have used Lemma 6.4 to bound the first term and utilizing Proposition 3.2 for the last. By the continuous embedding of Proposition 4.3.11 in [9] and Remark 3.1, $L_\infty \subset B^0_{\infty,\infty}$ and hence $\|f - f_0\|_{B^0_{\infty,\infty}} \leq \|f - f_0\|_\infty$. We then conclude that

$$\inf_{f \in F: \|f - f_0\| > M\epsilon_n} \|f - f_0\|_n \lesssim \inf_{\theta \in A_n} \|\theta - \theta_0\| + \frac{d}{\sqrt{n}}$$

with $A_n := \{ \theta \in F_n : \max_j 2\Sigma_{l=1}^d j^{1/2} \max_k |\theta_{j,k} - \theta_{0,j,k}^*| > M\epsilon_n \}$. Let $J_{n,t}(\alpha)$ be

$$\frac{1}{2} \left( \frac{n}{\log n} \right) \frac{\alpha^*}{\alpha_l(2\alpha_l + d)} \leq \left( \frac{M}{R} \right) \left[ \frac{\Sigma_{l=1}^d \alpha_l(\frac{1}{2} + \frac{1}{2\alpha_l})}{\log n} \right]^{-1} 2^{J_{n,t}(\alpha)} \leq \left( \frac{n}{\log n} \right) \frac{\alpha^*}{\alpha_l(2\alpha_l + d)}.$$ \tag{6.37}

Consider a $\theta^* \in F_n$ such that $|\theta_{j,k}^* - \theta_{0,j,k}^*| = R 2^{-\Sigma_{l=1}^d \alpha_l J_{n,t}(\alpha)(\frac{1}{2} + \frac{1}{2\alpha_l})}$ for $j_l = J_{n,t}(\alpha), l = 1 \ldots, d$ with $k = 0$, but $|\theta_{j,k}^* - \theta_{0,j,k}^*| = 0$ for all the other multi-indices. Then,

$$\max_j 2\Sigma_{l=1}^d j^{1/2} \max_k |\theta_{j,k}^* - \theta_{0,j,k}^*| > M2\Sigma_{l=1}^d J_{n,t}(\alpha)^2 \sqrt{\log n/n} \geq M\epsilon_n,$$

and this implies that $\theta^* \in A_n$, but since $J_{n,t}(\alpha) \leq J_{n,t}$,

$$\|\theta^* - \theta_0\| + \frac{d}{\sqrt{n}} = R 2^{-\Sigma_{l=1}^d \alpha_l J_{n,t}(\alpha)(\frac{1}{2} + \frac{1}{2\alpha_l})} + \frac{d}{\sqrt{n}} \lesssim \sqrt{\log n/n}.$$

Therefore, $\inf_{f \in F: \|f - f_0\| > M\epsilon_n} \|f - f_0\|_n^2 \lesssim \log n/n$ and plugging this back into (6.36) gives the result. \hfill \Box
Proof of Proposition 5.2. Write \( f_{J_n}(x) = \psi_{J_n}(x)^T \theta \), where \( \psi_{J_n}(x) \) is constructed by concatenating \( \psi_{j,k}(x) \) across \( \mathcal{N} \leq j \leq J_n - 1_d \) and all \( k \) in lexicographic order. Construct the wavelet basis matrix as \( \Psi := (\psi_{J_n}(X_1)^T, \cdots, \psi_{J_n}(X_n)^T)^T \). Let us consider the most natural plug-in test \( \phi_n = 1\{\|\hat{f}_{J_n} - f_0\|_\infty > M_0 \rho_n \epsilon_n \} \) for some constant \( M_0 < M \), by using the least squares estimate \( \hat{f}_{J_n}(x) := \psi_{J_n}(x)^T (\Psi^T \Psi)^{-1} \Psi^T Y \).

Let \( \rho_n \to \infty \) as \( n \to \infty \) to be determined below and assume that \( \epsilon_n \geq \sum_{l=1}^{d} 2^{-\alpha_l J_n} \). By the triangle inequality and in view of (6.38) of Lemma 6.3,

\[
E_{0} \phi_n = P_0 \left( \|\hat{f}_{J_n} - f_0\|_\infty > M_0 \rho_n \epsilon_n \right) \leq P_0 \left( \|\hat{f}_{J_n} - E_0 \hat{f}_{J_n}\|_\infty > M_0 \rho_n \epsilon_n - \|E_0 \hat{f}_{J_n} - f_0\|_\infty \right)
\]

\[
\leq P_0 \left( \|\hat{f}_{J_n} - E_0 \hat{f}_{J_n}\|_\infty > M_0 \rho_n \epsilon_n/2 \right),
\]

when \( n \) is large enough. Now apply (6.39) of Lemma 6.3 with \( x = C \gamma n \epsilon_n^2 \), and if

\[
(M_0/2) \rho_n \epsilon_n \geq Q_1 \sqrt{\frac{2 \sum_{l=1}^{J_n} \log n}{n}} + \sqrt{2 Q_2 C_\epsilon \sum_{l=1}^{d} (2^{J_n l})^2 \epsilon_n},
\]

then \( E_0 \phi_n \leq e^{-C_\epsilon n \epsilon_n^2} \). Take \( 2 \sum_{l=1}^{J_n} \epsilon_n = (n/ \log n)^2 \) and \( C_\epsilon = Q_1^2/(2 Q_2) \), then the above inequality is equivalent to \( \rho_n \geq \frac{Q_1}{M_0/2} (1 + 2 \sum_{l=1}^{d} (2^{J_n l})^2 \epsilon_n) \). For \( \epsilon_n = (n/ \log n)^{-\alpha} (2^{\alpha + d}) \) the targeted rate, it follows that we will have exponential Type I error if \( \rho_n \geq Q_1 (M_0/2)^{-1}[1 + (n/ \log n)^d (2^{\alpha + d})] \).

For the Type II error with \( f \in \mathcal{F}_n : \|f - f_0\|_\infty > M \rho_n \epsilon_n \) such that \( M > M_0 \), we apply the reverse triangle inequality twice to yield

\[
E_f (1 - \phi_n) = P_f \left( \|\hat{f}_{J_n} - f_0\|_\infty \leq M_0 \rho_n \epsilon_n \right)
\]

\[
\leq P_f \left( \|\hat{f}_{J_n} - E_f \hat{f}_{J_n}\|_\infty \geq \|f - f_0\|_\infty - M_0 \rho_n \epsilon_n - \|E_f \hat{f}_{J_n} - f\|_\infty \right).
\]

