Adaptive Bayesian density estimation in sup-norm

ZACHARIE NAULET

Université Paris-Saclay
Laboratoire de mathématiques d’Orsay
91405, Orsay, France
E-mail: zacharie.naulet@universite-paris-saclay.fr

We investigate the problem of deriving adaptive posterior rates of contraction on $L_\infty$ balls in density estimation. Although it is known that log-density priors can achieve optimal rates when the true density is sufficiently smooth, adaptive rates were still to be proven. Here we establish that the so-called spike-and-slab prior can achieve adaptive and optimal posterior contraction rates. Along the way, we prove a generic $L_\infty$ contraction result for log-density priors with independent wavelet coefficients. Interestingly, our approach is different from previous works on $L_\infty$ contraction and is reminiscent of the classical test-based approach used in Bayesian nonparametrics. Moreover, we require no lower bound on the smoothness of the true density, albeit the rates are deteriorated by an extra $\log(n)$ factor in the case of low smoothness.

MSC 2010 subject classifications: Primary 62G07; secondary 62G20.
Keywords: Bayesian density estimation, supremum norm, adaptation.

1. Introduction

We consider the problem of estimating a density $p$ with respect to Lebesgue measure on $[0,1]$ given $n$ independent and identically distributed samples $X_n := (X_1, \ldots, X_n)$ from the corresponding distribution $P$. We adopt the Bayesian paradigm and put a joint distribution on the log-density and the observations.

Over the decades, there has been a growing interest for the understanding of the frequentist behaviour of posterior distributions initiated by the seminal papers of Schwartz (1965); Barron et al. (1999); Ghosal et al. (2000). In particular Ghosal et al. (2000) states generic sufficient conditions for obtaining rates of concentration of the posterior distribution near the true model in some distance. The approach relies on the well-known existence of exponentially powerful test functions. The existence of such tests depends on the distance considered, and is guaranteed for the $L_1$ or Hellinger distance between densities, and also for the $L_2$ metric under supplementary assumptions. It is, however, now well understood that the test approach fails to give optimal rates when the risk is measured with respect to the $L_\infty$ distance, see Castillo (2014); Hoffmann et al. (2015); Yoo et al. (2017).
The failure of the classical approach for $L_\infty$ rates is unfortunate because one has in general a better intuition of the shape of $L_\infty$ balls rather than Hellinger balls, making the $L_\infty$ risk a more natural distance for evaluating performance of estimators. From a frequentist point of view, density estimation in sup-norm is now well understood. Minimax lower bounds can be found in Hasminskii (1979) while upper bounds can be found for instance in Ibragimov and Hasminskii (1980); Goldenshluger and Lepski (2014).

For Bayesian procedures, concentration on $L_\infty$ balls is much less understood. For the non-adaptive case, the first result goes back to Giné and Nickl (2011) where optimal rates are obtained in white-noise regression using conjugacy arguments. In the same paper, the authors obtained (possibly adaptive) rates for density estimation in sup-norm using a testing approach, but failed to achieve optimality. Using conjugacy arguments, Yoo and Ghosal (2016) also obtain non-adaptive but optimal rates for estimating a regression function. Scricciolo (2014) adapts the techniques of Giné and Nickl (2011) to obtain optimal rates when the true density is analytic. The first non-adaptive optimal result in density estimation for non ultra-smooth densities is to be credited to Castillo (2014), where the author uses techniques based on semi-parametric Bernstein–von Mises theorems. His approach, however, requires a minimal smoothness to be applicable. Recently, Castillo (2017) obtained non-adaptive but optimal rates for density estimation in sup-norm using Pólya trees prior, with no lower bound required on the smoothness.

The existence of adaptive and optimal results is, to our knowledge, even more limited. The first successful result is in Hoffmann et al. (2015) where the authors get adaptive optimal rates in $L_\infty$ norm for white-noise regression using a spike-and-slab prior. They also obtain adaptive and optimal rates in density estimation, though their result is rather an existence result as their abstract sieve prior is not computable. More recently, Yoo et al. (2017) obtained adaptive optimal rates in $L_\infty$ norm for estimating a regression function, using a white-noise approximation of the likelihood to adapt the techniques developed in Hoffmann et al. (2015). Since the first version of the current paper, Castillo and Mismer (2019) have introduced spike-and-slab Pólya trees and built upon the results of Castillo (2017) to obtain adaptive contraction rates, though the arguments we use are different.

In density estimation, it is not obvious to proceed as Yoo et al. (2017) and reduce the problem to white-noise regression, although it is known those models are equivalent (in the Le Cam sense) under certain assumptions. Here, instead, we propose a different approach. We obtain in Section 3 a general contraction result for log-density priors with independent wavelet coefficients. This result is the building block of our main Theorem 2.2 about the spike-and-slab log-density prior. The posterior spike-and-slab is known to be the Bayesian analogue of hard thresholding (Hàrdle et al., 2000), as already noticed by Hoffmann et al. (2015); Yoo et al. (2017). As such, it constitutes the prototypical example of model for which we expect adaptive and optimal $L_\infty$ contraction. Unlike Hoffmann et al. (2015); Yoo et al. (2017), however, the present paper does not exploit the thresholding property of the spike-and-slab posterior to establish the posterior support of the wavelet coefficients, but uses a more classical approach.
In the case of the spike-and-slab prior, we show that our method can be applied to obtain minimax optimal and adaptive posterior contraction. More precisely, we show that if \( L := \log p \in B_{\infty,\infty}^s[0,1] \), where \( B_{\infty,\infty}^s[0,1] \) denotes the Hölder-Zygmund space with smoothness \( s > 1/2 \) (see Definition 2.1), then there exists \( M > 0 \) such that as \( n \to \infty \),

\[
E_L \Pi(L' : \|L' - L\|_{\infty} \leq M \varepsilon_n^*(s) | X_n) = 1 + o(1),
\]

where \( \varepsilon_n^*(s) \) is the minimax rate over bounded balls in \( B_{\infty,\infty}^s[0,1] \) under \( L_\infty \) loss (Donoho et al., 1996)

\[
\varepsilon_n^*(s) := \left( \frac{\log n}{n} \right)^{\frac{s}{2s+1}}.
\]

Interestingly, our method can be applied to obtain adaptive rates in the region \( 0 < s \leq 1/2 \), which to the best of our knowledge is the first result of this type in the Bayesian literature for methods not relying on conjugacy arguments. The rates we obtain in this region are, however, slightly deteriorated by a factor \( \log(n) \).

In contrast with previous results in \( L_\infty \) loss, the approach used in this paper is somewhat less specific and uses the same kind of arguments as for the master theorem of Bayesian nonparametrics (Ghosal et al., 2000; Ghosal and van der Vaart, 2007b,a). In particular, it relies on the existence of suitable test functions and proving the prior positivity of some neighborhoods, in apparent contradiction with the folk wisdom that no test for the \( L_\infty \) loss has enough power to obtain optimal rates Hoffmann et al. (2015); Yoo et al. (2017). This contradiction is only apparent, as here we require to test only very specific kind of alternatives, and exponentially consistent tests are not needed. Although hard to generalize, we believe the present paper shows that the traditional approach of (Ghosal et al., 2000; Ghosal and van der Vaart, 2007b,a) is more powerful than we believed, giving hope for the existence of general contraction results in strong norms.

This article comes with a supplementary material, which contain additional proofs and various classical results about the spike-and-slab log-density prior. We adopt the convention that every section, subsection, theorem, etc. of the supplemental has label prefixed by S and is cited in cyan. References to the main document are cited in blue with no prefix.

2. Exponentiated random wavelet series

2.1. Log-density priors

We use the \( S \)-regular, orthogonal, boundary corrected wavelets of Cohen et al. (1993), referred to as the CDV basis. We denote this basis by \( \{\varphi_{j,k} : (j, k) \in \mathcal{V}\} \), where \( \mathcal{V} \subseteq \mathbb{Z}_2^2 \), and refer to Cohen et al. (1993); Giné and Nickl (2016); Castillo (2014) for details. Each index \((j, k)\) is a pair where \( j \geq 0 \) is the wavelet level and \( k \) the location index. The CDV basis is an orthogonal basis for \( L_2[0,1] \) equipped with the Lebesgue measure. We endow \( L_2[0,1] \) with the inner product \( \langle f, g \rangle := \int_{[0,1]} fg \). If \( f \in B_{\infty,\infty}^s[0,1] \) for some \( s > 0 \), then
the wavelet series \( \sum_{(j,k) \in V} \langle f, \varphi_{j,k} \rangle \varphi_{j,k} \) converges uniformly to \( f \). Conversely, for a given \( (\theta_{j,k})_{(j,k) \in V} \), the series \( \sum_{(j,k) \in V} \theta_{j,k} \varphi_{j,k} \) converges uniformly if \( \theta \) is in
\[
\Theta \subseteq \left\{ \theta \in \mathbb{R}^V : \sup_{x \in [0,1]} \sum_{(j,k) \in V} |\theta_{j,k}| |\varphi_{j,k}(x)| < \infty \right\}.
\]
Thus, we shall consider prior distributions over the space \( \Theta \). Such prior distribution induces a prior distribution on the space of densities on \([0,1]\) (with respect to the Lebesgue measure) through the mapping \( \theta \mapsto p_\theta \) such that
\[
p_\theta(x) := \frac{\exp\{\sum_{(j,k) \in V} \theta_{j,k} \varphi_{j,k}(x)\}}{\int_{[0,1]} \exp\{\sum_{(j,k) \in V} \theta_{j,k} \varphi_{j,k}\}}, \quad x \in [0,1].
\]
(2.1)
The coefficients \( \theta \) in equation (2.1) are immediately seen to be the basis coefficients of an unnormalized version of the log-density \( \log p_\theta \).

2.2. Spike-and-Slab log-density priors

To obtain adaptive and optimal rates of contraction, we consider the so-called spike-and-slab prior distribution over \( \Theta \) (Mitchell and Beauchamp, 1988). For some weights \((\omega_1, \omega_2, \ldots) \in [0,1]^N\)
\[
\theta_{j,k} \overset{\text{ind}}{\sim} \begin{cases} 
(1 - \omega_j) \delta_0 + \omega_j Q_j(\cdot) & \text{if } 0 \leq j \leq \frac{\log(n)}{\log(2)}, \\
\delta_0 & \text{if } j > \frac{\log(n)}{\log(2)}.
\end{cases}
\]
Here \( \delta_0 \) is the point mass at zero and \( Q_j \) are probability distributions on \( \mathbb{R} \). We assume \( Q_j \) have densities \( q_j \) such that for some \( 0 < s_0 < 1/2 \) and for some density \( f \), we have
\[
q_j(x) := 2^{j(s_0+1/2)} f(2^{j(s_0+1/2)} x) \text{ for every } j \geq 0.
\]
We write \( F \) the probability distribution with density \( f \).

In order to ensure that the prior puts enough mass on neighborhoods of the true log-density \( L \), we also assume that for all \( G > 0 \) there is \( g > 0 \) such that
\[
\inf_{x \in [-G,G]} f(x) \geq g.
\]
(2.4)
Adaptive Bayesian density estimation in sup-norm

We note that the assumptions of equations (2.2) and (2.4) are classical in the literature for rates of contraction in supremum loss. The equation (2.3) is however very strong, but guarantees that a priori $L_\theta$ has wavelet coefficients of reasonable magnitude, which guarantees that the posterior concentrates on nice neighborhoods of $L$, see in particular Section S4. As an example of distribution $F$ that satisfy the requirements of equations (2.3) and (2.4), one could take the distribution of the random variable $\log(Z)$ where $Z$ has an inverse-Gaussian distribution (or any distribution on $\mathbb{R}_+$ with an exponential behaviour both near 0 and $\infty$).

2.3. Adaptation and optimality under supremum loss

This paper considers adaptation over bounded Hölder-Zygmund balls, which we define below. First, we give a precise definition for the Hölder-Zygmund spaces of smoothness.

**Definition 2.1** (Hölder-Zygmund spaces). For any $s > 0$, the Hölder-Zygmund space $B^s_{\infty,\infty}[0,1]$ is the space of uniformly continuous functions $f : [0,1] \rightarrow \mathbb{R}$ such that $\|f\|_{\infty,\infty,s} < \infty$, where $\|f\|_{\infty,\infty,s} := \sup_{(j,k) \in \mathcal{V}} 2^{j(s+1/2)}|\langle f, \varphi_{j,k} \rangle|$.

We are now in position to define the bounded Hölder-Zygmund ball of log-densities with radius $R > 0$ and smoothness $s > 0$

$$\Sigma(R,s) := \left\{ L \in B^s_{\infty,\infty}[0,1] : \|L\|_{\infty,\infty,s} \leq R, \int_{[0,1]} \exp(L) = 1 \right\}.$$

We prove in the supplemental that spike-and-slab log-density priors satisfying equations (2.2) to (2.4) achieve adaptive and nearly optimal posterior contraction rates $\varepsilon^*_n(s)$ over $\Sigma(R,s)$ under Hellinger loss. In particular, the next theorem is proven in Section S3.

