Bayesian Analysis (2012) **7**, Number 2, pp. 311–334

Posterior Concentration Rates for Infinite Dimensional Exponential Families

Vincent Rivoirard* and Judith Rousseau†

Abstract. In this paper we derive adaptive non-parametric rates of concentration of the posterior distributions for the density model on the class of Sobolev and Besov spaces. For this purpose, we build prior models based on wavelet or Fourier expansions of the logarithm of the density. The prior models are not necessarily Gaussian.

Keywords: Bayesian non-parametric, rates of convergence, adaptive estimation, wavelets and Fourier Bases, Sobolev and Besov balls

1 Introduction

Frequentist properties of Bayesian nonparametric procedures have been increasingly studied in the last decade, following the seminal papers of Barron et al. (1999) and Ghosal et al. (2000) which established general conditions on the prior and on the true distribution to obtain posterior consistency for the former and posterior concentration rates for the latter. Consistency of the posterior distribution is admitted as a minimal requirement, both from a subjectivist and an objectivist view-point, see Diaconis and Freedman (1986). Studying posterior concentration rates allows for more refined results, in particular it helps in understanding some aspects of the prior and can be used to compare Bayesian procedures and to contrast them with frequentist procedures. In the frequentist nonparametric literature the usual optimality criterion that is considered is the minimaxity over a given functional class, typically indexed by a smoothness index, for instance a Sobolev or a Besov ball. However the smoothness is generally unknown a priori and it is important to construct estimators which adapt to these smoothness indexes, i.e. which do not depend on these indexes but which achieve the optimal rate within each class.

In this paper we derive posterior concentration rates for density estimation, when the prior is based on wavelet or Fourier expansions of the logarithm of the density. We consider $X^n = (X_1, ..., X_n)$ which, given a distribution $\mathbb{P}$ with a compactly supported density f with respect to the Lebesgue measure, are independent and identically distributed according to $\mathbb{P}$. Without loss of generality we assume that for any i, $X_i \in [0, 1]$ and we set

$$\mathcal{F} = \left\{ f : [0,1] \to \mathbb{R}^+ \ \text{ s.t. } \ \int_0^1 f(x)dx = 1 \right\}.$$

*CEREMADE, Université Paris Dauphine and INRIA project-team CLASSIC, Paris, France, Vincent.Rivoirard@dauphine.fr

†CREST-ENSAE and Université Paris Dauphine, Malakoff Cedex, France, rousseau@ensae.fr

 DOI:10.1214/12-BA710

As already mentioned, we restrict our attention to families of priors on $\mathcal{F}$ built from Fourier and wavelet expansions of $\log f$ assumed to be square-integrable. Wavelets are localized in both time and frequency whereas the standard Fourier basis is only localized in frequency. We recall that Fourier bases constitute unconditional bases of periodized Sobolev spaces W^γ where γ is the smoothness parameter. Wavelet expansions of any periodized function h take the following form:

$$h(x) = \theta_{-10}\mathbb{1}_{[0,1]}(x) + \sum_{j=0}^{+\infty}\sum_{k=0}^{2^j-1}\theta_{jk}\varphi_{jk}(x), \quad x \in [0,1]$$

where $\theta_{-10} = \int_0^1 h(x)dx$, $\theta_{jk} = \int_0^1 h(x)\varphi_{jk}(x)dx$ and $\mathbb{1}_A$ is the indicatrix function : $\mathbb{1}_A(x) = 1$ if $x \in A$ and $\mathbb{1}_A(x) = 0$ if $x \notin A$. We recall that the functions φ_{jk} are obtained by periodizing dilations and translations of a mother wavelet φ that can be assumed to be compactly supported. Unlike the Fourier basis, under standard properties of φ involving its regularity and its vanishing moments, wavelet bases constitute unconditional bases of Besov spaces $\mathcal{B}^\gamma_{p,q}$. We refer the reader to Härdle et al. (1998) for a good introduction to wavelets and to Section 5.2 for more details on Sobolev and Besov spaces. We just mention that the scale of Besov spaces includes Sobolev spaces: $W^\gamma = \mathcal{B}^\gamma_{2,2}$. In the sequel, to shorten notation, we use a unified framework including Fourier and wavelet bases. The considered orthonormal basis will be denoted $\Phi = (\phi_\lambda)_{\lambda\in\mathbb{N}}$, where $\phi_0 = \mathbb{1}_{[0,1]}$ and