Now by intersecting with \( \mathcal{F}_n \) and applying (6.38) with \( f_0 \) instead of \( f_0 \), \( \|E_f \hat{f}_{J_n} - f\|_\infty \lesssim \epsilon_n \) if \( \epsilon_n \geq \sum_{l=1}^{d} 2^{-\alpha_l J_n} \). Therefore for large enough \( n \), we can use the same argument as for the Type I error case to conclude that for any \( \rho_n \geq 2 Q_1 (M - M_0)^{-1}[1 + (n/ \log n)^d (2^{\alpha + d})] \), we will have \( \sup_{f \in \mathcal{F}_n : \|f - f_0\|_\infty > M \rho_n \epsilon_n} E_f (1 - \phi_n) \leq e^{-C_\epsilon n \epsilon_n^2} \) for constant \( C_\epsilon = Q_2^2/(2 Q_2) \). The second assertion is proved using the master theorem (see Theorem 3 of [7]) once we have test with exponential errors. The KL neighborhood and prior complement criteria follow the same steps as in Lemma 7.4 to prove \( L_2 \)-contraction. \( \square \)

Lemma 6.3. Project \( f \) unto wavelet basis at resolution \( J_n \) and write the model in (1.1) as \( Y = \Psi \theta + \epsilon \). Let \( \hat{f}_{J_n}(x) := \psi_{J_n}(x)^T (\Psi^T \Psi)^{-1} \Psi^T Y \) be the corresponding least squares estimator. If \( 2 \sum_{l=1}^{d} J_n l = o(n) \), then there exist constants \( C_0, Q_1, Q_2 > 0 \) such that the
following holds:

\[
\|E_0 \hat{f}_n - f_0\|_\infty \leq C_0 \sum_{l=1}^d 2^{-\alpha_I n,l}, \quad (6.38)
\]

\[
P_0 \left( \|\hat{f}_n - E_0 \hat{f}_n\|_\infty \geq Q_1 \sqrt{\frac{2\sum_{l=1}^d J_{n,l} \log n}{n}} + \sqrt{2Q_2 \frac{2\sum_{l=1}^d J_{n,l}}{n} x} \right) \leq e^{-x}. \quad (6.39)
\]

Furthermore for any \( f \in \mathcal{F}_n \), (6.38) still holds if we replaced \( f_0 \) for \( f \) and \( E_0 \) with \( E_f \). The same can be said for (6.39).

**Proof.** By Proposition 3.2, for any \( f_0 \in \mathcal{B}^\infty (R) \), there is a \( \theta_0 \) such that \( \|f_0 - \psi J_n(\cdot)^T \theta_0\|_\infty \lesssim \sum_{l=1}^d 2^{-\alpha_I n,l} \). Therefore by adding and subtracting \( \Psi \theta_0 \) and using the triangle inequality,

\[
\|E_0 \hat{f}_n - f_0\|_\infty = \|\psi J_n(\cdot)^T (\Psi^T \Psi)^{-1} \Psi^T F_0 - f_0\|_\infty \leq \|\psi J_n(\cdot)^T (\Psi^T \Psi)^{-1} \Psi^T (F_0 - \Psi \theta_0)\|_\infty + \|\psi J_n(\cdot)^T \theta_0 - f_0\|_\infty,
\]

where the second term is \( O(\sum_{l=1}^d 2^{-\alpha_I n,l}) \). For a matrix \( A \), let \( \|A\|_{(\infty, \infty)} = \max_{i} \sum_{j} |a_{ij}| \) (max of absolute row sums) and \( \|A\|_{(1,1)} = \max_{j} \sum_{i} |a_{ij}| \) (max of absolute column sums). Using H"older’s inequality \( \|x^T y\| \leq \|x\|_1 \|y\|_\infty \), definition of the induced matrix norm \( \|A x\|_1 \leq \|A\|_{(1,1)} \|x\|_1 \), the first term is bounded by

\[
\sup_{x \in [0,1]^d} \|\psi J_n(x)\|_1 \|((\Psi^T \Psi)^{-1})_{(1,1)}\| \|\Psi^T\|_{(1,1)} \|F_0 - \Psi \theta_0\|_\infty. \quad (6.40)
\]

By (6.3) with \( r = 0 \), it holds that \( \|\psi J_n(x)\|_1 \lesssim 2^{\sum_{l=1}^d J_{n,l}/2} \) uniformly in \( x \in [0,1]^d \). Note that since each entry of \( \Psi \) is a dilated and translated version of the base CDV wavelet with compact support, it follows that \( \Psi^T \Psi \) is banded. Furthermore by choosing \( 2^{\sum_{l=1}^d J_{n,l}} = o(n) \), all eigenvalues of \( \Psi^T \Psi \) are \( \approx n \) by virtue of Lemma 6.4. Therefore by appealing to Lemma A.4 of [25], we conclude \( \|((\Psi^T \Psi)^{-1})_{(\infty, \infty)} \| \lesssim n^{-1} \). Since \( (\Psi^T \Psi)^{-1} \) is symmetric, it follows that \( \|((\Psi^T \Psi)^{-1})_{(1,1)}\| \|((\Psi^T \Psi)^{-1})_{(\infty, \infty)} \| \lesssim n^{-1} \). Note that \( \|\Psi^T\|_{(1,1)} = \max_{1 \leq i \leq n} \sum_{j,k} |\psi_{j,k}(X_i)| \) and this is \( O(2^{\sum_{l=1}^d J_{n,l}/2}) \) as shown above. It now remains to bound \( \|F_0 - \Psi \theta_0\|_\infty \), and we know it is \( O(2^{\sum_{l=1}^d 2^{-\alpha_I n,l}}) \). Combine everything and use the assumption that \( 2^{\sum_{l=1}^d 2^{-\alpha_I n,l}} \leq n \) to conclude (6.38). Notice that if \( \|f - KJ_n(f)\|_\infty \lesssim \sum_{l=1}^d 2^{-\alpha_I n,l} \), the same conclusion still holds by replacing \( f_0 \) for \( f \) and \( E_0 \) with \( E_f \).

By Assumption 1, \( \hat{f}_n - E_0 \hat{f}_n \sim \text{GP}(0, \sigma^2 \Sigma_{J_n}) \), where the covariance kernel \( \Sigma_{J_n}(x, y) = \psi J_n(x)^T (\Psi^T \Psi)^{-1} \psi J_n(y) \) for any \( x, y \in [0,1]^d \). Note the fact that for any \( x \in [0,1]^d \), \( \|\psi J_n(x)\|^2 \leq \max_{j \leq J_n - 1, k} |\psi_{j,k}(x)| \sum_{j \leq J_n - 1, k} |\psi_{j,k}(x)| \lesssim \prod_{l=1}^d 2^{J_{n,l}} \), since the wavelets are uniformly bounded and applying (6.3) with \( r = 0 \). It then follows by appealing to Lemma 6.4 that

\[
\sup_{x \in [0,1]^d} \Sigma_{J_n}(x, x) \lesssim \|((\Psi^T \Psi)^{-1})_{(2,2)}\| \sup_{x \in [0,1]^d} \|\psi J_n(x)\|^2 \leq Q_2 n^{-1} 2^{\sum_{l=1}^d J_{n,l}},
\]
for some constant $Q_2 > 0$. By the Borell's inequality (see Proposition A.2.1 from [24] or Theorem 2.5.8 in [9]), we have for any $x \geq 0$ that

$$P_0 \left( \| \hat{f}_{J_n} - E_0 \hat{f}_{J_n} \|_{\infty} \geq E_0 \| \hat{f}_{J_n} - E_0 \hat{f}_{J_n} \|_{\infty} + \sqrt{2Q_2 n^{-1} 2\sum_{l=1}^{J_n} x} \right) \leq e^{-x}.$$  