**Theorem 2.1.** Let $\Pi$ be the spike-and-slab log-density prior satisfying equations (2.2) to (2.4) and let $H(P,Q)$ denote the Hellinger distance between probability distributions $P$ and $Q$. Then for all $0 < s_0 \leq s \leq S$ and for all $R > 0$ there exists a constant $M > 0$ such that

$$\lim_{n \rightarrow \infty} \sup_{L \in \Sigma(R,s)} \mathbb{E}_L \Pi(\theta : H(P_\theta, P_L) > M \varepsilon^*_n(s) | X_n) = 0.$$

The main theorem of the paper establishes that spike-and-slab log-density priors can achieve the optimal posterior contraction rates if we consider the $L_\infty$ loss. The rate is optimal when $s > 1/2$, with a slight deterioration in the region of small smoothness.

**Theorem 2.2.** Let $\Pi$ be the spike-and-slab log-density prior satisfying equations (2.2) to (2.4). Also let $\varepsilon^*_n(s) := \varepsilon^*_n(s)$ if $s > 1/2$ or $\varepsilon^*_n(s) := \log(n)\varepsilon^*_n(s)$ if $0 < s \leq 1/2$. Then for all $0 < s_0 \leq s \leq S$ and for all $R > 0$ there exists a constant $M > 0$ such that

$$\lim_{n \rightarrow \infty} \sup_{L \in \Sigma(R,s)} \mathbb{E}_L \Pi(\theta : \|L_\theta - L\|_\infty > M \varepsilon^*_n(s) | X_n) = 0.$$
We emphasize that the Theorem 2.2 also entails posterior concentration of $\|p_\theta - p_L\|_\infty$ at rate $\tilde{\epsilon}_n(s)$, and thus because $L \in \Sigma(R,s)$ implies that $e^{-R} \leq p_L \leq e^R$. Hence, $\|p_\theta - p_L\|_\infty$ and $\|L_\theta - L\|_\infty$ are equivalent distances when the latter is small enough. This assumption that $L$ belongs to a H"older ball of smoothness is stronger than the classical frequentist assumption that only $p_L$ does. In particular, we see that it entails that $p_L$ is bounded from below, which is of great help in the proofs. Of course, this begs the question of what can be said when $p_L$ is smooth but not bounded from below, which is outside the scope of this paper. We note that assuming smoothness on $L$ rather than $p_L$ is classical in the Bayesian community (see Castillo, 2014).

The rest of the paper is organized as follows. In the Section 3 we establish the main notations and give the main ideas behind the proof of the Theorem 2.2. In particular, we give guidelines for the proof and state a central contraction result in Theorem 3.1 which is at the core of the proof of Theorem 2.2. Then, in Section 4 we discuss the main implications of our results. Finally, proofs are given in Sections 5 to 7, respectively for the Theorem 2.2, the Theorem 3.1, and for auxiliary results. Many secondary proofs are deferred to the Section S5 of the supplemental.

3. Heuristic and main ideas behind the proof of Theorem 2.2

3.1. Notations

We let $\mathbb{N} := \{1,2,\ldots\}$ denote the set of natural numbers, and we let $\mathbb{Z}_+ := \{0,1,\ldots\}$ denote the set of positive integers. The symbols $\lesssim$ and $\gtrsim$ are used to denote inequalities up to generic constants. If $a \lesssim b$ and $b \lesssim a$, we write $a \asymp b$. For two sequences $(a_n)_{n \in \mathbb{Z}_+}$ and $(b_n)_{n \in \mathbb{Z}_+}$, the notation $a_n = o(b_n)$ means $\lim \sup_{n \to \infty} |a_n/b_n| = 0$, and $a_n = O(b_n)$ means $\lim \sup_{n \to \infty} |a_n/b_n| = C$ for some $C \geq 0$. For $a,b \in \mathbb{R}$, we let $a \wedge b$ denote the minimum of $a$ and $b$, and $a \vee b$ stands for the maximum.

Densities are understood with respect to the Lebesgue measure. Lower-case notations $p,q,\ldots$ are used to denote densities, while upper-case $P,Q,\ldots$ denote the corresponding distributions. Given a log-density $L$ on $[0,1]$, we write $p_L := \exp\{L\}$ the corresponding density and $P_L$ the corresponding distribution. When $L = L_\theta$ for some $\theta \in \Theta$, we abbreviate $P_\theta$ for $P_{L_\theta}$, etc.

We see $X_n = (X_1,\ldots,X_n)$ as the beginning of an infinite sequence $X_\infty = (X_1,X_2,\ldots)$ defined on a measurable space $(\Omega,\mathcal{A})$ and such that under $L$, the variables $X_1,X_2,\ldots$ are independent and identically distributed (iid) with distribution $P_L$. We write $\mathbb{P}_L$ the distribution of $X_\infty$, and we write indistinctly $\mathbb{E}_L$ the expectation under $\mathbb{P}_L$ or under $P_L$. We write $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ the empirical measure of $X_n$.

We use the standard definitions for the $L_p$ spaces of (equivalence classes) of functions with finite $\| \cdot \|_p$ norm, with $\|f\|_p^p := \int |f|^p$ if $1 \leq p < \infty$, and $\|f\|_\infty := \text{ess sup}_x |f(x)|$. 
We will also make use of the Hellinger distance between two probability distributions $P$ and $Q$ having respective densities $p$ and $q$, defined as $H(P, Q) := \frac{1}{\sqrt{2}} \|\sqrt{p} - \sqrt{q}\|_2$.

### 3.2. Change of parameterization

For some integer $J_0$ to be chosen sufficiently large, we define $B_0 := \{(j, k) \in V : j \leq J_0\}$. The indices in $B_0$ correspond to small scales wavelets and will require special cares. To ease the proof, it is convenient to relabel the wavelets with indices not in $B_0$. We let $\psi : N \to V \setminus B_0$ be the bijection corresponding to the lexicographical reordering of the index set $V \setminus B_0$, i.e. writing $\psi(m) = (j, k)$ and $\psi(m') = (j', k')$

$$m \leq m' \iff (j < j') \text{ or } (j = j' \text{ and } k \leq k').$$

For all $m \geq 1$ we write $J_m := \psi(m)$ the scale-index of the wavelet $\varphi_{\psi(m)}$. By construction $J_1 = J_0 + 1$ and $J_1 \leq J_2 \leq \ldots$. For proofs, it is also convenient to define $B_m := \{\psi(m)\}$ for all $m \geq 1$.

Given this re-indexing of the wavelets, we are now in position to define a change of parameterization which is convenient for proofs. We pick an arbitrary reference log-density $L \in \Sigma(R, s)$, and we establish the concentration under $L$ by taking care that the results are uniform over $\Sigma(R, s)$. Given $L$, we let $\theta_{j,k}^L := \langle L, \varphi_{j,k} \rangle$, and we define

$$F_{m}^\theta := \sum_{(j,k) \in B_m} (\theta_{j,k} - \theta_{j,k}^L)(\varphi_{j,k} - E_{L}[\varphi_{j,k}]).$$

Clearly $L_\theta - L$ can be written uniquely in term of the $(F_{m}^\theta)_{m \geq 0}$, so that we might as well consider $(F_{m}^\theta)_{m \geq 0}$ as the parameter of the model. For each $m \geq 0$ we will write $F_m := \text{span}\{\varphi_{j,k} - E_{L}[\varphi_{j,k}] : (j, k) \in B_m\}$, and $\mathcal{F} := \mathcal{F}_1 \times \mathcal{F}_2 \times \ldots$ the infinite cartesian product of these spaces. We also let $\mathcal{F}_0 := \{F \in \mathcal{F} : E_{L}[\exp\{\sum_{m \geq 0} F_m\}] < \infty\}$ denote the subset of proper parameters. Then, we now parameterize the model by $B \in \mathcal{F}_0$. Using the constraint that $E_{L}[\exp\{L_{F} - L\}] = 1$, we determine that the log-likelihood of the model $F \in \mathcal{F}_0$ is given by

$$L_{F} - L = \sum_{m \geq 0} F_m - \log E_{L}[\exp\{\sum_{m \geq 0} F_m\}]. \quad (3.1)$$

### 3.3. Guidelines for the proof of Theorem 2.2 and intermediate contraction results

Here we present the main ideas behind the proof of Theorem 2.2 and the main intermediate results that are used in the proof. First of all, it is convenient to assume that the posterior concentrates on nice neighborhoods of $L$. We will prove that the posterior concentrates on the set

$$\mathcal{C}_* := \{F \in \mathcal{F} : \sup_{x \in [0, 1]} \sum_{m \geq 0} |F_m(x)| \leq \delta\},$$
where $0 < \delta \leq 1$ is a constant to be chosen sufficiently small. Once it has been shown the posterior is concentrated on $\mathcal{C}_s$, the analysis of the log-likelihood difference $L_F - L$ is easier. Posterior contraction on $\mathcal{C}_s$ can be obtained by the classical machinery à la Ghosal and van der Vaart (Ghosal et al., 2000; Ghosal and van der Vaart, 2007a,b) and is essentially a corollary of the Theorem 2.1. It can be done using similar arguments as those already found in Castillo (2014); Rivoirard and Rousseau (2012), as we do in Section S5.1 to prove the following lemma.

**Lemma 3.1.** Let $\Pi$ be the prior described in Section 2.2. Then for all $0 < s_0 \leq s \leq S$, all $R > 0$, and all $\delta > 0$,

$$\lim_{n \to \infty} \sup_{L \in \Sigma(R,s)} \mathbb{E}_L(\mathcal{C}_s^c | X_n) = 0.$$  

To obtain $L_\infty$ rates, the goal is to relate the distance $\|L_F - L\|_\infty$ to the parameter $F$. In particular, we shall seek to relate $\|L_F - L\|_\infty$ to $\{\|F_m\|_2 : m \geq 0\}$, which is motivated by the fact that $\{\|F_m\|_2 : m \geq 0\}$ essentially drives the behaviour of the posterior distributions. The following lemma serves this purpose.

**Lemma 3.2.** Let $F \in \mathcal{C}_s$. Then, there exists a universal constant $C > 0$ such that for all choice of $J_0$ we have

$$\|L_F - L\|_\infty \leq C \sum_{j \geq J_0} 2^{j/2} \sup_{m : J_m = j} \|F_m\|_2.$$  

In view of Lemmas 3.1 and 3.2, to prove the Theorem 2.2 it is enough to prove that the posterior concentrates on a set where the rhs in Lemma 3.2 is smaller than a multiple of $\varepsilon_n^*(s)$. Our strategy is to build a partition of $\mathcal{F}$, where on each part we have a fine control of $\{\|F_m\|_2 : m \geq 0\}$. We build the partition $(\mathcal{S}_I)_{I \subseteq \mathbb{Z}_+}$, such that for every $I \subseteq \mathbb{Z}_+$,

$$\mathcal{S}_I := \{F \in \mathcal{F} : m \in I \implies \|F_m\|_2 > H_I(m), \ m \notin I \implies \|F_m\|_2 \leq H_I(m)\},$$

where we choose $H_m(I)$ as follows. We let $\Gamma, \gamma > 0$ be constants to be determined, and we define the optimal truncation level $J_n \equiv J_n(s)$ as the only integer satisfying

$$\begin{cases}
\gamma 2^{-(j_n + 1)(s + 1/2)} < \Gamma \sqrt{\log(n)/n} \leq \gamma 2^{j_n(s + 1/2)} & \text{if } s > 1/2, \\
\gamma 2^{-(j_n + 1)(s + 1/2)} < \Gamma 2^{-j_n/2 \varepsilon_n^*(s)} \leq \gamma 2^{-j_n(s + 1/2)} & \text{if } 0 < s \leq 1/2.
\end{cases} \tag{3.2}$$

Then, for $\xi > 1$ also to be chosen accordingly,

$$H_I(m) := \begin{cases}
\Gamma \xi^{-1} \rho_m \text{ if } J_m \leq j_n, \\
\gamma 2^{-(j_m + 1)(s + 1/2)} \text{ if } J_m > j_n,
\end{cases} \quad \rho_m := \begin{cases}
\sqrt{\log(n)/n} & \text{if } s > 1/2, \\
2^{-j_m/2 \varepsilon_n^*(s)} & \text{if } 0 < s \leq 1/2.
\end{cases}$$

At this point, it might look obscure why the definition of $H_I(m)$ differs according to whether $s > 1/2$ or not, and also according to whether $0 \in I$ or not. The subtle reason
Adaptive Bayesian density estimation in sup-norm

of this choice will be found when proving the Theorem 3.1. In fact, we will require to control some covariance terms involving $F_m$ and $F_m'$ (see also Section 6.3). The control of these covariance terms can get tricky, and this particular choice of $H_I(m)$ permits to obtain the desired control.

Since $H_I(m)$ is function of $I$, it is not immediate that $(\mathcal{S}_I)_{I \subseteq \mathbb{Z}^+}$ is a proper partition. We establish this fact in the next lemma, together with a useful property of this partition.