- for the Fourier basis, if $\lambda \geq 1$,

$$\phi_{2\lambda-1}(x) = \sqrt{2}\sin(2\pi\lambda x), \quad \phi_{2\lambda}(x) = \sqrt{2}\cos(2\pi\lambda x),$$

- for the wavelet basis, if $\lambda = 2^j + k$, with $j \in \mathbb{N}$ and $k \in \{0, \dots, 2^j - 1\}$,

$$\phi_\lambda = \varphi_{jk}.$$

Here and in the sequel, $\mathbb{N}$ denotes the set of non negative integers, and $\mathbb{N}^*$ the set of positive integers. Now, the decomposition of each periodized function $h \in \mathbb{L}_2[0,1]$, the set of square-integrable functions on $[0,1]$ with respect to the Lebesgue measure, on $(\phi_\lambda)_{\lambda\in\mathbb{N}}$ is written as follows:

$$h(x) = \sum_{\lambda\in\mathbb{N}}\theta_\lambda\phi_\lambda(x), \quad x \in [0,1],$$

where $\theta_\lambda = \int_0^1 h(x)\phi_\lambda(x)dx$.

We use such expansions to build non-parametric priors on $\mathcal{F}$ in the following way: For any $k \in \mathbb{N}^*$, we set

$$\mathcal{F}_k = \left\{ f_\theta = \exp\left(\sum_{\lambda=1}^k \theta_\lambda\phi_\lambda - c(\theta)\right) \text{ s.t. } \theta \in \mathbb{R}^k \right\},$$

where

$$c(\theta) = \log\left(\int_0^1 \exp\left(\sum_{\lambda=1}^{k} \theta_\lambda \phi_\lambda(x)\right) dx\right). \tag{1}$$

So, we define a prior π on the set $\mathcal{F}_\infty = \cup_k \mathcal{F}_k \subset \mathcal{F}$ by defining a prior p on $\mathbb{N}^*$ and then, once k is chosen, we fix a prior π_k on $\mathcal{F}_k$. Such priors are often considered in the Bayesian non-parametric literature and our study can be strongly connected to the well-known papers by van der Vaart and van Zanten (2008) (see Section 4 for a detailed comparison with our results) and by Shen and Wasserman (2001) who considered the problem of estimating a regression function decomposed on an orthonormal basis. The prior model proposed by Shen and Wasserman is a special case of the prior described in Section 2. However, they assume that all the functions of the basis are uniformly bounded by a constant. This is of course satisfied by the Fourier basis but not by wavelet bases. Furthermore, once the prior is fixed, depending on a hyperparameter p, rates are only achieved on Sobolev balls of regularity p. So, adaptation is not handled by Shen and Wasserman (2001). See below for specific results we obtain in our paper. The family of priors defined in Section 2 has also been used in the infinite-means model (equivalently in the white noise model) by Zhao (2000) where minimax but non adaptive rates were obtained for the $\mathbb{L}_2$-risk (see for instance Theorem 6.1 of Zhao (2000)). We also mention the special case of log-spline priors studied by Ghosal et al. (2000) and the prior model based on Legendre polynomials considered by Verdinelli and Wasserman (1998).