(6.41)

Define $\xi := \Psi^T \epsilon$. Observe that $\hat{f}_{J_n} - E_0 \hat{f}_{J_n}$ can be expressed as $\psi_{J_n}(x)^T (\Psi^T \Psi)^{-1} \xi$, and by Hölder's inequality,

$$E_0 \| \hat{f}_{J_n} - E_0 \hat{f}_{J_n} \|_{\infty} \leq \sup_{x \in [0,1]^d} \| \psi_{J_n}(x) \|_1 \| (\Psi^T \Psi)^{-1} \|_{1,\infty} \| \xi \|_\infty \lesssim n^{-1} 2^{\sum_{l=1}^{J_n} x} E_0 \| \xi \|_{\infty},$$

in view of the bounds established in (6.40). We adopt the convention by indexing the rows and columns of $\Psi^T \Psi$ with arrays of the form $(j, k)$. Keeping in mind that by Assumption 1, $\xi \sim N(0, \sigma_0^2 \Psi^T \Psi)$, we can apply Lemma 2.3.4 of [9] to conclude that for $Z_{j,k} \sim N(0,1)$ i.i.d with $(j, k)$ running across all the indices in the wavelet series up to resolution $J_n$,

$$E_0 \| \xi \|_\infty \lesssim \max_{j \leq J_n} \max_{l \leq J_n} \sqrt{(\Psi^T \Psi)(j,k),(j,k)} E \left( \max_{j \leq J_n} \max_{l \leq J_n} | Z_{j,k} | \right) \approx \sqrt{n \log n},$$

where we have utilized Lemma 6.4 to upper bound the diagonals of $\Psi^T \Psi$, and used the assumption $2^{\sum_{l=1}^{J_n} x} = o(n)$. Combine all these established bounds back into (6.41) to deduce (6.39). Notice that (6.39) still holds by replacing $P_0$ and $E_0$ with $P_f$ and $E_f$ because $\hat{f}_{J_n} - E_f \hat{f}_{J_n}$ is also a Gaussian process under our working model assumption in (1.1).

\begin{lemma}
For any $\theta$, there exist constants $C_1, C_2 > 0$ such that

$$C_1 \| \theta \|^2 \left( n - \prod_{l=1}^{d} 2^{J_{n,l}} \right) \leq n \| K_{J_n}(f) \|^2 = \theta^T \Psi^T \Psi \theta \leq C_2 \| \theta \|^2 \left( n + \prod_{l=1}^{d} 2^{J_{n,l}} \right).$$  

(6.42)

In particular if $2^{\sum_{l=1}^{J_n} x} = o(n)$, then the maximum eigenvalue of $\Psi^T \Psi$ is $O(n)$, while its minimum eigenvalue is $\gtrsim n$.
\end{lemma}

\begin{proof}
By definition, $n^{-1} \theta^T \Psi^T \Psi \theta = \int_{[0,1]^d} K_{J_n}(f)(x)^2 dG_n(x)$. Apply Lemma 7.1 below with $f = K_{J_n}(f)$. Then $\int_{[0,1]^d} K_{J_n}(f)(x)^2 d\lambda = \| \theta \|^2$ by orthonormality. By the Cauchy-Schwarz inequality,

$$\int_{[0,1]^d} \left| \frac{\partial^d}{\partial x_1 \cdots \partial x_d} K_{J_n}(f)(x) \right|^2 dx \leq 2 \| K_{J_n}(f) \|_2 \left\| \frac{\partial^d}{\partial x_1 \cdots \partial x_d} K_{J_n}(f) \right\|_2,$$

where $\| K_{J_n}(f) \|_2 = \| \theta \|$ again by orthonormality, while

$$\left\| \frac{\partial^d}{\partial x_1 \cdots \partial x_d} K_{J_n}(f) \right\|_2 \lesssim \sqrt{\sum_{j_1=J_{n,1}}^{J_{n,d}-1} \cdots \sum_{j_d=J_{n,d}^{-1}}^{J_{n,d}^{-1}} \sum_{j_{d+1}=J_{n,d+1}}^{J_{n,d+1}^{-1}} 2 \sum_{l=1}^{d} 2^{J_{n,l}} \| \theta \|},$$

follows by applying the third display of Section 5 in [2]. The last statement follows since the maximum or minimum eigenvalue is the maximization or minimization of $\theta^T \Psi^T \Psi \theta/\| \theta \|^2$ over $\theta \neq 0$.
\end{proof}
7 Technical lemmas

The lemma below quantifies the error in approximating Riemann’s sum with integral, and is useful to give size estimates of discrete sums found in this paper. Let \( \langle f, g \rangle \) be the inner product of two functions \( f, g \) in Hilbert space.

**Lemma 7.1.** Suppose \( \partial^d f / (\partial x_1 \cdots \partial x_d) \in L_1 \), then if the fixed design points are chosen such that (3.3) holds, we will have for some constant \( C > 0 \),

\[
\left| \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{X}_i) - \int_{[0,1]^d} f(x) dx \right| \leq C \frac{1}{n} \int_{[0,1]^d} \left| \partial^d f(x) \right| dx.
\]

*Proof.* Note that \( n^{-1} \sum_{i=1}^{n} f(\mathbf{X}_i) = \int_{[0,1]^d} f(x) dG_n(x) \) where \( G_n(x) \) is the empirical distribution. By the triangle inequality,

\[
\left| \int_{[0,1]^d} f(x) dG_n(x) \right| \leq \left| \int_{[0,1]^d} f(x) dU(x) \right| + \left| \int_{[0,1]^d} f(x) d(G_n - U)(x) \right|,
\]

where \( U(x) \) is the Uniform\([0, 1]^d\) cumulative distribution function. Thus, the first term is \( \left| \int_{[0,1]^d} f(x) dx \right| \). By the multivariate integration by parts, \( \int_{[0,1]^d} f(x) d(G_n - U)(x) \) is

\[
f(x)(G_n - U)(x)|_0^1 + \int_{[0,1]^d} \partial^d f(x) (G_n - U)(x) dx.
\]

Since \( (G_n - U)(1_d) = (G_n - U)(0) = 0 \), it follows by assumption (3.3) that the second term in (7.1) is bounded above by

\[
\|G_n - U\|_\infty \int_{[0,1]^d} \left| \partial^d f(x) \right| dx \lesssim \frac{1}{n} \int_{[0,1]^d} \left| \partial^d f(x) \right| dx.
\]