**Lemma 3.3.** The collection $(\mathcal{S}_I)_{I \subseteq \mathbb{Z}^+}$ is a partition of $\mathcal{F}$ and

$$F \in \left( \bigcup_{I \neq \emptyset} \mathcal{S}_I \right)^c \implies \|F_m\|_2 \leq \begin{cases} \Gamma \rho_m & \text{if } J_m \leq j_n, \\ \gamma 2^{-J_m(s+1/2)} & \text{if } J_m > j_n. \end{cases}$$

The previous lemma is one of the key result. In conjunction with Lemmas 3.1 and 3.2, it implies the following corollary which is the starting point of the proof of Theorem 2.2.

**Corollary 3.1.** For all choice of $J_0, \Gamma, \gamma, \xi$ and for all $(R, s)$ there exists $M > 0$ such that the following bound is true.

$$\mathbb{E}_L \Pi(\theta : \|L\theta - L\|_\infty > M \varepsilon_n^*(s) | X_n) \leq \sum_{I \subseteq \mathbb{Z}^+ \backslash I \neq \emptyset} \mathbb{E}_L \Pi(\mathcal{C}_s \cap \mathcal{S}_I | X_n).$$

Our strategy is then to bound each of the terms $\mathbb{E}_L \Pi(\mathcal{C}_s \cap \mathcal{S}_I | X_n)$ for $I \neq \emptyset$; which is done in the Theorem 3.1 below. Interestingly, the technique is reminiscent to the classical testing approach of (Ghosal et al., 2000; Ghosal and van der Vaart, 2007b,a) with extra cares. Large parts of the proofs of Theorems 2.2 and 3.1 rely on the fine tuning of the constants $\delta, J_0, \Gamma, \gamma, \xi$, as well as the relation between those constants, and also on taking $n$ sufficiently large. Since the proofs are quite long, it can be challenging to keep track all along of the constraints those constants must satisfy. To facilitate the understanding of the theorems and their proof, we summarize in the next assumption how $\delta, J_0, \Gamma, \gamma, \xi$ and $n$ must be taken at the end of the day for the theorems to hold true.

**Assumption 1.** We assume that there are constants $K_0, K_1, K_2, K_3, K_4 > 0$ eventually large and eventually depending on $(R, s)$ but solely on $(R, s)$, such that

1. $J_0 \geq K_0 \max\{1, \log(1/\delta), \log(\gamma)\}$;
2. $\log(n) \geq K_1 \max\{2^{h_0}, \delta^{-2}, \log(\Gamma), \log(\gamma), \log(\xi)\}$;
3. $\xi \geq K_2 \delta^{-1} \max\{1, \gamma\} 2^{h_0}$;
4. $\gamma \geq K_3 \max\{1, \delta^{-1}\}$;
5. $\Gamma \geq \max\{\gamma, K_4 2^{h_0}, K_4 \xi\}$.

The constant $\delta$ will be taken as small as needed.

Finally, bounding $\mathbb{E}_L \Pi(\mathcal{C}_s \cap \mathcal{S}_I | X_n)$ rely on splitting the parameter $F$ into two parts $F = (F_I, \bar{F}_I)$ where $F_I := (F_m)_{m \in I}$ and $\bar{F}_I := (F_m)_{m \not\in I}$. The following functions will
also be used:

\[ S_{F_I} := \sum_{m \in I} F_m, \text{ and } S_{\bar{F}_I} := \sum_{m \notin I} F_m. \]

**Theorem 3.1.** Suppose Assumption 1 is satisfied with constants \( K_0, K_1, K_2 > 0 \) sufficiently large and let \( \Pi \) be a prior such that \( F_I \) and \( \bar{F}_I \) are independent for all \( I \subseteq \mathbb{Z}_+ \). Then, there are constants \( c_0, c_1, c_2, \delta_0 > 0 \) such that for all \( 0 < \delta \leq \delta_0 \), for all \( \alpha \leq 1/2 \), and for all \( I \subseteq \mathbb{Z}_+ \),

\[
E_L[\Pi(\mathcal{E} \cap \mathcal{A}_I | X_n)] \leq \left\{ \exp \left\{ -c_2 n \mathcal{E}_f^2 + c_1 2^L_b |I| \right\} \alpha \frac{\Pi(\mathcal{A}_I)}{\Pi(\mathcal{A}_I)} \right\}^{1-\alpha},
\]

where \( \mathcal{E}_I := \inf\{E_L[S_{F_I}^2]^{1/2} : F \in \mathcal{I}_I\}, \mathcal{A}_I := \{F_I : \|F_m\|_{L^2} > H_I(m)\}, \text{ and } \hat{\mathcal{A}}_I := \{F_I : E_L[S_{F_I}^2] \leq \delta^2 \mathcal{E}_f^2, \|S_{F_I}\|_{\infty} \leq \delta\}; \) provided that \( \Pi(\hat{\mathcal{A}}_I) > 0 \) and \( \mathcal{E}_I > 0 \) for all \( I \subseteq \mathbb{Z}_+ \).

We note that the fact that \( \mathcal{E}_I > 0 \) and \( \Pi(\hat{\mathcal{A}}_I) > 0 \) for the spike-and-slab prior are consequences of Lemma 5.1 and Proposition 7.2 that will be established later. Also, we point out that in the whole paper we make the abuse of notations of writing \( \Pi \) to denote the prior on \( \theta, L_\theta, L_F, F_\cdot \), as well as for the restricted parameters \( F_I \) or \( \bar{F}_I \). The proof of the Theorem 2.2 consists on specializing the bound of Theorem 3.1 to the spike-and-slab prior and using it in conjunction with Corollary 3.1 to conclude.

4. Discussion

**The master theorem of Bayesian nonparametrics** The current state-of-the-art method in calculating posterior contraction rates is the master theorem developed by Ghosal et al. (2000); Ghosal and van der Vaart (2007a); Shen et al. (2001). This theorem relies on two main ingredients:

- The existence of tests for the hypothesis \( H_0 : L' = L \) against the alternative \( H_1 : \|L' - L\|_\infty > \alpha \nu_n(s) \), with Type I and Type II errors decreasing as \( \exp\left\{ -K n \nu_n(s)^2 \right\} \);
- The prior puts enough mass on certain Kullback-Leibler neighborhoods of \( L \).

In the context of \( L_\infty \) contraction, it is known that the master theorem yields suboptimal contraction rates (Giné and Nickl, 2011; Hoffmann et al., 2015; Yoo et al., 2017). The issue is discussed thoroughly in (Hoffmann et al., 2015; Yoo et al., 2017): no test has enough power to obtain the optimal rate of contraction in \( L_\infty \). In particular, the Type II error has to decay polynomially in \( n \), unless we deteriorate the rate. It is known that not all the alternative \( H_1 \) has to be tested – only a suitable sieve – but this does not help either to get optimal rates, the root of the problem being deeper.

The arguments in Hoffmann et al. (2015); Yoo et al. (2017) are strong, and it is natural to ask what is wrong in the current paper such that the tests we use in the proofs of Theorems 2.2 and 3.1 permit optimal contraction rates. This indeed relies on the nature
of the alternative we test. We are not constructing tests for $H_1 : \|L' - L\|_\infty > M\varepsilon_n^*(s)$, but instead for each $I \subseteq \mathbb{N}$, we build a test for $H_1 : L' \in \{L_F : F \in \mathcal{F}_I\}$. Those tests (see the proof of Lemma 6.2) have Type I and Type II errors decreasing as $\exp\{-Kn\varepsilon^2_n\}$, which is typically polynomial in $n$ when $|I|$ is small, and thus not in contradiction with the arguments of the aforementioned papers. We remark that $|I|$ small corresponds exactly to those log-densities $L'$ that can be far from $L$ in $L_\infty$ but close in $L_2$, and thus hard to separate. When $|I|$ gets large, however, the powers of the tests increase, which is what saves us.

The main drawback of the method is getting a sharp enough bound on the denominator of the Bayes rule, which seems hard to do beyond the scope of independent wavelet coefficients, or at least having a nice structure. Anyhow, we believe the approach of the current paper shows that the master theorem of Bayesian nonparametrics can be still useful for $L_\infty$ contraction, giving some hope toward a general $L_\infty$ contraction result of the same flavour.

**Suboptimality when $0 < s \leq 1/2$** The rates of Theorem 2.2 are slightly suboptimal in the region $0 < s \leq 1/2$. The problem is indeed not inherent to the spike-and-slab prior, and as such not surprising as it is known density estimation on the interval behave very differently when $0 < s \leq 1/2$ or $s > 1/2$, see for instance Brown et al. (1998). Our troubles come from the impossibility of taking $\rho_m = \sqrt{\log(n)/n}$ when $s \leq 1/2$ and we have instead to take a much larger threshold $\rho_m = 2^{-j_m/2}\varepsilon_n^*(s)$. The reasons for this impossibility are to be found in controlling some covariance terms when decomposing the likelihood, see Section 6.3. In fact, this exhibits a major difference on the strength of the result we prove here: in the case $s > 1/2$ the control of the posterior is much tighter. In particular, we prove that every wavelet coefficients of $L_\theta$ at level $j \leq j_n$ is within $\sqrt{\log(n)/n}$ distance of the coefficients of $L$ if $s > 1/2$, while we are only able to get a distance of $2^{-j/2}\varepsilon_n^*(s)$ otherwise (which is much larger when $j$ is small).

To the best of our knowledge, no method based on asymptotic expansions of the log-likelihood succeeded before in getting posterior $L_\infty$ rates when $s \leq 1/2$. Thereby, the strategy developed here shed new light on our understanding of $L_\infty$ contraction. In view of the recent result of Castillo and Mismer (2019), however, methods based on conjugacy arguments are able to obtain adaptivity and optimality over all $0 < s \leq 1$, with no extra log$(n)$ factor. This shows that we don’t really understand yet enough the behaviour of the log-likelihood when $0 < s \leq 1/2$, which should be investigated in a near future by the author.

**Estimation of the derivatives** The spike-and-slab prior of Section 2.2 also achieves optimal contraction rates for estimating the derivatives of the density. We remark that if $L$ has derivatives $L^{(r)}$, $r \geq 1$ integer, then $s \geq 1 > 1/2$. Then, in this case, investigation of the proof of Theorem 2.2 shows that the posterior contracts on the set $\{L' : |\langle L' - L, \varphi_{j,k} \rangle| \lesssim \sqrt{\log(n)/n}1_{j \leq j_n} + 2^{-j(s+1/2)}1_{j > j_n} \forall (j,k) \in \mathcal{V}\}$. Then, it is a classical result
that this implies for all 1 ≤ r ≤ s ≤ S with r integer,
\[
\sup_{L \in \Omega(R,s)} \mathbb{E}_L \Pi(\theta : \|L^{(r)}_\theta - L^{(r)}\|_\infty > M\varepsilon_n(s)^{s-r} | X_n) = o(1).
\]

5. Proof of the Theorem 2.2

As explained in Section 3.3, the proof of Theorem 2.2 consists on plugging the bound of Theorem 3.1 into the bound of Corollary 3.1. The first step is to obtain an upper estimate on \(\Pi(A_1)/\Pi(A_I)\). The next lemma is proved in Section S5.3.

**Lemma 5.1.** Suppose Assumption 1 is satisfied with constants \(K_0, K_1, K_2, K_3 > 0\) sufficiently large. Then, there are universal constants \(\nu_1, \nu_2 > 0\) such that for all \(I \subseteq \mathbb{Z}_+\),
\[
\frac{\Pi(A_I)}{\Pi(A_I)} \leq \nu_1 \exp \left\{ \nu_2 \log(n) \sum_{m \in I} 2^{J_m - \nu_3 K_2} 1_{J_m \leq j_n} - (1 + \mu_s) \log(2) \sum_{m \in I} J_m 1_{J_m > j_n} \right\}.
\]
Furthermore, if \(I \cap \{m : J_m > \log(n)/\log(2)\} \neq \emptyset\), then \(\Pi(A_I)/\Pi(A_I) = 0\).

Then, we can plug the bound of Lemma 5.1 into the Theorem 3.1 and fine-tune \(\alpha\) in function of \(I\) to obtain a clean bound on \(\mathbb{E}_L \Pi(\mathcal{C}_n \cap \mathcal{J}_I | X_n)\). We do so in the following lemma, also proved in Section S5.3.