Our results are concerned with adaptation that has been partially studied in the literature. Let us cite Scricciolo (2006) who considers infinite dimensional exponential families and derives minimax and adaptive posterior concentration rates. Her results differ from ours in two main aspects. Firstly she restricts her attention to the case of Sobolev spaces and Fourier basis, whereas we consider Besov spaces. Secondly she obtains adaptivity by putting a prior on the smoothness of the Sobolev class whereas we obtain adaptivity by constructing a prior on the size k of the parametric spaces, which to our opinion is a more natural approach. Moreover Scricciolo (2006) merely considers Gaussian priors. Also related to adaptation are the works of Huang (2004) and Ghosal et al. (2008) who derive a general framework to obtain adaptive posterior concentration rates, the former applies her results to the Haar basis case. The limitation in her case, apart from the fact that she considers the Haar basis and no other wavelet basis is that she constrains the θ_λ's in each k-dimensional model to belong to a ball with fixed radius.

In this paper we give general conditions on families of priors briefly described previously to obtain adaptive minimax rates (up to a $\log n$ term) for the estimation of f. In the next section we introduce the prior model and Theorem 2.1 gives the posterior rates on Sobolev balls $W^\gamma(R)$ and Besov balls $\mathcal{B}^\gamma_{p,q}(R)$ $p \geq 2$. Section 3 gives the proof of Theorem 2.1 and Section 4 contains conclusions we can draw from our results. We state in the Appendix a result useful for establishing concentration rates. The Appendix also contains the proof of a technical result and an overview of Sobolev and Besov spaces.

Notation: In the sequel, we denote by $\ell_n(f)$ the log-likelihood associated with the density f. The Kullback-Leibler divergence and the Hellinger distance between two positive densities f_1 and f_2 will be respectively denoted $K(f_1, f_2)$ and $h(f_1, f_2)$. We recall that

$$K(f_1, f_2) = \int_0^1 f_1(x) \log\left(\frac{f_1(x)}{f_2(x)}\right) dx \tag{2}$$

and

$$h(f_1, f_2) = \left[\int \left(\sqrt{f_1(x)} - \sqrt{f_2(x)}\right)^2 dx\right]^{1/2}. \tag{3}$$

In the sequel, we shall also use

$$V(f_1, f_2) = \int_0^1 f_1(x) \left(\log\left(\frac{f_1(x)}{f_2(x)}\right)\right)^2 dx. \tag{4}$$

The minimum between reals a and b is denoted by $a \wedge b$. Let $\mathbb{P}_0$ be the true distribution of the observations X_i whose density and cumulative distribution function are respectively denoted f_0 and F_0. For any positive sequence u_n, $x_n = o_{\mathbb{P}_0}(u_n)$ means that x_n/u_n converges to 0 in probability with repect to the distribution $\mathbb{P}_0$. We denote by $\|.\|_\gamma$ and $\|.\|_{\gamma,p,q}$ the norms associated with W^γ and $\mathcal{B}^\gamma_{p,q}$ respectively. The integer r will denote the number of vanishing moments of the wavelet basis. When the Fourier basis is considered, we set $r = +\infty$.

2 Prior models and concentration rates

Given $\beta > 1/2$, the prior p on k satisfies one of the following conditions:

[Case (PH)] There exist two positive constants c_1 and c_2 and $s \in \{0, 1\}$ such that for any $k \in \mathbb{N}^*$,

$$\exp\left(-c_1 k L(k)\right) \leq p(k) \leq \exp\left(-c_2 k L(k)\right),$$

where $L(x) = (\log x)^s$.

[Case (D)] If $k_n^* = \lfloor n^{1/(2\beta+1)} \rfloor$, i.e. the largest integer smaller than $n^{1/(2\beta+1)}$,

$$p(k) = \delta_{k_n^*}(k),$$

where $\delta_{k_n^*}$ denotes the Dirac mass at the point k_n^*.