For the other direction, we apply the reverse triangle inequality to yield

\[
\left| \int_{[0,1]^d} f(x) dG_n(x) \right| \geq \left| \int_{[0,1]^d} f(x) dx \right| - \left| \int_{[0,1]^d} f(x) d(G_n - U)(x) \right|,
\]

and the assertion follows by virtue of the upper bound established for the second term. \( \square \)

**Lemma 7.2.** Under the assumption of (3.3), there exists constant \( C > 0 \) such that

\[
\| (\Psi^T_a \Psi_b)_{c,e} - n \| (\psi_{a,c}, \psi_{b,e}) \| \leq C \prod_{l=1}^{d} 2^{(\alpha_l + \beta_l)/2}, \quad \| (\Psi^T_a B)_{c,m} \| \leq C \prod_{l=1}^{d} 2^{\alpha_l/2}.
\]

In particular, for \( j_l \leq \tilde{J}_{n,l}, l = 1, \ldots, d \) where \( \tilde{J}_{n,l} \) is increasing with \( n \), then if \( 2 \sum_{l=1}^{d} \tilde{J}_{n,l} = o(n) \), this implies that \( (\Psi_j^T \Psi_j)_{k,k} \approx n \) by the orthonormality of \( \psi_{j,k}(\cdot) \), and

\[
\sum_{u=1}^{n} |\psi_{j,k}(\mathbf{X}_u)| \lesssim n \prod_{l=1}^{d} 2^{-j_l/2}.
\]

29
Proof. To prove the first statement of (7.2), apply Lemma 7.1 with \( f(x) = \psi_{a,c}(x)\psi_{b,e}(x) \). Let us denote \( I(f) := \int_{[0,1]^d} \frac{\partial^d}{\partial x_1 \cdots \partial x_d} f \, dx \). Then,

\[
I(f) = \int_{[0,1]^d} \left| \psi_{a,c}(x) \frac{\partial^d}{\partial x_1 \cdots \partial x_d} \psi_{b,e}(x) + \psi_{b,e}(x) \frac{\partial^d}{\partial x_1 \cdots \partial x_d} \psi_{a,c}(x) \right| dx.
\]

Recall that \( \psi_{j,k} \) is a dilated and translated tensor product of \( \psi_l, l = 1, \ldots, d \). Since the wavelets are compactly supported by assumption, we can write the support of \( \psi_l \) as \([A_{1,l}, A_{2,l}]\), and the support of \( \psi_{a,c} \) is then \( \mathcal{I}_a = \prod_{l=1}^d [(A_{1,l} + c_l)2^{-a_l}, (A_{2,l} + c_l)2^{-a_l}] \). Therefore, we can restrict the domain of integration to \( \mathcal{I}_a \cap \mathcal{I}_b \), which is the intersection of the supports of \( \psi_{a,c} \) and \( \psi_{b,e} \). Thus, \( I(f) \) is bounded above up to some constant by

\[
\prod_{l=1}^d 2^{(a_l+b_l)/2} \left( \int_{\mathcal{I}_b} \int_{\mathcal{I}_a} 2^h \|\psi_l\|_\infty \|\psi'_l\|_\infty dx + \int_{\mathcal{I}_a} \int_{\mathcal{I}_b} 2^h \|\psi_l\|_\infty \|\psi'_l\|_\infty dx \right),
\]

which has the order \( \prod_{l=1}^d 2^{(a_l+b_l)/2} \) since the wavelets and their derivatives are uniformly bounded by construction.

For the second assertion of (7.2), take \( f(x) = \psi_{a,c}(x)\varphi_{N,m}(x) \). Then \( \langle \psi_{a,c}, \varphi_{N,m} \rangle = 0 \) by orthonormality, and \( I(f) \) is bounded above by some constant by \( \prod_{l=1}^d 2^{a_l/2}(1 + \int_{\mathcal{I}_a} \prod_{l=1}^d 2^{a_l} \, dx) \lesssim \prod_{l=1}^d 2^{a_l/2}, \) and the assertion follows by Lemma 7.1.

We choose \( f(x) = |\psi_{j,k}(x)| \) to prove (7.3). Then, \( \int_{[0,1]^d} f(x) \, dx = \prod_{l=1}^d 2^{-j_l/2} \|\psi_l\|_1 \), where \( \|\psi_l\|_1 < \infty \) since \( \psi_l, \varphi_l \in L_\infty \subseteq L_1, l = 1, \ldots, d \) on \([0,1]\) by assumption. By restricting the domain of integration to \( \mathcal{I}_j \),

\[
I(f) \leq \int_{\mathcal{I}_j} \left| \int_{\mathcal{I}_l} 2^{j_l/2+j_l} \psi'_l(2^j x_l - k_l) \text{sgn}[\psi(2^j x_l - k_l)] \right| \, dx \leq \prod_{l=1}^d 2^{j_l/2},
\]

with \( \text{sgn}(\cdot) \) denoting the sign function, i.e., \( \text{sgn}(x) = 1 \) if \( x \geq 0 \) and \( -1 \) if \( x < 0 \). Therefore if \( j_l \leq \tilde{J}_{n,l} \) where \( 2\sum_{l=1}^d \tilde{J}_{n,l} = o(n) \), then the above is \( o(1) \) and the result follows.

\[ \square \]

Remark 7.3. For random design points, the stochastic version of Lemma 7.1 can be deduced from Bernstein’s inequality (see (3.24) in Theorem 3.1.7 of [9]). In this case, (7.2) is:

\[
P \left( \left| \langle \Psi^T_a \Psi_b \rangle_{c,e} - n \langle \psi_{a,c}, \psi_{b,e} \rangle \right| \lesssim \sqrt{n} \prod_{l=1}^d 2^{(a_l+b_l)/2} \right) \geq 1 - 2e^{-2^{\sum_{l=1}^d \frac{(a_l+b_l)}{2}}},
\]

\[
P \left( \left| \langle \Psi^T_a B \rangle_{c,m} \right| \lesssim \sqrt{n} \prod_{l=1}^d 2^{a_l/2} \right) \geq 1 - 2e^{-2^{\sum_{l=1}^d a_l}}.
\]

The extra \( \sqrt{n} \) is due to the fact that \( \|G_n - U\|_\infty = O_P(n^{-1/2}) \) in the random case, instead of the rate \( O(n^{-1}) \) in the fixed design case. In particular, for \( j_l \leq \tilde{J}_{n,l} \), \( l = 1, \ldots, d \) where \( \tilde{J}_{n,l} \) is increasing with \( n \), then if \( 2\sum_{l=1}^d \tilde{J}_{n,l} = o(\sqrt{n}) \), this implies that \( \langle \Psi^T_j \Psi_j \rangle_{k,k} \approx n \) by the orthonormality of \( \psi_{j,k}(\cdot) \) and \( \sum_{l=1}^n |\psi_{j,k}(X_l)| \lesssim n \prod_{l=1}^d 2^{-j_l/2} \) with probability at least \( 1 - 2e^{-2^{\sum_{l=1}^d j_l}} \).
The lemma below gives the $L_2$-posterior contraction rate for spike-and-slab prior in non-parametric regression models. It shows in particular that there is an extra logarithmic factor in the rate, and is a reflection of the fact that separable selection rules (coefficient-wise spike-and-slab) will have at least a logarithmic penalty when trying to estimate $f$ adaptively under a global $L_2$-loss.