**Lemma 5.2.** Suppose Assumption 1 is satisfied with constants \(K_0, K_1, K_2, K_3, K_4 > 0\) sufficiently large and \(\delta > 0\) is taken sufficiently small. Then, there are universal constants \(\nu_3, \nu_4 > 0\) such that for all \(I \subseteq \mathbb{Z}_+\),
\[
\mathbb{E}_L \Pi(\mathcal{C}_n \cap \mathcal{J}_I | X_n) \leq \nu_4 \prod_{m \in I} n^{-\nu_3 K_2} \prod_{m \in I} 2^{-J_m (1 + \mu_s/2)}.
\]

Now we are in position to use the bound established in Lemma 5.2 with the inequality of Corollary 3.1 to finish the proof. Define for simplicity \(g_m := \nu_3 K_2 \log(n)\) if \(0 \leq J_m \leq j_n\) and \(g_m := J_m (1 + \mu_s/2) \log(2)\) if \(J_m > j_n\). Then,
\[
\sum_{I \subseteq \mathbb{Z}_+ \atop I \neq \emptyset} \mathbb{E}_L \Pi(\mathcal{C}_n \cap \mathcal{J}_I | X_n) \leq \nu_4 \sum_{b \in \{0,1\}^{\mathbb{Z}_+}} 1_{\{\sum_{m} b_m \geq 1\}} \prod_{m \in \mathbb{Z}_+} e^{-g_m b_m}
\]
\[
\leq \nu_4 \sum_{m' \in \mathbb{Z}_+} \sum_{b \in \{0,1\}^{\mathbb{Z}_+}} b_{m'} \prod_{m \in \mathbb{Z}_+} e^{-g_m b_m}
\]
\[
= \nu_4 \sum_{m' \in \mathbb{Z}_+} \sum_{b \in \{0,1\}^{\mathbb{Z}_+}} e^{-g_m b_{m'} b_m} \prod_{m \in \mathbb{Z}_+} e^{-g_m b_m}
\]
\[
= \nu_4 \sum_{m' \in \mathbb{Z}_+} e^{-g_{m'} \prod_{m \in \mathbb{Z}_+} (1 + e^{-g_m})}.
\]
Hence we get the bound,
\[
\sum_{I \subset \mathbb{Z}^+} \mathbb{E}_I \Pi(\mathcal{E}_* \cap \mathcal{F}_I \mid X_n) \leq \nu_4 \exp \left( \sum_{m \in \mathbb{Z}^+} e^{-g_m} \right) \sum_{m \in \mathbb{Z}^+} e^{-g_m}.
\]

The previous display is \(o(1)\) whenever \(\sum_{m \in \mathbb{Z}^+} e^{-g_m} = o(1)\), which we prove now. Indeed,
\[
\sum_{m \in \mathbb{Z}^+} e^{-g_m} = \sum_{m \in \mathbb{Z}^+} e^{-g_m} 1_{J_m \leq j_n} + \sum_{m \in \mathbb{Z}^+} e^{-g_m} 1_{J_m > j_n}
\]
\[
\lesssim n^{-\nu_3 J_1^2} \sum_{m \in \mathbb{Z}^+} 1_{J_m \leq j_n} + \sum_{m \in \mathbb{Z}^+} 2^{-j(1+\mu_\star/2)} 1_{J_m > j_n}
\]
\[
\lesssim n^{-\nu_3 J_1^2} \sum_{j=0}^{j_n} 2^j + \sum_{j > j_n} 2^{-j(1+\mu_\star/2)} 2^j
\]
\[
\lesssim n^{-\nu_3 J_1^2} 2^{j_n} + 2^{-j_n \mu_\star/2},
\]
where the third line follows as there are no more than \(\lesssim 2^j\) wavelets at level \(j\). Now, we remark that \(2^{j_n} \asymp (n/\log(n))^{2/(2s+1)}\). Hence, if \(K_4\) is taken large enough, \(\sum_{m \in \mathbb{Z}^+} e^{-g_m} = o(1)\), as claimed.

6. Proof of the Theorem 3.1

6.1. Main ideas

We already know that \(\mathcal{F}_I \subseteq \{F_I : \mathbb{E}_I[S_{F_I}^2] \geq \mathcal{E}_I^2\}\) by construction. We obtain a finer result by further slicing the set \(\mathcal{F}_I\). For \(y \geq 1\) integer, we let
\[
\mathcal{F}_I^y := \mathcal{F}_I \cap \{F : y \mathcal{E}_I^2 \leq \mathbb{E}_I[S_{F_I}^2] < (y+1) \mathcal{E}_I^2\}.
\]

Similarly, we define \(\mathcal{A}_I^y := \mathcal{A}_I \cap \{F : y \mathcal{E}_I^2 \leq \mathbb{E}_I[S_{F_I}^2] < (1+y) \mathcal{E}_I^2\}\). Clearly \((\mathcal{A}_I^y)_{y \geq 1}\) is a partition of \(\mathcal{A}_I\). The first lemma establishes a first bound on the posterior mass of \(\mathcal{F}_I^y\).

Lemma 6.1. Suppose Assumption 1 is satisfied with constants \(K_0, K_1, K_2 > 0\) sufficiently large. Also suppose \(\Pi\) is such that \(F_I\) and \(\bar{F}_I\) are independent for all \(I \subseteq \mathbb{N}\), and \(\Pi(\bar{A}_I) > 0\) for all \(I \subseteq \mathbb{Z}^+\). Then, there are universal constants \(c_0, \delta_0 > 0\) such that for all \(t > 0\) there is an event \(\Omega_t\) with \(\mathbb{P}_t^f(\Omega_t) \leq e^{-t}\) and if \(X_n \in \Omega_t\), for all \(0 < \delta \leq \delta_0\), for all \(I \subseteq \mathbb{Z}^+,\) for all \(y \geq 1\),
\[
\Pi(\mathcal{E}_* \cap \mathcal{F}_I^y \mid X_n) \leq 2e^{2\delta t + c_0 \delta y \mathcal{E}_I^2} \int_{\mathcal{A}_I^y \cap \{\|S_{F_I}\|_{\infty} \leq y\}} \prod_{i=1}^n \frac{q_{F_I}(X_i) \Pi(\Pi(dF_i))}{\Pi(L_i) \Pi(\mathcal{A}_I^y)},
\]
where \(q_{F_I}\) is a probability density on \([0,1]\) whose exact expression is known but deferred to the proof of the lemma for convenience.
The last lemma is the key result of the proof. Interestingly, the classical approach to concentration results à la Ghosal et al. (2000) consists on establishing a similar relation, but with $q_{F_I}$ replaced by $\exp\{L_F\}$ and $A_I$ replaced by a Kullback-Leibler neighborhood of $p_L$. We use the estimate of Lemma 6.1 to bound $E_L[\Pi(\mathcal{C}_* \cap \mathcal{F}_I^y \mid X_n)]$ using the standard testing approach à la Ghosal et al. (2000), coupled with the square-root trick of Lijoi et al. (2005); Walker et al. (2007); Ghosal and van der Vaart (2007b). This step is rather immediate in view of the existing literature and it boils down to bound $\inf\{H(Q_{S_{F_I}}, P_L)^2 : F_I \in A_I^y, \|S_{F_I}\|_\infty \leq \delta\}$ and the metric entropy (in the Hellinger distance) of the set of densities $\mathcal{P}_I^y := \{q_{F_I} : F_I \in A_I^y, \|S_{F_I}\|_\infty \leq \delta\}$. This gives the following lemma.

**Lemma 6.2.** Suppose Assumption 1 is satisfied with constants $K_0, K_1, K_2 > 0$ sufficiently large, and let everything as in Lemma 6.1. Then, there are universal constants $c_1, \delta_0 > 0$ such that for all $0 < \delta \leq \delta_0$, for all $I \subseteq \mathbb{Z}_+$, for all $y \geq 1$, for all $t > 0$,

$$E_L[\Pi(\mathcal{C}_* \cap \mathcal{F}_I^y \mid X_n)1_{\Omega_t}] \leq \left\{ \frac{8\exp\left(\frac{-2E_I^2}{256} + 25t + c_12^{5y}|I|\right) \Pi(A_I^y)}{1 - e^{-8yE_I^2/256} \Pi(A_I)} \right\}^{1/2}.$$  

We can obtain a bound on $E_L[\Pi(\mathcal{C}_* \cap \mathcal{F}_I^y \mid X_n)1_{\Omega_t}]$ by summing over $y \geq 1$ the bound obtained in Lemma 6.2. This gives a valid bound, but it is in not sharp enough in cases where $|I|$ gets too large or $nE_I^2$ is too small. Indeed, in those cases, we can improve the bound to give more importance to the prior by remarking that taking the expectation both sides of the expression in Lemma 6.1 and applying Fubini’s theorem gives

$$E_L[\Pi(\mathcal{C}_* \cap \mathcal{F}_I^y \mid X_n)1_{\Omega_t}] \leq 2e^{25t + c_1\delta y nE_I^2} \frac{\Pi(A_I^y)}{\Pi(A_I)}, \quad (6.2)$$

This improvement permits to assume only $\mu_* > 0$, otherwise we would have to assume $\mu_* > 1$, which may be undesirable in practice (as it may cause over-shrinkage). The next lemma leverages that $E_L[\Pi(\mathcal{C}_* \cap \mathcal{F}_I^y \mid X_n)1_{\Omega_t}]$ is bounded above by the minimum between the expression in Lemma 6.2 and the last display to get a sharp bound on $E_L[\Pi(\mathcal{C}_* \cap \mathcal{F}_I^y \mid X_n)1_{\Omega_t}]$.

**Lemma 6.3.** Suppose Assumption 1 is satisfied with constants $K_0, K_1, K_2 > 0$ sufficiently large, and let everything as in Lemmas 6.1 and 6.2. Then, there are universal constants $c_2, \delta_0 > 0$ such that for all $0 < \delta \leq \delta_0$, for all $I \subseteq \mathbb{Z}_+$, for all $\frac{512c_0^2}{1 + 312c_0^2} < \alpha < 1/2$, for all $t > 0$,

$$E_L[\Pi(\mathcal{C}_* \cap \mathcal{F}_I^y \mid X_n)1_{\Omega_t}] \leq \frac{\sqrt{8e^{25t} \exp\left\{-c_2nE_I^2 + c_12^{5y}|I|\right\}} \Pi(A_I)}{(1 - e^{-c_2nE_I^2})^{2\alpha}} \Pi(A_I)^{1-\alpha}.$$

Finally, to obtain the bound in the statement of the theorem, we note that,

$$E_L[\Pi(\mathcal{C}_* \cap \mathcal{F}_I^y \mid X_n) \leq \inf_{t \geq 0} \left\{ e^{-t} + E_L[\Pi(\mathcal{C}_* \cap \mathcal{F}_I \mid X_n)1_{\Omega_t}] \right\}].$$

Plugging the bound obtained in Lemma 6.3 into the previous display and solving to find the infimum gives the bound of the theorem when choosing $\delta$ small enough.
6.2. Proofs of Lemmas 6.1 to 6.3

Proof of Lemma 6.1. Let define \( \Phi(f) := f - \mathbb{E}_L[f] - \log \mathbb{E}_L[e^{f - \mathbb{E}_L[f]}] \) and \( C(f, g) := \log \mathbb{E}_L[e^{\Phi(f)}e^{\Phi(g)}] \). We will see that \(-C(S_{F_l}, S_{F_l})\) is asymptotically equivalent to the covariance of \( S_{F_l} \) and \( S_{F_l} \), and thus we will refer abusively to this term as the covariance from now on. It is easily seen that the log-likelihood can be rewritten as

\[
L_F - L = \Phi(S_{F_l}) + \Phi(S_{F_l}) - C(S_{F_l}, S_{F_l}).
\]

Then, by the Bayes rule,

\[
\Pi(\mathcal{G}_s \cap \mathcal{J}_l^y | X_n) = \frac{\int_{\mathcal{G}_s \cap \mathcal{J}_l^y} e^{n \bar{Y}_n \Phi(S_{F_l})} e^{n \bar{Y}_n \Phi(S_{F_l})} e^{-n C(S_{F_l}, S_{F_l})} \Pi(dF)}{\int e^{n \bar{Y}_n \Phi(S_{F_l})} e^{n \bar{Y}_n \Phi(S_{F_l})} e^{-n C(S_{F_l}, S_{F_l})} \Pi(dF)}. \tag{6.3}
\]

Recall that \( F_l := (F_m)_{m \in I} \) and \( \bar{F}_l := (F_m)_{m \not\in I} \). Also, in addition to \( A_l^y \), we let

\[
N_l := \left\{ \bar{F}_l : \forall m \not\in I, \|F_m\|_2 \leq H_l(m) \right\}.
\]

It is immediate that if \( F \in \mathcal{G}_s \cap \mathcal{J}_l^y \) then \( F_l \in A_l^y \) and \( \bar{F}_l \in N_l \). Also, if \( F \in \mathcal{G}_s \), then

\[
\max\{|S_{F_l}(x)|, |S_{F_l}(x)|\} \leq \max\left\{ \sum_{m \in I} |F_m(x)|, \sum_{m \not\in I} |F_m(x)| \right\} \leq \sum_{m \geq 0} |F_m(x)| \leq \delta.
\]

So for all \( F \in \mathcal{G}_s \) and for all \( I \subseteq \mathbb{Z}_+ \), we have \( \|S_{F_l}\|_\infty \leq \delta \) and \( \|S_{F_l}\|_\infty \leq \delta \). It follows by equation (6.3) that \( \Pi(\mathcal{G}_s \cap \mathcal{J}_l^y | X_n) \) is bounded from above by

\[
\frac{\int 1_{A_l^y}(F_l)1_{N_l}(\bar{F}_l)1_{\|S_{F_l}\|_\infty \leq \delta}1_{\|S_{F_l}\|_\infty \leq \delta}e^{n \bar{Y}_n \Phi(S_{F_l})} e^{n \bar{Y}_n \Phi(S_{F_l})} e^{-n C(S_{F_l}, S_{F_l})} \Pi(dF)}{\int 1_{A_l^y}(F_l)1_{N_l}(\bar{F}_l)1_{\|S_{F_l}\|_\infty \leq \delta}1_{\|S_{F_l}\|_\infty \leq \delta}e^{n \bar{Y}_n \Phi(S_{F_l})} e^{n \bar{Y}_n \Phi(S_{F_l})} e^{-n C(S_{F_l}, S_{F_l})} \Pi(dF)}.
\]