Conditionally on k the prior π_k on $\mathcal{F}_k$ is defined by

$$\frac{\theta_\lambda}{\sqrt{\tau_\lambda}} \overset{iid}{\sim} g, \quad \tau_\lambda = \tau_0 \lambda^{-2\beta} \quad 1 \leq \lambda \leq k,$$

where τ_0 is a positive constant and g is a continuous density on $\mathbb{R}$ such that for any x,

$$A_* \exp\left(-\tilde{c}_* |x|^{p_*}\right) \leq g(x) \leq B_* \exp\left(-c_* |x|^{p_*}\right),$$

where p_*, A_*, B_*, $\tilde{c}_*$ and c_* are positive constants.

Observe that the prior is not necessarily Gaussian since we allow for densities g to have different tails. In the Dirac case (D), the prior on k is non random. For the case (PH), $L(x) = \log(x)$ typically corresponds to a Poisson prior on k and the case $L(x) = 1$ typically corresponds to geometric priors. The density g can be for instance the Laplace or the Gaussian density.

Assume that f_0 is 1-periodic and $f_0 \in \mathcal{F}_\infty$. Let $\Phi = (\phi_\lambda)_{\lambda\in\mathbb{N}}$ be one of the bases introduced in Section 1, then there exists a sequence $\theta_0 = (\theta_{0\lambda})_{\lambda\in\mathbb{N}^*}$ such that

$$f_0(x) = \exp\left(\sum_{\lambda\in\mathbb{N}^*} \theta_{0\lambda}\phi_\lambda(x) - c(\theta_0)\right).$$

We have the following result.

Theorem 2.1. *We assume that* $\|\log(f_0)\|_\infty < \infty$ *and* $\log(f_0) \in \mathcal{B}^\gamma_{p,q}(R)$ *with* $p \geq 2$, $1 \leq q \leq \infty$ *and* $1/2 < \gamma < r+1$ *is such that*

$$\beta < 1/2 + \gamma \quad \textit{if} \quad p_* \leq 2 \qquad \textit{and} \qquad \beta < \gamma + 1/p_* \quad \textit{if} \quad p_* > 2.$$

Then,

$$\mathbb{P}^\pi\left\{f_\theta \quad s.t. \quad h(f_0, f_\theta) \leq \sqrt{\frac{\log n}{L(n)}}\epsilon_n | X^n\right\} = 1 + o_{\mathbb{P}_0}(1), \tag{5}$$

and

$$\mathbb{P}^\pi\left\{f_\theta \quad s.t. \quad \|\theta_0 - \theta\|_{\ell_2} \leq \log n \sqrt{\frac{\log n}{L(n)}}\epsilon_n | X^n\right] = 1 + o_{\mathbb{P}_0}(1), \tag{6}$$

where in the case (PH),

$$\epsilon_n = \epsilon_0 \left(\frac{\log n}{n}\right)^{\frac{\gamma}{2\gamma+1}},$$

in the case (D), $L(n) = 1$,

$$\epsilon_n = \epsilon_0 (\log n) n^{-\frac{\beta}{2\beta+1}}, \quad \textit{if } \gamma \geq \beta$$

$$\epsilon_n = \epsilon_0 n^{-\frac{\gamma}{2\beta+1}}, \quad \textit{if } \gamma < \beta$$

and ϵ_0 *is a constant large enough.*

If the density g only satisfies a tail condition of the form

$$g(x) \leq C_g |x|^{-p_*},$$

where C_g is a constant and $p_* > 1$, then, in the case (PH), if $\gamma > 1$, the rates defined by (5) and (6) remain valid.

Note that in the case (PH) the posterior concentration rate is, up to a $\log n$ term, the minimax rate of convergence, whereas in the case (D) the minimax rate is achieved only when $\gamma = \beta$. The interpretation is the following: In the case (PH), k is random, which allows one to determine the appropriate space $\mathcal{F}_k$ by using the data, i.e. the Bayesian procedure automatically adapts to the unknown approximation space containing the unknown signal. In particular, if f_0 belongs to $\mathcal{B}^{\gamma}_{p,q}$ then the "optimal" approximation space $\mathcal{F}_k$ corresponds to $k = O\left(n^{1/(2\gamma+1)}\right)$. Roughly speaking, this optimal k is obtained by some trade-off between minimizing the Kullback-Leibler divergence between f_0 and $\mathcal{F}_k$ and maximizing the prior mass of neighbourhoods of f_0 in $\mathcal{F}_k$.