**Lemma 7.4.** Under the hierarchical spike-and-slab prior in (3.1), there exist constants $M, P_4 > 0$ such that uniformly in $0 < \alpha_i < \eta_i + 1, l = 1, \ldots, d$,

$$
\sup_{f_0 \in B_{2,\infty}(R)} E_0 \left( \|f - f_0\|_n + |\sigma^2 - \sigma_0^2| > M(n / \log n)^{-\alpha/(2\alpha + d)} \|Y\| \right) \leq n^{-P_4}.
$$

**Proof.** We will use the master theorem (see Theorem 3 of [7]) by constructing test function with exponential error probabilities, and verify that prior gives sufficient mass on Kullback-Leibler neighborhood around $(f_0, \sigma_0)$. Let $\epsilon_n \to 0$ and $\alpha n_2 \to \infty$. Define $J_{n,l}(\alpha)$ such that $2^{J_{n,l}(\alpha)} \asymp (\alpha / (n / \log n))^{\alpha/(2\alpha + d)}, l = 1, \ldots, d$. Form sieves $F_n := \{f : \theta_j / \epsilon_n = \theta_M, j \leq J_{n,l}(\alpha), k_l\}$, where $\theta_M := \sup_j \{\theta_j : \theta_j / \epsilon_n = \theta, j \leq J_{n,l}(\alpha)\}$, where the last inequality follows from Proposition 3.2 since $f_0 \in B^\alpha_{2,\infty}(R)$. In view of Lemma 6.4, $\|f - K_{J_n(\alpha)}(f_0)\|_n \leq \|f - f_0\|_n \leq \|f - K_{J_n(\alpha)}(f_0)\|_n + \|K_{J_n(\alpha)}(f_0) - f_0\|_n \leq \|f - K_{J_n(\alpha)}(f_0)\|_n + \sum_{l=1}^d 2^{-\alpha_{J_{n,l}(\alpha)}}$, where the last inequality follows from Proposition 3.2 since $2^{J_{n,l}(\alpha)} = o(n)$. Here the tilde in $\tilde{\theta}$ represents the truncated mother wavelet coefficients. We then conclude that there are constants $W_1, W_2 > 0$ such that for $f \in F_n$,

$$
W_1 \|\tilde{\theta} - \tilde{\theta}_0\| \leq \|f - f_0\|_n \leq W_2 \|\theta - \theta_0\| + \|\tilde{\theta} - \tilde{\theta}_0\| + (\log n / n)^{\alpha/(2\alpha + d)},
$$

by the definition of $J_{n,l}(\alpha)$. Let us define sieve slices $F^j_n := \{f \in F_n : j \epsilon_n < \|f - f_0\|_n + |\sigma^2 - \sigma_0^2| \leq (j + 1) \epsilon_n\}$ for any integer $j \geq M$. It follows from (7.4) above that

$$
F^j_n \subset \{\|\tilde{\theta} - \tilde{\theta}_0\| \leq (2/W_1) j \epsilon_n, |\sigma^2 - \sigma_0^2| \leq 2 j \epsilon_n\}.
$$

By calculating the covering number of the Euclidean space on the right hand side, we conclude that $F^j_n$ has a $\epsilon_n$-net of at most $e^{C \epsilon_n^2}$ points for some constant $C > 0$ if $2^{J_{n,l}(\alpha)} \leq \alpha n_2$. Then by Lemma 1 of [15], there exists a test $\phi_{n,j}$ with exponentially small error probabilities for testing $f = f_0$ against $f \in F^j_n$, by combining 3 tests corresponding to the cases where $|\sigma^2 - \sigma_0^2| \leq 1/2, \sigma^2 > 3\sigma_0^2/2$ and $\sigma^2 \leq \sigma_0^2/2$. Using the arguments outlined in the proof of Theorem 9 of [7], we conclude that $\phi_n = \sup_{j > M} \phi_{n,j}$ is a test with exponentially small Type I and II errors, thus fulfilling the testing requirement of the master theorem.

To characterize prior concentration, let $K(p, q) := \int p \log (p/q) d\mu$ be the Kullback-Leibler divergence and $V(p, q) := \int p \log (p/q) - K(p, q)^2 d\mu$, where $\mu$ is the Lebesgue measure. Define the Kullback-Leibler neighborhood $B_n(\epsilon_n) := \{(f, \sigma^2) : n^{-1} \sum_{i=1}^n K(p_{f_0,i}, p_{f,i}) \leq \epsilon_n^2, n^{-1} \sum_{i=1}^n V(p_{f_0,i}, p_{f,i}) \leq \epsilon_n^2 \}$ with $p_{g,i}$ being the density of $N[g(X_i), \sigma^2]$. After some
calculations, we have
\[
\frac{1}{n} \sum_{i=1}^{n} K(p_{f_0,i}, p_{f,i}) = \frac{1}{2} \log \left( \frac{\sigma^2}{\sigma_0^2} \right) - \frac{1}{2} \left( 1 - \frac{\sigma_0^2}{\sigma^2} \right) + \frac{\| f - f_0 \|^2}{2\sigma^2},
\]
\[
\frac{1}{n} \sum_{i=1}^{n} V(p_{f_0,i}, p_{f,i}) = \frac{1}{2} \left( 1 - \frac{\sigma_0^2}{\sigma^2} \right)^2 + \frac{\sigma_0^2 \| f - f_0 \|^2}{\sigma^4}.
\]

Hence, there are constants $W_3, \tilde{W}_3 > 0$ such that $B_n(\epsilon_n) \supset \{ \| f - f_0 \| \leq W_3 \epsilon_n, |\sigma^2 - \sigma_0^2| \leq \tilde{W}_3 \epsilon_n \}$. In view of (7.4), we have for any $\epsilon_n \geq (3W_3/W_3)(\log n/n)^{\alpha^*/(2\alpha^*+d)}$ that $B_n(\epsilon_n) \supset \{ \| \theta - \theta_0 \| \leq W_3 \epsilon_n/(3W_2), \| \tilde{\theta} - \tilde{\theta}_0 \| \leq W_3 \epsilon_n/(3W_2), |\sigma^2 - \sigma_0^2| \leq \tilde{W}_3 \epsilon_n \}$. Therefore by the assumed independence of the priors,
\[
\Pi[B_n(\epsilon_n)] \geq \Pi \left[ \| \theta - \theta_0 \| \leq \frac{W_3}{3W_2} \epsilon_n \right] \Pi \left[ \| \tilde{\theta} - \tilde{\theta}_0 \| \leq \frac{W_3}{3W_2} \epsilon_n \right] \Pi \left( |\sigma^2 - \sigma_0^2| \leq \tilde{W}_3 \epsilon_n \right). \tag{7.5}
\]