The main challenge in the proof of the theorem is to control the term \( C(S_{F_l}, S_{F_l}) \) both in the numerator and denominator, which is deferred to Section 6.3. In fact, by Corollaries 6.1 and 6.2, if the constants \( K_0, K_1, K_2 \) in Assumption 1 are taken sufficiently large, there is a universal \( C > 0 \) such that taking \( \delta \) small enough gives

\[
2e^{-C \delta n^{\gamma}} \frac{\int 1_{A_l^y}(F_l)1_{N_l}(\bar{F}_l)1_{\|S_{F_l}\|_\infty \leq \delta}1_{\|S_{F_l}\|_\infty \leq \delta}e^{n \bar{Y}_n \Phi(S_{F_l})} e^{n \bar{Y}_n \Phi(S_{F_l})} \Pi(dF)}{\int 1_{A_l^y}(F_l)1_{N_l}(\bar{F}_l)1_{\|S_{F_l}\|_\infty \leq \delta}1_{\|S_{F_l}\|_\infty \leq \delta}e^{n \bar{Y}_n \Phi(S_{F_l})} e^{n \bar{Y}_n \Phi(S_{F_l})} \Pi(dF)},
\]

But \( F_l \) is independent of \( \bar{F}_l \) and \( S_{F_l} \) is solely function of \( F_l \) (respectively \( S_{F_l} \) and \( \bar{F}_l \)), thus

\[
\Pi(\mathcal{G}_s \cap \mathcal{J}_l^y | X_n) \leq 2e^{-C \delta n^{\gamma}} \frac{\int 1_{\|S_{F_l}\|_\infty \leq \delta}e^{n \bar{Y}_n \Phi(S_{F_l})} \Pi(dF)}{\int e^{n \bar{Y}_n \Phi(S_{F_l})} \Pi(dF)} = 2e^{-C \delta n^{\gamma}} \frac{\int 1_{\|S_{F_l}\|_\infty \leq \delta}e^{n \bar{Y}_n \Phi(S_{F_l})} \Pi(dF)}{\int e^{n \bar{Y}_n \Phi(S_{F_l})} \Pi(dF)}, \tag{6.4}
\]
where the second line follows because \( F_I \in \tilde{A}_I \Rightarrow \| S_{F_I} \|_{\infty} \leq \delta \) by construction. It is interesting that \( x \mapsto p_L(x)e^{\Phi S_{F_I}(x)} \) is indeed a proper density function, i.e. it is non-negative and integrates to 1. We write \( q_{F_I}(x) := p_L(x)e^{\Phi S_{F_I}(x)} \). We then can bound the expectation of equation (6.4) using the standard approach à la Ghosal et al. In particular, we arrive at the bound of equation (6.1) by controlling the \( \mathbb{P}_L \)-probability of the event

\[
\Omega_t := \left\{ \mathbf{x}_n : \prod_{i=1}^{n} \frac{q_{F_{\mathcal{I}}}(X_{i})}{p_L(X_{i})} \geq e^{-n\delta^2\varepsilon_t^2}e^{-\sqrt{2n\delta^2\delta t} - \delta t} \right\}.
\]

**Proposition 6.1.** Let \( \Pi \) be any probability measure supported on the set \( \tilde{A}_I \). For all \( I \subset \mathbb{Z}_+ \), for all \( 0 < \delta \leq \log(2) \), for all \( t > 0 \), and for all \( n > 0 \), \( \mathbb{P}_L(\Omega_t) \geq 1 - e^{-t} \).

Then, on the event that \( \mathbf{x}_n \in \Omega_t \), the equation (6.4) becomes

\[
\Pi(\mathcal{C}_s \cap \mathcal{Y}_t \mid \mathbf{x}_n) \leq 2e^{C\delta_n\varepsilon_t^2 + n\delta^2\varepsilon_t^2 + \sqrt{2n\delta^2\delta t} + \delta t} \prod_{i=1}^{n} \frac{q_{F_{\mathcal{I}}}(X_{i})}{p_L(X_{i})} \, \Pi(\mathcal{A}_I) \blacktriangleleft \Pi(\mathcal{A}_{\mathcal{I}}) \mathcal{S}_F \mathcal{S}_F^2 \leq \delta \}
\]

The conclusion follows because \( y \geq 1 \), and because if \( t \leq 2n\varepsilon_t^2 \) then \( \sqrt{2n\delta^2\varepsilon_t^2} + \delta t \leq 2\delta n\varepsilon_t^2 + \delta t \), while if \( t > 2n\varepsilon_t^2 \) then \( \sqrt{2n\delta^2\delta t} + \delta t \leq 2\delta t \). Hence, we can take \( \delta_0 = \log(2) \) and \( c_0 = C + 2 + \delta_0 \).

**Proof of Lemma 6.2.** First we obtain a lower bound on \( \inf \{ H(Q_{F_I}, P_L)^2 : F_I \in A_I \} \). The following proposition helps.

**Proposition 6.2.** As \( \eta \to 0 \) it holds \( H(Q_{F_I}, P_L)^2 \geq \frac{1}{8} \mathbb{E}_L[\Pi^2(y)]e^{O(n)} \) for all \( S_{F_I} \) satisfying \( \| S_{F_I} \|_{\infty} \leq \eta \). Then \( \inf \{ H(Q_{F_I}, P_L)^2 : F_I \in A_I, \| S_{F_I} \|_{\infty} \leq \delta \} \geq \frac{y^2}{10} \) for \( \delta \) small enough (but not depending on \( I \) nor on \( y \)).

For \( \epsilon > 0 \) and any subset \( A \) of a metric space equipped with metric \( d \), we let \( N(\epsilon, A, d) \) denote the \( \epsilon \)-covering number of \( A \), i.e. the smallest number of balls of radius \( \epsilon \) needed to cover \( A \). If \( d \) is induced by some norm \( \| \cdot \| \), we write \( N(\epsilon, A, \| \cdot \| ) \). By Ghosal and van der Vaart (2007b, Corollary 1) our Proposition 6.2 implies that for all \( D > 0 \), all \( y \geq 1 \), and all \( n \geq 1 \) there exists a test \( \phi_{n,y} \) such that

\[
\mathbb{E}_L[\phi_{n,y}] \leq \frac{N(\sqrt{\frac{\pi^2}{16}}, P_I^y, H)}{D} e^{-\frac{y\varepsilon_t^2}{16}}, \quad \sup_{F_I \in A_I^y, \| S_{F_I} \|_{\infty} \leq \delta} \mathbb{E}_{F_I}[1 - \phi_{n,y}] \leq De^{-\frac{y\varepsilon_t^2}{2\sqrt{\delta}}},
\]

where \( \mathbb{E}_{F_I} \) is understood as the expectation under \( Q_{F_I}^y \), and where \( P_I^y := \{ q_{F_I} : F_I \in A_I^y, \| S_{F_I} \|_{\infty} \leq \delta \} \). Using the estimate of Lemma 6.1 we find that \( \mathbb{E}_L[\Pi(\mathcal{C}_s \cap \mathcal{Y}_t \mid \mathbf{x}_n)1_{\Omega_t}] \)
is bounded by

\[
\mathbb{E}_L[\phi_{n,y}(\mathcal{C}_* \cap \mathcal{F}^y I | X_n)] + \mathbb{E}_L[(1 - \phi_{n,y}) \Pi(\mathcal{C}_* \cap \mathcal{F}^y I | X_n)1_{\Omega}]
\]

\[
\leq \mathbb{E}_L[\phi_{n,y}] + 2e^{256\cdot|H|} \int \mathbb{E}_{F_e}[1 - \phi_{n,y}] \Pi(dF_e) \Pi(\mathcal{A}_I)
\]

\[
\leq \frac{N(\sqrt{\frac{y}{16}}, \mathcal{F}^y I | H)}{D} e^{-\frac{y\mathcal{E}^2}{256}} + 2De^{256\cdot|H|} \frac{\Pi(\mathcal{A}_I)}{\Pi(\mathcal{A}_I)}
\]

The previous display is true for any \(D > 0\) and thus we can optimize over \(D\), which is also known as the square-root trick (Lijoi et al., 2005; Walker et al., 2007). Doing so gives the bound,

\[
\mathbb{E}_L[\Pi(\mathcal{C}_* \cap \mathcal{F}^y I | X_n)1_{\Omega}] \leq \left\{ \frac{\frac{8e^{256\cdot|H|}}{256} N(\sqrt{\frac{y}{16}}, \mathcal{F}^y I | H)}{1 - e^{-\frac{y\mathcal{E}^2}{256}}} \right\} e^{-\frac{y\mathcal{E}^2}{256}}.
\]

To obtain the bound in the statement of the lemma, it is enough to prove that \(\sup_{y \geq 1} N(\sqrt{\frac{y}{16}}, \mathcal{F}^y I | H) \leq \exp(c_1 2^{L_0} |I|)\). The following lemma helps.

**Proposition 6.3.** There exists \(\delta_0 > 0\) such that for all \(\delta \leq \delta_0\), all \(I \subseteq \mathbb{Z}_+\), and all \(F_I, F_I' \in \mathcal{A}_I \cap \{\|S_{F_I}\|_\infty \leq \delta\}\), it holds \(H(Q_{F_I}, Q_{F_I'})^2 \leq \frac{1}{2} \mathbb{E}_L[(S_{F_I} - S_{F_I'})^2]\).

Observe that for \(F_I, F_I' \in \mathcal{A}_I\) we have \(S_{F_I} - S_{F_I'} = \sum_{m \in I} (F_m - F_m')\), where by construction each \(F_m - F_m'\) is in \(\mathcal{F}_m\). Then, by Proposition 7.1-(4), and then by Proposition 7.1-(3),

\[
\mathbb{E}_L[(S_{F_I} - S_{F_I'})^2] \lesssim \sum_{m \in I} \|F_m - F_m'\|_2^2 \lesssim \sum_{m \in I} \sum_{v \in B_m} \langle F_m - F_m', \varphi_v \rangle^2.
\]

On the other hand, for all \(F_I \in \mathcal{A}_I^y\), we have by Proposition 7.1 that,

\[
\sum_{m \in I} \sum_{v \in B_m} \langle F_m, \varphi_v \rangle^2 \lesssim \sum_{m \in I} \|F_m\|_2^2 \lesssim \mathbb{E}_L[S_{F_I}^2] \lesssim y\mathcal{E}^2.
\]

By equations (6.5) and (6.6) and Proposition 6.3, we find that \(N(\sqrt{\frac{y}{16}}, \mathcal{F}^y I | H)\) is no more than the covering number of a ball of radius \(\lesssim \sqrt{\mathcal{E}I}\) with balls of radius \(\sqrt{\mathcal{E}I}_I\) in \(\mathbb{R}^p\) equipped with the euclidean distance, with \(p = \sum_{m \in I} |B_m| \leq |B_0| \cdot |I|\). By Pollard (1990, Lemma 4.1), this implies that there is a universal \(K > 0\) such that

\[
N\left(\sqrt{\frac{y}{16}}, \mathcal{F}^y I | H\right) \leq \max\left\{1, \left(\frac{3K\sqrt{\mathcal{E}I}}{\sqrt{\mathcal{E}I}_I}\right)^p\right\} \leq \max\{1, (3K)^p\}.
\]

Finally, by construction it is true that \(|B_0| \lesssim 2^{L_0}\). \(\square\)
Proof of Lemma 6.3. We use the fact that for all $u,v \geq 0$ we have $\min\{u,v\} = \inf_{\beta \in (0,1)} \{u^{1-\beta} v^\beta\}$. Then, combining the bounds of Lemma 6.2 with the bound of equation (6.2), we get that for any $\beta \in (0,1)$ and $y \geq 1$

\[
\frac{E_L[\Pi(\mathcal{E}_* \cap \mathcal{J}^{(y)}_1 | X_n) 1_{\Omega_i}]}{\sqrt{8e^{2\delta t}}} \leq \min \left\{ e^{-y \mu_n \mathcal{E}_1^2} \left( \frac{\Pi(A^y_1)}{\Pi(A_1)} \right)^{1-\beta/2} \left( \frac{\exp\left(\frac{yn\mathcal{E}_1^2}{256} + c_1 2^{J_0 |I|} \right)}{1 - e^{-y \mu_n \mathcal{E}_1^2/256}} \right)^{1/2} \right\} \leq \frac{\exp\left\{ - y \mu_n \mathcal{E}_1^2 \left( \frac{\beta}{512} + c_0 \beta - c_0 \delta \right) + \frac{\beta c_2 2^{J_0 |I|}}{2} \right\}}{(1 - e^{-y \mu_n \mathcal{E}_1^2/256})^{3/2}} \left( \frac{\Pi(A^y_1)}{\Pi(A_1)} \right)^{1-\beta/2}.
\]