A glance at the proof of Theorem 2.1 allows us to go further. We have the following result enhancing our results with respect to adaptation.

Corollary 2.1. *Let $p^* > 0$ and $1/2 < \beta \leq (1/2 + 1/p^*) \wedge 1$, then for all $R_0 > 0$ and $\gamma_0 > 1/2$, the posterior distribution associated to the prior (PH) achieves the adaptive minimax rate up to a logarithmic term on the whole class*

$$\mathcal{C}(\gamma_0, R_0) = \left\{\mathcal{B}^{\gamma}_{p,q}(R) : \gamma_0 \leq \gamma < r+1,\ p \geq 2,\ 1 \leq q \leq \infty,\ 0 < R \leq R_0\right\},$$

meaning that there exists $\epsilon_0 > 0$ such that

$$\sup_{\substack{\gamma_0 \leq \gamma < r+1 \\ 0 < R \leq R_0}} \ \sup_{\substack{p \geq 2 \\ q \geq 1}} \ \sup_{f_0 \in \mathcal{B}^{\gamma}_{p,q}(R)} \mathbb{E}_{f_0}\left[\mathbb{P}^{\pi}\left\{h(f_0, f_\theta) \leq \epsilon_0 \sqrt{\frac{\log n}{L(n)}} \left(\frac{\log n}{n}\right)^{\frac{\gamma}{2\gamma+1}} |X^n\right\}\right] = o(1).$$

A similar result holds for the ℓ_2-loss.

3 Proof of Theorem 2.1

In the sequel, C denotes a generic positive constant whose value is of no importance and may change from line to line. To simplify some expressions, we omit at some places the integer part $\lfloor \cdot \rfloor$. Remember that $L(n) = 1$ for any n in the case (D). To prove Theorem 2.1 it is enough to verify conditions (A), (B) and (C) of Theorem 5.1, which is presented in the appendix. We consider $(\Lambda_n)_n$ the increasing sequence of subsets of $\mathbb{N}^*$ defined by $\Lambda_n = \{1, 2, \dots, l_n\}$ with $l_n \in \mathbb{N}^*$ (defined below). For any n, we set:

$$\mathcal{F}^*_n = \left\{f_\theta \in \mathcal{F}_{l_n} \ \text{ s.t. } \ f_\theta = \exp\left(\sum_{\lambda \in \Lambda_n} \theta_\lambda \phi_\lambda - c(\theta)\right),\ \|\theta\|_{\ell_2} \leq w_n\right\},$$

with

$$w_n = \exp(w_0 n^{\rho} (\log n)^q), \quad \rho > 0,\ q \in \mathbb{R}.$$

For l_0 a constant,

- $\epsilon_n = \epsilon_0 n^{-\frac{\gamma}{2\gamma+1}} (\log n)^{\frac{\gamma}{2\gamma+1}}$ and we set $l_n = \frac{l_0 n \epsilon_n^2}{L(n)}$ in the case (PH),
- $\epsilon_n = \epsilon_0 (\log n)^{1_{\{\gamma \geq \beta\}}} n^{-\frac{\beta \wedge \gamma}{2\beta+1}}$ and we set $l_n = l_0 k^*_n = l_0 n^{\frac{1}{2\beta+1}}$ in the case (D).