Since $\pi_\sigma$ is continuous and $\pi_\sigma(\cdot) > 0$ by assumption, we have
\[
\Pi(|\sigma^2 - \sigma_0^2| \leq \tilde{W}_3 \epsilon_n) \geq 2\tilde{W}_3 \epsilon_n \inf_{|u - \sigma_0^2| \leq \tilde{W}_3 \epsilon_n} \pi_\sigma(u) = 2\tilde{W}_3 \epsilon_n \pi_\sigma(\sigma_0^2)[1 + o(1)] \geq e^{-H_1 \log n},
\]
for some constant $H_1 > 0$, where the last inequality follows since $\epsilon_n \geq n^{-1/2}$ by assumption. For a set $A$ in some Euclidean space, we denote $\text{vol}(A)$ to be the volume of $A$ with respect to the Lebesgue measure. Let $\tilde{N} = \prod_{l=1}^{d} 2^{N_l}$. The first prior factor on the right hand side of (7.5) is
\[
\Pi \left[ \| \theta - \theta_0 \| \leq \frac{W_3}{3W_2} \epsilon_n \right] = \int_{\| \theta - \theta_0 \| \leq W_3 \epsilon_n/(3W_2)} \pi_\theta(d\theta) \pi_\sigma(d\sigma) \geq \frac{\pi_{\max} W_3}{3W_2} \epsilon_n \cdot \prod_{l=1}^{d} 2^{N_l} - \sum_{l=1}^{d} 2^{N_l} - \sum_{l=1}^{d} 2^{N_l} - \sum_{l=1}^{d} 2^{N_l} - \sum_{l=1}^{d} 2^{N_l}
\]
\[
\geq \pi_{\max} \text{vol}(\theta \in \mathbb{R}^{d}) \geq e^{-H_2 \log n},
\]
for some constant $H_2 > 0$. We lower bound the second factor in (7.5) by
\[
\Pi \left[ \| \tilde{\theta} - \tilde{\theta}_0 \| \leq W_3 \epsilon_n/(3W_2) \right] \geq \Pi \left[ \| \tilde{\theta} - \tilde{\theta}_0 \| \leq W_3 \epsilon_n/(3W_2) \right] \pi(\tilde{\theta}_0) \Pi(\tilde{\theta}_0),
\]
where $\pi(\tilde{\theta}_0) = \{ (j, k) : \theta_{j,k} \neq 0, \text{ if } j < J_{n,l}(\alpha), l = 1, \ldots, d \text{ and } \theta_{j,k} = 0 \text{ if for some } l = 1, \ldots, d, J_{n,l}(\alpha) \leq j \leq J_{n,l}(\alpha) - 1, \text{ with } 0 \leq k \leq 2^{j_{l}} - 1 \}$. Denote $K_n(\alpha) = \{ (j_1, \ldots, j_d) : 0 \leq j_1 \leq J_{n,1}(\alpha) - 1, l = 1, \ldots, d \}$. Recall that $n^{-\lambda} \leq \omega_{j,n} \leq \min \{ \prod_{l=1}^{d} 2^{-j_{l}(1+\mu_{l})}, 1/2 \}$. Using
the fact that $\log(1 - x) \geq -(2 \log 2)x$ for $0 \leq x \leq 0.5$, we have

$$
\Pi(\tilde{P}_n) = \exp \left\{ \sum_{j_1 = N_1}^{J_{n,1}(\alpha) - 1} \cdots \sum_{j_d = N_d}^{J_{n,d}(\alpha) - 1} 2^{\sum_{l=1}^d j_l} \log \omega_{j,n} + \sum_{j \in K_n(\alpha)^c} 2^{\sum_{l=1}^d j_l} \log \left(1 - \omega_{j,n}\right) \right\}
$$

$$
\geq \exp \left\{ -\lambda \log n \prod_{l=1}^d \sum_{j_l = N_l}^{J_{n,l}(\alpha) - 1} 2^{j_l - 2 \log 2} \sum_{j \in K_n(\alpha)^c} 2^{\sum_{l=1}^d j_l} \omega_{j,n} \right\} \tag{7.6}
$$

Define sets $Q_l$, $l = 1, \ldots, d$ where $Q_l$ can be $\{ j_l < J_{n,l}(\alpha) \}$ or $\{ j_l \geq J_{n,l}(\alpha) \}$, but with the constraint that not all $Q_l$’s are $\{ j_l < J_{n,l}(\alpha) \}$. Then the summation over $j \in K_n(\alpha)^c$ is such that $j$ takes on all $2^d - 1$ possible combinations of the $Q_l$’s, and each combination has the form

$$
\sum_{j_1 \in Q_1} \cdots \sum_{j_d \in Q_d} 2^{\sum_{l=1}^d j_l} \omega_{j,n} \leq \sum_{j_1 \in Q_1} \cdots \sum_{j_d \in Q_d} 2^{-\sum_{l=1}^d j_l} \mu_l.
$$

Among these $2^d - 1$ combinations, the configuration with one $Q_l = \{ j_l \geq J_{n,l}(\alpha) \}$ and the rest $Q_l = \{ j_l < J_{n,l}(\alpha) \}$, $l \neq i, l = 1, \ldots, d$ will dominate the sum, and they are exactly $d$ such configurations. Thus, the sum over $j \in K_n(\alpha)^c$ in (7.6) is bounded above up to some universal constant by

$$
\sum_{i=1}^d \sum_{j_i \geq J_{n,i}(\alpha)} 2^{-j_i \mu_i} \prod_{l \neq i} \sum_{j_l < J_{n,l}(\alpha)} 2^{-j_l \mu_l} \lesssim \sum_{i=1}^d 2^{-J_{n,i}(\alpha)/2},
$$

since $\mu_i > 1/2, l = 1, \ldots, d$. Hence, the expression in the exponential function in (7.6) is bounded below up to some constant multiple by $-\log n 2^{\sum_{l=1}^d J_{n,l}(\alpha)} - \sum_{i=1}^d 2^{-J_{n,i}(\alpha)/2}$. We then conclude that $\Pi(\tilde{P}_n) \geq e^{-H_1 \log n 2^{\sum_{l=1}^d J_{n,l}(\alpha)^c}}$ for some constant $H_3 > 0$.