Hence, for any $\frac{512 - c_0 \delta}{1 + 512 - c_0 \delta} < \beta \leq 1$, writing $\mu := \frac{\beta}{512} + c_0 \beta - c_0 \delta > 0$ for simplicity, by Hölder’s inequality

\[
\sum_{y \geq 1} e^{-y \mu_n \mathcal{E}_1^2} \left( \frac{\Pi(A^y_1)}{\Pi(A_1)} \right)^{1-\beta/2} \leq \sum_{y \geq 1} e^{-2y \mu_n \mathcal{E}_1^2/\beta} \left( \frac{\Pi(A^y_1)}{\Pi(A_1)} \right)^{1-\beta/2} = \frac{e^{-\mu \mathcal{E}_1^2}}{(1 - e^{-2\mu \mathcal{E}_1^2/\beta})^{3/2}} \left( \frac{\Pi(A^y_1)}{\Pi(A_1)} \right)^{1-\beta/2},
\]

where the second line follows since $(A^y_1)_{y \geq 1}$ is a partition of $A_1$. Therefore,

\[
\sum_{y \geq 1} \frac{E_L[\Pi(\mathcal{E}_* \cap \mathcal{J}^{(y)}_1 | X_n) 1_{\Omega_i}]}{\sqrt{8e^{2\delta t}}} \leq \frac{\exp\left\{ - y \mu_n \mathcal{E}_1^2 + \frac{\beta c_2 2^{J_0 |I|}}{2} \right\}}{(1 - e^{-y \mu_n \mathcal{E}_1^2/256})^{3/2}} \left( \frac{\Pi(A^y_1)}{\Pi(A_1)} \right)^{1-\beta/2}.
\]

By taking $\alpha = \beta/2$ and $\beta > \frac{1024 \cdot C_\delta}{1 + 512 - C_\delta}$ we have that $\mu > \frac{1}{2} (\frac{\beta}{512} + C\delta) \beta$, whence the conclusion. \qed

6.3. Control of the covariance terms

The major difficulty in establishing the Theorem 3.1 is to prove estimates on the covariance terms $C(S_{F_1}, S_{F_2})$ that are sharp enough. The estimate used in the proof of Theorem 3.1 are established in the Corollaries 6.1 and 6.2 below, which are consequences of the next lemma. The proof of the Lemma 6.4 is quite long and is deferred to Section S5.2.

Lemma 6.4. Suppose Assumption 1 is satisfied with constants $K_0, K_1, K_2 > 0$ sufficiently large. Then, there is a constant $C > 0$ such that for all $1 < \delta \leq 1$, for all $I \subseteq \mathbb{Z}_+$, for all $\|S_{F_1}\|_\infty \leq \delta$, and for all $S_{F_1} \in N_I$, if $s > 1/2$

\[
|C(S_{F_1}, S_{F_2})| \leq C_\delta E_L[|S_{F_1}^2|] + \delta \Gamma \sqrt{\frac{\log(n)}{n}} \|F_0\|_2 1_{0 \in I} + \delta \Gamma e^{-1 \alpha \eta} \left\{ \sum_{m \in I} \|F_m\|_2^2 1_{J_m \leq J_n} \right\}^{1/2} \sqrt{\frac{\log(n)}{n}} + \frac{\delta}{\sqrt{n}} \left\{ \sum_{m \in I} \|F_m\|_2^2 1_{J_m > J_n} \right\}^{1/2},
\]
and if $0 < s \leq 1/2$, 
\[
|C(S_{F_1}, S_{F_1})| \leq C \delta \mathbb{E}_L[|S_{F_1}^2|] + \delta \Gamma 2^{-J_0/2} \varepsilon_n^*(s)\|F_0\|_2 1_{0\in I} \\
+ \delta \left\{ \sum_{m \in I} \|F_m\|^2 1_{J_m \leq j_n} \right\}^{1/2} \left\{ \Gamma^2 \xi^{-2} \delta_{m \in I} \varepsilon_n^*(s)^2 \sum_{m \in I} 2^{-J_m} 1_{J_m \leq j_n} \right\}^{1/2} \\
+ \delta \left\{ \sum_{m \in I} \|F_m\|^2 1_{J_m > j_n} \right\}^{1/2} \left\{ \gamma \sum_{m \in I} 2^{-J_m(2s+1)} 1_{J_m > j_n} \right\}^{1/2}.
\]

Corollary 6.1. Suppose Assumption 1 is satisfied with constants $K_0, K_1, K_2 > 0$ sufficiently large. Then, there is a constant $C > 0$ such that for all $1 < \delta \leq 1$, for all $I \subseteq \mathbb{Z}_+$, for all $y \geq 1$, for all $F_i \in A_i^y$, and for all $F_i \in N_I$

\[
C(S_{F_1}, S_{F_1}) \geq -C\delta y\varepsilon_I^2 - \frac{\delta}{n}.
\]

Corollary 6.2. Suppose Assumption 1 is satisfied with constants $K_0, K_1, K_2 > 0$ sufficiently large. Then, there is a constant $C > 0$ such that for all $1 < \delta \leq 1$, for all $I \subseteq \mathbb{Z}_+$,

\[
\sup \{ C(S_{F_1}, S_{F_1}) : F_i \in \hat{A}_i, F_i \in N_I \} \leq C\delta^2 \varepsilon_I^2.
\]

6.4. Proofs of Propositions 6.1 to 6.3

Proof of Proposition 6.1. The proof is an adaptation of the classical Ghosal et al. (2000, Lemma 8.1). The first step is to remark that by Jensen’s inequality applied to the logarithm

\[
\log \int_{\hat{A}_i} e^{\pi_n \Phi(S_{F_1})} \Pi(dF_i) \geq \sum_{i=1}^n \int_{\hat{A}_i} \left( \Phi(S_{F_1}(X_i) - \mathbb{E}_L[\Phi(S_{F_1})]) \right) \Pi(dF_i) \\
+ n \int_{\hat{A}_i} \mathbb{E}_L[\Phi(S_{F_1})] \Pi(dF_i).
\]

For all $F_i \in \hat{A}_i$, since $\mathbb{E}_L[S_{F_1}^2] = 0$, we have whenever $0 < \delta \leq \log(2)$,

\[
\mathbb{E}_L[\Phi(S_{F_1})] = -\log \mathbb{E}_L[e^{S_{F_1}}] \\
\geq -\log \mathbb{E}_L \left[ 1 + S_{F_1} + \frac{1}{2} S_{F_1}^2 e^{\|S_{F_1}\|_\infty} \right] \\
\geq -\log \left( 1 + \mathbb{E}_L[|S_{F_1}^2|] \right).
\]

By definition $\mathbb{E}_L[S_{F_1}^2] \leq \delta^2 \varepsilon_I^2$ whenever $F_i \in \hat{A}_i$. That is $\mathbb{E}_L[\Phi(S_{F_1})] \geq -\mathbb{E}_L[S_{F_1}^2] \geq -\delta^2 \varepsilon_I^2$ for any $F_i \in \hat{A}_i$. Hence,

\[
\log \int_{\hat{A}_i} e^{\pi_n \Phi(S_{F_1})} \Pi(dF_i) \geq \sum_{i=1}^n \int_{\hat{A}_i} \left( \Phi(S_{F_1}(X_i) - \mathbb{E}_L[\Phi(S_{F_1})]) \right) \Pi(dF_i) - n\delta^2 \varepsilon_I^2. \quad (6.7)
\]
Now we define the random variables $Z_i := \int_{A_i} (\Phi(S_{F_i}(X_i)) - \mathbb{E}_L[\Phi(S_{F_i})]) \Pi(dF_i)$. Observe that $\mathbb{E}_L[Z_i] = 0$, and $\Phi(S_{F_i}) - \mathbb{E}_L[\Phi(S_{F_i})] = S_{F_i}$, so we have $|Z_i| \leq \delta$ because $\Pi$ is a probability measure. Further, by an application of Jensen’s inequality and Fubini’s theorem

$$\mathbb{E}_L[Z_i^2] = \mathbb{E}_L\left[\left( \int_{A_i} S_{F_i}(X_i) \Pi(dF_i) \right)^2 \right]$$

$$\leq \mathbb{E}_L\left[ \int_{A_i} S_{F_i}(X_i)^2 \Pi(dF_i) \right]$$

$$= \int_{A_i} \mathbb{E}_L[S_{F_i}^2] \Pi(dF_i).$$

Therefore $\mathbb{E}_L[Z_i^2] \leq \delta^2 \mathcal{E}_i^2$, because of the definition of $A_i$, and because $\Pi$ is a probability measure. By the equation (6.7), the probability of $\sum_{i=1}^{n} Z_i \leq -\delta t - \sqrt{2n\delta^2 \mathcal{E}_i^2 t}$. The conclusion of the proposition then follows by Bernstein’s inequality (Boucheron et al., 2013, Theorem 2.10).

**Proof of Proposition 6.2.** Observe that by definition $q_{F_i}(x) = p_{F_i}(x)e^{\Phi(S_{F_i})}$. Also we have $\mathbb{E}_L[S_{F_i}] = 0$ and thus $\Phi(S_{F_i}) = S_{F_i} - \log \mathbb{E}_L[e^{S_{F_i}}]$. We lower bound $H(Q_{F_i}, P_L)$ by obtaining an upper bound on the Hellinger affinity $R(Q_{F_i}, P_L) := \int_{[0,1]} \sqrt{q_{F_i}p_L} \Pi$ and using that $H(Q_{F_i}, P_L)^2 = 1 - R(Q_{F_i}, P_L)$. Clearly $R(Q_{F_i}, P_L) = \mathbb{E}_L[e^{\frac{1}{2}\Phi(S_{F_i})}] = \mathbb{E}_L[e^{\frac{1}{2}S_{F_i}}]/\mathbb{E}_L[e^{S_{F_i}}]^{1/2}$. But $\|S_{F_i}\|_{\infty} \leq \eta$, thus $\mathbb{E}_L[e^{S_{F_i}}] \geq 1 + \mathbb{E}_L[S_{F_i}] + \frac{1}{2}\mathbb{E}_L[S_{F_i}^2]e^{-\eta} = 1 + \frac{1}{2}\mathbb{E}_L[S_{F_i}^2]e^{-\eta}$. Consequently, $\mathbb{E}_L[e^{S_{F_i}}]^{1/2} \geq 1 + \frac{1}{2}\mathbb{E}_L[S_{F_i}^2]e^{O(\eta)}$. Similarly, $\mathbb{E}_L[e^{\frac{1}{2}S_{F_i}}] \leq 1 + \frac{1}{2}\mathbb{E}_L[S_{F_i}^2]e^{\eta}$. It follows $R(Q_{F_i}, P_L) \leq 1 - \frac{1}{4}\mathbb{E}_L[S_{F_i}^2]e^{O(\eta)}$.

**Proof of Proposition 6.3.** Remark that $\mathbb{E}_L[S_{F_i}] = 0$, so $\Phi(S_{F_i}) = S_{F_i} - \log \mathbb{E}_L[e^{S_{F_i}}]$, and similarly for $\Phi(S_{F_i})$. Since $\|S_{F_i}\|_{\infty} \leq \delta$ then $\|\Phi(S_{F_i})\|_{\infty} \leq 2\delta$, similarly for $S_{F_i}$. By a Taylor expansion there is $u \in (\Phi(S_{F_i}), \Phi(S_{F_i}))$, and hence $|u| \leq 2\delta$, such that $e^{\frac{1}{4}\Phi(S_{F_i})} = e^{\frac{1}{4}\Phi(S_{F_i})} + \frac{1}{2}(\Phi(S_{F_i}) - \Phi(S_{F_i}))e^{u}$. That is, $(e^{\frac{1}{4}\Phi(S_{F_i})} - e^{\frac{1}{4}\Phi(S_{F_i})})^2 \leq \frac{1}{16}e^{4\delta}(\Phi(S_{F_i}) - \Phi(S_{F_i}))^2$. Then we can bound the Hellinger distance as follows.

$$H(Q_{F_i}, Q_{F_i})^2 = \frac{1}{2} \left( \sqrt{p_{F_i}e^{\Phi(S_{F_i})}} - \sqrt{p_{F_i}e^{\Phi(S_{F_i})}} \right)^2$$

$$= \frac{1}{2} \mathbb{E}_L\left[ \left( e^{\frac{1}{4}\Phi(S_{F_i})} - e^{\frac{1}{4}\Phi(S_{F_i})} \right)^2 \right]$$

$$\leq \frac{1}{8}e^{4\delta}\mathbb{E}_L[(\Phi(S_{F_i}) - \Phi(S_{F_i}))^2].$$

Expanding the square in the last equation and using that $\mathbb{E}_L[S_{F_i}] = \mathbb{E}_L[S_{F_i}] = 0$, we find that

$$H(Q_{F_i}, Q_{F_i})^2 \leq \frac{1}{8}e^{4\delta}\mathbb{E}_L[(S_{F_i} - S_{F_i})^2] + \frac{1}{8}e^{4\delta}\log \frac{\mathbb{E}_L[e^{S_{F_i}}]}{\mathbb{E}_L[e^{S_{F_i}}]}. \quad (6.8)$$
Proof of Corollary 6.2.