By the assumption in (3.2) and denoting $\tilde{J} = \prod_{l=1}^d [2^{J_{n,l}(\alpha) - 2 N_l}]$,

$$
\Pi \left[ \|\tilde{\theta} - \tilde{\theta}_0\| \leq \frac{W_3}{3W_2} \epsilon_n \right] \tilde{P}_n = \int_{\|\tilde{\theta} - \tilde{\theta}_0\| \leq W_3 \epsilon_n / (3W_2)} \prod_{l=1}^d \prod_{j_l = N_l}^{J_{n,l}(\alpha) - 1} \prod_{k_l = 0}^{J_{n,l}(\alpha) - 1 - 2 j_l - 1} p(\theta_{j,k}) d\theta_{j,k}
$$

$$
\geq p_{\min} \frac{\tilde{J}}{\Gamma(\tilde{J}/2 + 1)} \frac{\pi^{\tilde{J}/2}}{3W_2^\tilde{J} \epsilon_n} \geq e^{-H_4 \log n 2^{\sum_{l=1}^d J_{n,l}(\alpha)^c}},
$$

for some constant $H_4 > 0$. By multiplying the lower bounds obtained for (7.5), it follows that $\Pi[B_n(\epsilon_n)] \geq e^{-C n_0^2}$ for some constant $C > 0$ only if $(\log n/n) 2^{\sum_{l=1}^d J_{n,l}(\alpha) - 2 N_l} \lesssim e_n^2$. This implies that $\epsilon_n \gtrsim (\log n/n)^{\alpha^*} / (2^{\alpha^* + d})$ for $2^{J_{n,l}(\alpha)} \asymp (n/\log n)^{\alpha^*} / (2^{\alpha^* + d})$, $l = 1, \ldots, d$.

It now remains to show that $E_0 \Pi(F_n | Y) \to 0$. By continuous embedding, this is equivalent to showing that $E_0 \Pi(\Theta_n^c | Y) \to 0$. Observe that $\Theta_n^c = \bigcup_{j \geq J_{n}(\alpha)} |\theta(j) \neq 0|$, with $|\theta(j) \neq 0|$
representing the set such that \( \theta_{j,k} \neq 0 \) for at least one \( k_l \) at some \( l = 1, \ldots, d \). Let \( A_j(m) = \{ \theta_j : \text{exactly } m \text{ among all } 2\sum_{i=1}^j m_i \text{ elements are not zero, while the rest are zeroes} \} \). Then \( [\theta(j) \neq 0] \) can be expressed as a union of \( A_j(m) \) across \( m = 1, \ldots, 2\sum_{i=1}^j m_i \) and resulting in \( \Pi(\Theta^c_n|Y) \leq \sum_{j \geq J_n(\alpha)} [\Pi(A_j(1)|Y) + \ldots + \Pi \left( A_j \left( \frac{2\sum_{i=1}^j m_i}{n} \right) \right)] \). After some calculations, it turns out that the first sum is \( O_{P_0}(e^{-C\log n}) \) while the rest of the terms are \( o_{P_0}(e^{-C\log n}) \) for some large enough constant \( C > 0 \). We then conclude that \( E_0 \Pi(F_n^c|Y) \lesssim e^{-C\log n} \to 0 \) as \( n \to \infty \).}

**Corollary 7.5.** As a consequence of Lemma 7.4 above, we have with posterior probability at least \( 1 - n^{-F_3} \) that

\[
\| \theta - \theta_0 \| \lesssim (n/ \log n)^{-\alpha*/(2\alpha^*+d)}, \quad \| \theta - \theta_0 \| \lesssim (n/ \log n)^{-\alpha*/(2\alpha^*+d)}. \tag{7.7}
\]

**Proof.** If \( f \) has wavelet expansion as in (3.1) at resolution \( J_n \), then by the property of \( L_2 \)-projection and Lemma 7.2, we have \( \| f - f_0 \|_n \geq \| f - K_{J_n}(f_0) \|_n \gtrsim (\| \theta - \theta_0 \|^2 + \| \theta - \theta_0 \|^2)^{1/2} \) since \( 2\sum_{i=1}^j m_i = o(n) \) by assumption. The result follows by applying Lemma 7.4. \( \square \)

As discussed earlier in Section 6.2 and also in the proofs of Lemmas 6.1 and 6.2, we need \( W_n \) as defined in (6.17) to hold with posterior probability tending to 1. This then allows us to bound posterior probabilities under the regression model with posterior under the sequence white noise model. To accomplish this, we need to control simultaneously the discrete approximation error of Lemma 7.2, the truncation error of Proposition 3.2 and the model stochastic error of (1.1).

**Lemma 7.6.** We have uniformly over \( \alpha \in A \) with \( f_0 \in B_{\infty,\infty}^\alpha(R), R > 0 \) that

\[
E_0 \Pi \left( |\tilde{\beta}_n(\Theta)| > \tau_n \sqrt{n \log n} \right) \leq n^{-F_4},
\]

for \( \tau_n \to 0 \) given in (7.8) below, \( P_4 \) the same as in Lemma 7.4, and for any \( N_1 \leq j_1 \leq J_{n,l} - 1, 0 \leq k_l \leq 2^{j_1} - 1, l = 1, \ldots, d \).

**Proof.** By the triangle inequality, \( |\tilde{\beta}_n(\Theta)| \) is bounded above by

\[
|\langle \Psi_j^T \Psi_j \rangle_{k,-k}(\theta_j - \theta_0)_{j,-k}| + \sum_{\alpha \neq j} \|\langle \Psi_j^T \Psi_j \rangle_{k,-k} \theta_a - \theta_0 \rangle \| + \|\langle \Psi_j^T B \rangle_{k,-k} \theta - \theta_0 \| + \|\langle \xi^T \Psi_j \rangle_{j,k} \|
\]

Let \( \epsilon_n = (n/ \log n)^{-\alpha*/(2\alpha^*+d)} \). By the Cauchy-Schwartz inequality, the first term on the right hand side is bounded by \( \|\langle \Psi_j^T \Psi_j \rangle_{k,-k} \theta - \theta_0 \| \). By Lemma 7.2, \( \|\langle \Psi_j^T \Psi_j \rangle_{k,-k} \| \lesssim \prod_{i=1}^d 2^{3j_i/2} \). Then by Corollary 7.5, the first term is \( O_{P_4}(2^{3\sum_{i=1}^d j_i/2} \epsilon_n) \). Similarly, we can bound the third term using the same lemma and corollary by \( \|\langle \Psi_j^T B \rangle_{k,-k} \theta - \theta_0 \| \leq C_3 \|\theta - \theta_0 \| \prod_{i=1}^d 2^{N_j/2} = O_{P_4}(2^{3\sum_{i=1}^d j_i \epsilon_n}) \).