Proof of Corollary 6.1: the case where $s > 1/2$. By construction, if $m \in I$ and $F \in \mathcal{A}_I$, then $\|F_m\|_2 \geq H_I(m)$. Thus, if $I \cap \{m : J_m \leq J_n\} \neq \emptyset$, for all $F \in \mathcal{A}_I$,

$$
(\Gamma \xi^{-10\varepsilon I})^2 \frac{\log(n)}{n} \leq \sum_{m \in I} H_I(m)^2 1_{J_m \leq J_n} \leq \sum_{m \in I} \|F_m\|_2^2 1_{J_m \leq J_n}.
$$

Also if $0 \in I$, then $\Gamma^2 \log(n)/n \leq \|F_0\|_2^2$. Therefore Lemma 6.4 and Young’s inequality imply

$$
|C(S_{F_I}, S_{F_I})| \leq C \delta \mathbb{E}_L[S_{F_I}^2] + \delta \|F_0\|_2^2 1_{0 \in I} + \delta \sum_{m \in I} \|F_m\|_2^2 1_{J_m \leq J_n} + \frac{\delta}{2} \sum_{m \in I} \|F_m\|_2^2 1_{J_m > J_n} + \frac{\delta}{2n}.
$$

The conclusion follows since $\sum_{m \in I} \|F_m\|_2^2 \geq \mathbb{E}_L[S_{F_I}^2]$ by Proposition 7.1-(4).

Proof of Corollary 6.1: the case where $0 < s \leq 1/2$. By construction, if $m \in I$ and $F \in \mathcal{A}_I$, then $\|F_m\|_2 \geq H_I(m)$. Thus, if $I \cap \{m : J_m \leq J_n\} \neq \emptyset$, for all $F \in \mathcal{A}_I$,

$$
(\Gamma \xi^{-10\varepsilon I})^2 \varepsilon_n^2 \sum_{m \in I} 2^{-J_m} 1_{J_m \leq J_n} \leq \sum_{m \in I} \|F_m\|_2^2 1_{J_m \leq J_n}.
$$

Similarly if $I \cap \{m : J_m > J_n\} \neq \emptyset$, we have by construction for all $F \in \mathcal{A}_I$

$$
\gamma^2 \sum_{m \in I} 2^{-J_m(2s+1)} 1_{J_m > J_n} \leq \sum_{m \in I} \|F_m\|_2^2 1_{J_m > J_n}.
$$

Also of $0 \in I$, then $\Gamma^2 2^{-J_0} \varepsilon_n^2 \leq \|F_0\|_2^2$. Therefore Lemma 6.4 implies

$$
|C(S_{F_I}, S_{F_I})| \leq C \delta \mathbb{E}_L[S_{F_I}^2] + \delta \|F_0\|_2^2 1_{0 \in I} + \delta \sum_{m \in I} \|F_m\|_2^2.
$$

The conclusion follows since $\sum_{m \in I} \|F_m\|_2^2 \geq \mathbb{E}_L[S_{F_I}^2]$ by Proposition 7.1-(4).

Proof of Corollary 6.2. The proof follows immediately from Lemma 6.4, from the fact that $\sum_{m \in I} \|F_m\|_2^2 \geq \mathbb{E}_L[S_{F_I}^2]$ by Proposition 7.1-(4), and from the definition of $\mathcal{A}_I$. 

Adaptive Bayesian density estimation in sup-norm
7. Auxiliary results and remaining proofs

7.1. Relations between norms

In many places we need to relate norm of various functions. In this section we collect the propositions that serve this purpose.

Proposition 7.1. Let $J_0$ be chosen large enough. Then the following are true.

1. For all $m \in \mathbb{Z}_+$ and all $F \in \mathcal{F}_m$, $F = \sum_{(j,k) \in B_m} \langle F, \varphi_{j,k} \rangle (\varphi_{j,k} - \mathbb{E}_L[\varphi_{j,k}])$;
2. For all $m \in \mathbb{Z}_+$, all $F \in \mathcal{F}_m$, and all $(j,k) \notin B_0 \cup B_m \implies \langle F, \varphi_{j,k} \rangle = 0$.
3. There exist constants $C_1, C_2 > 0$ such that for all $m \in \mathbb{Z}_+$ and all $F \in \mathcal{F}_m$,
   $C_1 \sum_{v \in B_m} \langle F, \varphi_v \rangle^2 \leq \|F\|_2^2 \leq C_2 \sum_{v \in B_m} \langle F, \varphi_v \rangle^2$.
4. There exist constants $C_1, C_2 > 0$ such that for all $J \subseteq \mathbb{Z}_+$, for all collections $\{F_m : m \in J\}$, $C_1 \sum_{m \in J} \|F_m\|_2^2 \leq \mathbb{E}_L[\sum_{m \in J} F_m^2] \leq C_2 \sum_{m \in J} \|F_m\|_2^2$.
5. There exist constants $C_1, C_2 > 0$ such that $\sup_x \sum_k |\varphi_{j,k}(x)| \leq C_1 2^{j/2}$ for all $j \geq 0$, and $\sum_k |\mathbb{E}_L[\varphi_{j,k}]| \leq C_2 2^{j/2}$ for all $j \geq 0$. Consequently, $\sup_x \sum_k |\varphi_{j,k}(x) - \mathbb{E}_L[\varphi_{j,k}]| \leq 2 \max\{C_1, C_2\} 2^{j/2}$ for all $j \geq 0$.

Proposition 7.2. Suppose Assumption 1 is valid. There exists a universal constant $C > 0$ such that for all $I \subseteq \mathbb{Z}_+$,

$$\mathcal{E}_I^2 \geq C \left( \frac{\Gamma^2 \log(n)}{n} \sum_{m \in I} \xi^{-2\log(1-m^2)} \mathbf{1}_{m \leq j_m} + \gamma^2 \sum_{m \in I} 2^{-j_m(2\gamma + 1)} \mathbf{1}_{m > j_m} \right).$$

7.2. Proofs of the lemmas used in the guidelines of Section 3 and proof of the Corollary 3.1

Proof of Lemma 3.2. Write $g = \sum_{m \geq 0} F_m$ for simplicity. Then, remark that $\mathbb{E}_L[g] = 0$, and thus $\mathbb{E}_L[e^g] \geq \mathbb{E}_L[1 + g + \frac{1}{2} g^2 e^{-\|g\|_\infty}] = 1 + \frac{1}{2} \mathbb{E}_L[g^2] e^{-\|g\|_\infty} \geq 1$, and with the same argument $1 \leq \mathbb{E}_L[e^g] \leq 1 + \frac{1}{2} \mathbb{E}_L[g^2] e^{-\|g\|_\infty}$. It follows from equation (3.1) that $|L_F - L| \leq |g| + |\log \mathbb{E}_L[e^g]| = |g| + \log \mathbb{E}_L[e^g] \leq |g| + \frac{1}{2} \mathbb{E}_L[g^2] e^{\|g\|_\infty}$. Since $\mathbb{E}_L[g^2] \leq \|g\|_\infty^2$,

$$\|L_F - L\|_\infty \leq \left\| \sum_{m \geq 0} F_m \right\|_\infty \left( 1 + \frac{1}{2} \left\| \sum_{m \geq 0} F_m \right\|_\infty e^{\|\sum_{m \geq 0} F_m\|_\infty} \right) \leq \left\| \sum_{m \geq 0} F_m \right\|_\infty,$$

(7.1)
because by assumption $F \in \mathcal{V}_s$. Further, by Proposition 7.1,

$$ \left\| \sum_{m \geq 0} F_m \right\|_\infty \leq \sup_{x \in [0,1]} \sum_{m \geq 0} \sum_{(j,k) \in B_m} |\langle F_m, \varphi_{j,k} \rangle| |\varphi_{j,k}(x)| - \mathbb{E}_L [\varphi_{j,k}]| $$

$$ \leq \sup_{x \in [0,1]} \sum_{j \geq j_0} \sup_{m : j_m = j} \sup_{(j,k) \in B_m} |\langle F_m, \varphi_{j,k} \rangle| \| \varphi_{j,k}(x) - \mathbb{E}_L [\varphi_{j,k}]| $$

$$ \lesssim \sum_{j \geq j_0} \sup_{m : j_m = j} \| F_m \|_2 2^{j/2}. \tag{7.2} $$

The conclusion follows by combining equations (7.1) and (7.2). □

**Proof of Lemma 3.3.** We first establish that $(\mathcal{I}_I)_{I \subseteq \mathbb{Z}_+}$ is a partition of $\mathcal{F}$. Pick $F \in \mathcal{F}$ arbitrary. We want to show that there exists a unique $I \subseteq \mathbb{Z}_+$ such that $F \in \mathcal{I}_I$. We have the following two possibilities:

- If $\|F_0\|_2 \geq \rho_0 \Gamma$, choose $I = \{0\} \cup \{m \geq 1 : \|F_m\|_2 > \rho_m \Gamma \xi^{-1}, J_m \geq j_n\} \cup \{m \geq 1 : \|F_m\|_2 > \gamma 2^{-J_m(s+1/2)}, J_m > j_n\}$.
- If $\|F_0\|_2 < \rho_0 \Gamma$, choose $I = \{m \geq 1 : \|F_m\|_2 > \rho_m \Gamma, J_m \leq j_n\} \cup \{m \geq 1 : \|F_m\|_2 > \gamma 2^{-J_m(s+1/2)}, J_m > j_n\}$.

The index set $I$ is uniquely defined by $F$, and $F \in \mathcal{I}_I$. We now prove the second claim. Let $A := \bigcup_{I \neq \emptyset} \mathcal{I}_I$. We can decompose $A$ as $A = A_1 \cup A_2$ where $A_1 := \bigcup_{I \subseteq \mathbb{Z}_+, 0 \in I} \mathcal{I}_I$ and $A_2 := \bigcup_{I \subseteq \mathbb{Z}_+, 0 \notin I} \mathcal{I}_I$. Remark that $A_1 = \{F \in \mathcal{F} : \|F_0\|_2 \geq \rho_0 \Gamma\}$, and $A_2 = A_1^c \cap \{F \in \mathcal{F} : \exists m \geq 1, \|F_m\|_2 > H_1(m)\}$. Note that if $0 \in I$ then $A_1^c$ is empty, so

$$ A_2 = A_1^c \cap \{F \in \mathcal{F} : \exists m \geq 1, \|F_m\|_2 \geq \rho_m \Gamma \xi^{-1} 1_{J_m \leq j_n} + \gamma 2^{-J_m(s+1/2)} 1_{J_m > j_n}\}. $$

The conclusion follows since $A^c = A_1^c \cup A_2$ and since $\xi > 1$. □

**Proof of Corollary 3.1: the case where $s > 1/2$.** In view of Lemmas 3.2 and 3.3, it is sufficient to show that

$$ \sum_{j = j_0}^{j_n} \rho_m 2^{j/2} + \sum_{j > j_n} \gamma 2^{-j/s} \lesssim \bar{\varepsilon}^*_n(s). \tag{7.3} $$

But, $\gamma \sum_{j > j_n} 2^{-j/s} \lesssim \gamma 2^{-j_n/s} \lesssim (\gamma/\Gamma) 2^{-j_n/s} \bar{\varepsilon}^*_n(s)$ by the definition of $j_n$ in equation (3.2).

On the other hand, $\sum_{j = j_0}^{j_n} \gamma \rho_m 2^{j/2} \leq \Gamma \sqrt{\log(n)/n} 2^{j_n/2} \lesssim \Gamma (\gamma/\Gamma) 2^{-j_n/s} \bar{\varepsilon}^*_n(s)$, still by equation (3.2). □

**Proof of Corollary 3.1: the case where $0 < s \leq 1/2$.** As for the other case, it is enough to show that equation (7.3) holds true. In this case, $\gamma \sum_{j > j_n} 2^{-j/s} \lesssim \gamma 2^{-j_n/s} \lesssim \Gamma \bar{\varepsilon}^*_n(s)^2$ by equation (3.2). Also, $\sum_{j = j_0}^{j_n} \gamma \rho_m 2^{j/2} \leq \gamma j_n \bar{\varepsilon}^*_n(s) \lesssim \log(\Gamma \bar{\varepsilon}^*_n(s)/\gamma) \Gamma \bar{\varepsilon}^*_n(s) \leq \log(n) \bar{\varepsilon}^*_n(s)$, again by equation (3.2). □
7.3. Proofs of Propositions 7.1 and 7.2

Proof of Proposition 7.1, Item (1). By construction we know that there are numbers $a_{j,k} \in \mathbb{R}$ such that $F = \sum_{(j,k) \in B_m} a_{j,k}(\varphi_{j,k} - \mathbb{E}_L[\varphi_{j,k}])$. We note that if $m \geq 1$ the coefficients $(a_{j,k})$ are uniquely determined by $a_{j,k} = \langle F, \varphi_{j,k} \rangle$, because $J_m > J_0$ is large enough such that all $\langle \varphi_{j,k}, 1 \rangle = 0$ for all $(j, k) \in B_m$. Thus $(F, \varphi_{j,k}) = \sum_{(j',k') \in B_m} a_{j',k'}(\varphi_{j',k'}, \varphi_{j,k}) = a_{j,k}$, for any $(j, k) \in B_m$. This establishes the proof for $m \geq 1$. For $m = 0$, it is the case that $F_0$ is in the span of $\{\varphi_{j,k} : (j, k) \in B_0\}$ (because the constants are included in the span), and thus $F_0$ can be uniquely written as $F_0 = \sum_{(j,k) \in B_0} \langle F_0, \varphi_{j,k} \rangle \varphi_{j,k}$. But by construction, $\mathbb{E}_L[F_0] = 0$, so in fact $F_0 = \sum_{(j,k) \in B_0} \langle F_0, \varphi_{j,k} \rangle (\varphi_{j,k} - \mathbb{E}_L[\varphi_{j,k}])$. □