To bound the last bias term, observe that \( \|\langle \xi^T \Psi_j \rangle_{j,k} \| \leq \|\xi\|_{\infty} \sum_{u=1}^d |\psi_{j,k}(X_u)| \). By Proposition 3.2 with \( r = 0 \), we have that \( \|\xi\|_{\infty} \lesssim \sum_{l=1}^d 2^{-\alpha_l j_{l,1}} \) for \( f_0 \in B_{\infty,\infty}^\alpha(R) \). By (7.3) in
Lemma 7.2, \( \sum_{u=1}^{n} |\psi_{j,k}(X_u)| \lesssim n \prod_{l=1}^{d} 2^{-j_l/2} \) since \( 2^{\sum_{l=1}^{d} j_l} = o(n) \). We then conclude that 
\[ |(\xi_j^{T} \Psi_j)_{k}| \lesssim n \prod_{l=1}^{d} 2^{-j_l/2} \sum_{l=1}^{d} 2^{-\alpha_l J_{n,l}}. \]

It now remains to bound the second term. By another application of the Cauchy-Schwartz inequality, this term is bounded above by \( \|\theta - \theta_0\| \sum_{(a,b) \neq (j,i)} \|[(\Psi_j)^T \Psi^b_{a}]_{k,\cdot}\| \). Then by Lemma 7.2, 
\[ \sum_{(a,b) \neq (j,i)} \|[(\Psi_j)^T \Psi^b_{a}]_{k,\cdot}\| \lesssim \sum_{a \neq j} \prod_{l=1}^{d} 2^{a_l+j_l/2}. \]

Define sets \( T_l, l = 1, \ldots, d \) such that \( T_l \) can be \( \{a_l = j_l\} \) or \( \{a_l \neq j_l\} \), but with the constraint that not all \( T_l \)'s are \( \{a_l = j_l\} \). Then the sum on the right hand side above consists of \( 2^d - 1 \) terms of the form 
\[ \sum_{a_l \in T_l} \prod_{l=1}^{d} 2^{a_l+j_l/2} = \prod_{l=1}^{d} \sum_{a_l \in T_l} 2^{a_l+j_l/2}. \]

If \( T_l = \{a_l = j_l\} \), then \( \sum_{a_l = j_l} 2^{a_l+j_l/2} = 2^{j_l+1/2} \); and if \( T_l = \{a_l \neq j_l\} \), then \( \sum_{a_l \neq j_l} 2^{a_l+j_l/2} \lesssim 2^{J_{n,l}+j_l/2} \). It follows by Corollary 7.5 that the second term in the upper bound of \( |\tilde{\beta}_n(\Theta)| \) above is \( O_P(2^{\sum_{j=1}^{d}(j_l/2+J_{n,l})} \epsilon_n) \). Combining the bounds obtained, we have for any \( N_l \leq j_l \leq J_{n,l} - 1, l = 1, \ldots, d, \)
\[ |\tilde{\beta}_n(\Theta)| = O_P \left( 2^{\sum_{j=1}^{d}(j_l/2+J_{n,l})} \left( n/ \log n \right)^{-\alpha^*/(2\alpha^*+d)} \right) + O \left( n \prod_{l=1}^{d} 2^{-j_l/2} \sum_{l=1}^{d} 2^{-\alpha_l J_{n,l}} \right). \]

To optimize the upper bound with respect to \( j_l, l = 1, \ldots, d \), observe that \( 2^{\sum_{j=1}^{d} j_l} \asymp n^{2-\sum_{l=1}^{d} J_{n,l}} \epsilon_n^{-1} \sum_{l=1}^{d} 2^{-\alpha_l J_{n,l}} \) balances the two antagonistic terms and yields
\[ |\tilde{\beta}_n(\Theta)| = O_P \left( \sqrt{n} 2^{\sum_{j=1}^{d} J_{n,l}/2} \epsilon_n^{1/2} \sum_{l=1}^{d} 2^{-\alpha_l J_{n,l}/2} \right). \]

To proceed, we set \( 2^{J_{n,l}} = 2^{J_{l}} \), \( l = 1, \ldots, d \), and it follows that \( 2^{\sum_{j=1}^{d} J_{n,l}} = 2^{dJ_l/\alpha_l} \) and 
\[ \sum_{l=1}^{d} 2^{-\alpha_l J_{n,l}/2} = 2^{-J_{n,l}/2} \]. By letting \( 2^{\sum_{j=1}^{d} J_{n,l}} = \sqrt{n} \log n \), the above is with posterior probability tending to 1 bounded above by
\[ d \sqrt{n} 2^{\sum_{j=1}^{d} (j_l/2 + J_{n,l})} \epsilon_n^{1/2} \lesssim n^{-2\alpha^* - d(\alpha^* + d)} (\log n)^{2\alpha^* - d(\alpha^* + d)} \frac{1}{4} \sqrt{n} \log n =: \tau_n \sqrt{n} \log n. \quad (7.8) \]

Uniformly over \( \alpha \in \mathbb{A} \), we have \( 1/d + 1/(2\alpha^*) > 1/\alpha_l, l = 1, \ldots, d \), and this implies that \( \alpha^* > d/2 \) by summing both sides. Thus, \( \tau_n \to 0 \) as \( n \to \infty \). \( \square \)

**Remark 7.7.** In the random case, we have instead
\[ |\tilde{\beta}_n(\Theta)| = O_P \left( \sqrt{n} 2^{\sum_{j=1}^{d} (j_l/2 + J_{n,l})} (\log n/n)^{\alpha^*/(2\alpha^*+d)} + n \prod_{l=1}^{d} 2^{-j_l/2} \sum_{l=1}^{d} 2^{-\alpha_l J_{n,l}} \right) \]
\[ = O_P \left( \sqrt{n} n^{1/4d} 2^{\sum_{j=1}^{d} J_{n,l}/2} \epsilon_n^{1/2} \sum_{l=1}^{d} 2^{-\alpha_l J_{n,l}/2} \right), \]

35
by taking $2\sum_{l=1}^{d} J_{l} = \sqrt{n} \sum_{l=1}^{d} 2^{-\alpha_{l} J_{l}/2} \sum_{j=1}^{J_{l}} \epsilon_{n,j}^{-1}$. Then let $2J_{n,l} = 2^{1/\alpha_{l}}$, $l = 1, \ldots, d$, and take $2\sum_{l=1}^{d} J_{n,l} = (n/\log n)^{1/4}$, the above is with posterior probability tending to 1 bounded above by

$$d\sqrt{n} n^{1/4} 2\sum_{l=1}^{d} J_{n,l} [1/2 - \alpha_{l}^*/(2d)] \epsilon_{n}^{1/2} \lesssim (n/\log n)^{-3d(2\alpha_{l}^* + d)/(2d(2\alpha_{l}^* + d))} (\log n)^{-1/2} \sqrt{n \log n} =: \tau_{n} \sqrt{n \log n}.$$  

In view of Remark 4.3, we take $\alpha$ uniformly over $1/d + 1/(2\alpha_{l}^*) > 2/\alpha_{l}$, $l = 1, \ldots, d$, and this implies that $\alpha_{l}^* > 3d/2$ by summing both sides. Thus, $\tau_{n} \to 0$ as $n \to \infty$.

### References