Proof of Proposition 7.1, Item (2). This follows from the Item (1) and because for $(j, k) \notin B_0$ and $J_0$ large enough, we have $\langle 1, \varphi_{j,k} \rangle = 0$. Therefore, it it the case that $(F, \varphi_{j,k}) = \sum_{(j',k') \in B_m} \langle F, \varphi_{j',k'} \rangle (\varphi_{j',k'}, \varphi_{j,k})$. By orthogonality of the wavelet basis, the previous is either 0 if $(j, k) \notin B_m$, or $(F, \varphi_{j,k})$ otherwise. □

Proof of Proposition 7.1, Item (3). The lower bound is immediate because $\|F\|_2^2 = \sum_{(j,k) \in V} (F, \varphi_{j,k})^2 \geq \sum_{(j,k) \in B_m} (F, \varphi_{j,k})^2$, so indeed $C_1 = 1$ works. For the upper bound, we note that because $\|L\|_{\infty} \lesssim 1$ we have $\|F_m\|_2^2 \lesssim \mathbb{E}_L[F_m^2]$, and by Item (1)

$$
\mathbb{E}_L[F_m^2] = \mathbb{E}_L\left[\left( \sum_{(j,k) \in B_m} (F, \varphi_{j,k}) \varphi_{j,k} - \mathbb{E}_L \left[ \sum_{(j,k) \in B_m} (F, \varphi_{j,k}) \varphi_{j,k} \right] \right)^2 \right]
\leq \mathbb{E}_L\left[ \left( \sum_{(j,k) \in B_m} (F, \varphi_{j,k}) \varphi_{j,k} \right)^2 \right]
\lesssim \sum_{(j,k) \in B_m} \langle F, \varphi_{j,k} \rangle^2
= \sum_{(j,k) \in B_m} \langle F, \varphi_{j,k} \rangle^2,
$$

where the last line follows by the orthogonality of the wavelet basis. □

Proof of Proposition 7.1, Item (4). We start with the upper bound, which follows from similar arguments than those of the Item (3). Indeed, recall that $\mathbb{E}_L[g^2] \asymp \|g\|_2^2$ for
all \( g \) because \( \| L \|_\infty \lesssim 1 \), hence

\[
\mathbb{E}_L \left[ \left( \sum_{m \in J} F_m \right)^2 \right] = \mathbb{E}_L \left[ \left( \sum_{m \in J} \sum_{(j,k) \in B_m} (F_m, \varphi_{j,k}) \varphi_{j,k} - \mathbb{E}_L \left[ \sum_{m \in J} \sum_{(j,k) \in B_m} (F_m, \varphi_{j,k}) \varphi_{j,k} \right] \right)^2 \right]
\leq \mathbb{E}_L \left[ \left( \sum_{m \in J} \sum_{(j,k) \in B_m} (F_m, \varphi_{j,k}) \varphi_{j,k} \right)^2 \right]
\lesssim \left\| \sum_{m \in J} \sum_{(j,k) \in B_m} (F_m, \varphi_{j,k}) \varphi_{j,k} \right\|^2
= \sum_{m \in J} \sum_{(j,k) \in B_m} \langle F_m, \varphi_{j,k} \rangle^2.
\]

Then, by the Item (2),

\[
\mathbb{E}_L \left[ \left( \sum_{m \in J} F_m \right)^2 \right] \lesssim \sum_{m \in J} \| F_m \|^2_2.
\quad (7.4)
\]

We now proceed with the lower bound. By the Item (1),

\[
\left\| \sum_{m \in J} F_m \right\|^2_2 = \sum_{(j,k) \in V} \left\langle \sum_{m \in J} F_{m'} \varphi_{j',k'} \right\rangle^2
\geq \sum_{m \geq 1} \sum_{(j,k) \in B_m} \left\langle \sum_{m' \in J} \sum_{(j',k') \in B_{m'}} \langle F_{m'}, \varphi_{j',k'} \rangle \varphi_{j',k'} - \mathbb{E}_L[\varphi_{j',k'}] \varphi_{j,k} \right\rangle^2
= \sum_{m \geq 1} \sum_{(j,k) \in B_m} \left\langle \sum_{m' \in J} \sum_{(j',k') \in B_{m'}} \langle F_{m'}, \varphi_{j',k'} \rangle \varphi_{j',k'} \varphi_{j,k} \right\rangle^2
= \sum_{m \geq 1} \sum_{(j,k) \in B_m} \langle F_{m'}, \varphi_{j,k} \rangle^2 1_{m \neq 0},
\]

where the third line follows because \( \langle 1, \varphi_{j,k} \rangle = 0 \) for all \( (j,k) \notin B_0 \), and the last line by orthogonality of the wavelet basis. Therefore by the Item (3) it must be the case that

\[
\mathbb{E}_L \left[ \left( \sum_{m \in J} F_m \right)^2 \right] \gtrsim \sum_{m \in J} \| F_m \|^2_2 1_{m \neq 0}.
\quad (7.5)
\]

The last display gives the proof in the case where \( 0 \notin J \). We now assume that \( 0 \in J \), which is a more delicate case. In this situation, we have that \( F_0 = \sum_{m \in J} F_m - \sum_{m \in J} F_m 1_{m \neq 0} \), and thus

\[
\mathbb{E}_L[F_0^2] \leq 2\mathbb{E}_L \left[ \left( \sum_{m \in J} F_m \right)^2 \right] + 2\mathbb{E}_L \left[ \left( \sum_{m \in J} F_m 1_{m \neq 0} \right)^2 \right]
\lesssim \mathbb{E}_L \left[ \left( \sum_{m \in J} F_m \right)^2 \right] + \sum_{m \in J} \| F_m \|^2_2 1_{m \neq 0}.
\quad (7.6)
\]
where the second line follows from the upper bound of equation (7.4) applied to the index set \( J \setminus \{0\} \). Combining equations (7.5) and (7.6),

\[
\|F_0\|_2^2 \lesssim \mathbb{E}_L[F_0^2] \lesssim \mathbb{E}_L \left[ \left( \sum_{m \in J} F_m \right)^2 \right].
\]  

(7.7)

Now if we combine the equations (7.5) and (7.7), we have indeed

\[
\mathbb{E}_L \left[ \left( \sum_{m \in J} F_m \right)^2 \right] \gtrsim \max \left\{ \sum_{m \in J} \|F_m\|_2^2 \mathbf{1}_{m \neq 0}, \|F_0\|_2^2 \right\}
\]

\[
\geq \frac{1}{2} \sum_{m \in J} \|F_m\|_2^2 \mathbf{1}_{m \neq 0} + \frac{1}{2} \|F_0\|_2^2
\]

\[
= \frac{1}{2} \sum_{m \in J} \|F_m\|_2^2.
\]

Proof of Proposition 7.1, Item (5). The first claim is a well-known localization properties of the wavelet basis. The second fact follows because \( \mathbb{E}_L[\varphi_{j,k}] \leq \|p_L\|_\infty \|\varphi_{j,k}\|_1 \lesssim \|p_L\|_\infty 2^{-j/2} \), and because there are no more than \( 2^j \) wavelets at each level \( j \geq 0 \). The third fact is obvious.

Proof of Proposition 7.2. From the definition of \( \mathcal{E}_I \) and from Proposition 7.1-(4), it is immediate that \( \mathcal{E}_I^2 \gtrsim \sum_{m \in I} H_1(m)^2 \). If \( s > 1/2 \) then the result is immediate. In case \( 0 < s \leq 1/2 \), then we note that by definition of \( j_n \)

\[
\gamma 2^{-j_n(s+1/2)} \geq \Gamma 2^{-j_n/2} \varepsilon_n^*(s) \implies 2^{-j_n} \geq \left( \frac{\Gamma}{\gamma} \right)^{\frac{1}{2}} \left( \frac{\log(n)}{n} \right)^{\frac{1}{2s+1}}.
\]

Therefore,

\[
\sum_{m \in I} \rho_m^2 \mathbf{1}_{J_m \leq j_n} = \sum_{m \in I} 2^{-j_m} \varepsilon_n^*(s)^2 \mathbf{1}_{J_m \leq j_n}
\]

\[
\geq 2^{-j_n} \varepsilon_n^*(s)^2 \sum_{m \in I} \mathbf{1}_{J_m \leq j_n}
\]

\[
= \left( \frac{\Gamma}{\gamma} \right)^{\frac{1}{2}} \cdot \frac{\log(n)}{n} \sum_{m \in I} \mathbf{1}_{J_m \leq j_n}
\]

\[
\geq \frac{\log(n)}{n} \sum_{m \in I} \mathbf{1}_{J_m \leq j_n},
\]

where the last line is true under Assumption 1.
Acknowledgements

This work was supported by U.S. Air Force Office of Scientific Research grant #FA9550-15-1-0074. The author also thanks Daniel M. Roy for helpful discussions and the opportunity to work on this project.

Supplementary Material

Supplementary Material: Adaptive Bayesian density estimation in sup-norm (). This supplementary material contains additional proofs and results for the spike-and-slab log-density prior. In particular, it contains the missing proofs of Theorem 2.1 and Lemmas 3.1, 5.1, 5.2 and 6.4.

References

Andrew Barron, Mark J Schervish, and Larry Wasserman. The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27(2):536–561, 1999.
Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
Lawrence D Brown, Cun-Hui Zhang, et al. Asymptotic nonequivalence of nonparametric experiments when the smoothness index is 1/2. *The Annals of Statistics*, 26(1):279–287, 1998.

Ismaël Castillo. On Bayesian supremum norm contraction rates. *Ann. Statist.*, 42(5):2058–2091, 2014.

Ismaël Castillo. Pólya tree posterior distributions on densities. In *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, volume 53, pages 2074–2102. Institut Henri Poincaré, 2017.

Ismaël Castillo and Romain Mismer. Spike and slab pólya tree posterior distributions. *arXiv preprint arXiv:1911.12106*, 2019.

Albert Cohen, Ingrid Daubechies, and Pierre Vial. Wavelet bases on the interval and fast algorithms. *Journal of Applied and Computational Harmonic Analysis*, 1(1):54–81, 1993.

David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, pages 508–539, 1996.

Subhashis Ghosal and Aad van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):192–223, 02 2007a.

Subhashis Ghosal and Aad van der Vaart. Posterior convergence rates of dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723, 04 2007b. URL https://doi.org/10.1214/009053606000001271.

Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, Apr 2000.

Evarist Giné and Richard Nickl. Rates of contraction for posterior distributions in $L^r$-metrics, $1 \leq r \leq \infty$. *Ann. Statist.*, 39(6):2883–2911, 2011.
Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2016.

Alexander Goldenshluger and Oleg Lepski. On adaptive minimax density estimation on $\mathbb{R}^d$, *Probability Theory and Related Fields*, 159(3-4):479–543, 2014.

Wolfgang Härdle, Gerard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. *Wavelets, approximation, and statistical applications*. Springer Science & Business Media, 2000.

Rafail Zalmanovich Hasminskii. A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theory of Probability & Its Applications*, 23(4):794–798, 1979.

Marc Hoffmann, Judith Rousseau, and Johannes Schmidt-Hieber. On adaptive posterior concentration rates. *Ann. Statist.*, 43(5):2259–2295, 2015.

Il’dar Abdullovich Ibragimov and Rafail Zalmanovich Hasminskii. On estimate of the density function. *Zapiski Nauchnykh Seminarov POMI*, 98:61–85, 1980.

Antonio Lijoi, Igor Prünster, and Stephen G Walker. On consistency of nonparametric normal mixtures for bayesian density estimation. *Journal of the American Statistical Association*, 100(472):1292–1296, 2005.

Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

David Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.

Vincent Rivoirard and Judith Rousseau. Bernstein–von mises theorem for linear functionals of the density. *Ann. Statist.*, 40(3):1489–1523, 06 2012. . URL https://doi.org/10.1214/12-AOS1004.

Lorraine Schwartz. On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26, 1965.

Catia Scricciolo. Adaptive Bayesian density estimation in $L^p$ -metrics with Pitman–Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Analysis*, 9(2):475–520, Jun 2014.

Xiaotong Shen, Larry Wasserman, et al. Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714, 2001.

Stephen G Walker, Antonio Lijoi, and Igor Prünster. On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.*, pages 738–746, 2007.

William Weimin Yoo and Subhashis Ghosal. Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.*, 44(3):1069–1102, 2016.

William Weimin Yoo, Judith Rousseau, and Vincent Rivoirard. Adaptive supremum norm posterior contraction: Spike-and-slab priors and anisotropic Besov spaces. *arXiv preprint arXiv:1708.01909*, 2017.