Condition (A) Since $\beta > 1/2$, $\sum_\lambda \tau_\lambda < \infty$ and for the sake of simplicity, without loss of generality, we assume that $\sum_\lambda \tau_\lambda \leq 1$. Using the tail assumption on g,

$$\begin{aligned}
\pi\left\{\mathcal{F}_n^{*c}\right\} &\leq \sum_{\lambda > l_n} p(\lambda) + \mathbb{P}^\pi\left\{\sum_{\lambda \leq l_n} \theta_\lambda^2 > w_n^2\right\} \\
&\leq C\exp\left(-c_2 l_n L(l_n)\right) + \sum_{\lambda \leq l_n} \mathbb{P}^\pi\left\{\frac{\theta_\lambda^2}{\tau_\lambda} > w_n^2\right\} \\
&\leq C\exp\left(-c_2 l_0 n\epsilon_n^2\right) + \sum_{\lambda \leq l_n} \mathbb{P}^\pi\left\{\exp\left(\frac{c_*|\theta_\lambda|^{p_*}}{2\tau_\lambda^{p_*/2}}\right) > \exp\left(\frac{c_* w_n^{p_*}}{2}\right)\right\} \\
&\leq C\exp\left(-c_2 l_0 n\epsilon_n^2\right) + C l_n \exp\left(-\frac{c_* w_n^{p_*}}{2}\right) \\
&\leq C\exp\left(-c_2 l_0 n\epsilon_n^2\right) + C\exp\left(-n^H\right)
\end{aligned}$$

for any positive $H > 0$. For the second line, we have used

$$\mathbb{P}^\pi\left\{\sum_{\lambda \leq l_n} \theta_\lambda^2 > w_n^2\right\} \leq \mathbb{P}^\pi\left\{\sum_{\lambda \leq l_n} \theta_\lambda^2 > \sum_{\lambda \leq l_n} \tau_\lambda w_n^2\right\} \leq \sum_{\lambda \leq l_n} \mathbb{P}^\pi\left\{\theta_\lambda^2 > \tau_\lambda w_n^2\right\}$$

and for the fourth line the Markov inequality combined with

$$\frac{\theta_\lambda}{\sqrt{\tau_\lambda}} \stackrel{iid}{\sim} g$$

and

$$g(x) \leq B_* \exp\left(-c_*|x|^{p_*}\right).$$

Hence,

$$\pi\left\{\mathcal{F}_n^{*c}\right\} \leq C\exp\left(-c_2 l_0 n\epsilon_n^2\right)$$

and Condition (A) is proved for l_0 large enough.

Condition (B) In the framework of Theorem 5.1, we bound $H_{n,j}$ for any $j \geq 1$. Actually, since the Hellinger distance is uniformly bounded by $\sqrt{2}$, we can restrict our attention to the case $j \leq \sqrt{2}\epsilon_n^{-1}$. For this purpose, we show that the Hellinger distance between two functions of $\mathcal{F}_n^*$ is related to the ℓ_2-distance of the associated coefficients. Let

$$\tilde{c}_1 = \frac{1}{4c_{1,\Phi}},$$

where $c_{1,\Phi}$ is defined in Lemma 5.1, in the appendix. We consider f_θ and $f_{\theta'}$ belonging to $\mathcal{F}_n^*$ with

$$f_\theta = \exp\left(\sum_{\lambda \in \Lambda_n} \theta_\lambda \phi_\lambda - c(\theta)\right), \quad f_{\theta'} = \exp\left(\sum_{\lambda \in \Lambda_n} \theta'_\lambda \phi_\lambda - c(\theta')\right).$$

We show that for $j \leq \sqrt{2}\epsilon_n^{-1}$,

$$\|\theta - \theta'\|_{\ell_2} \leq \tilde{c}_1 j \epsilon_n l_n^{-1/2} \Rightarrow h(f_\theta, f_{\theta'}) \leq \frac{j\epsilon_n}{2}. \tag{7}$$

For this purpose, we apply Lemma 5.1 with $K_n = \Lambda_n$ and $k_n = l_n$ and if $\|\theta' - \theta\|_{\ell_2} \leq \tilde{c}_1 j \epsilon_n l_n^{-1/2}$, then,

$$\left\| \sum_{\lambda \in \Lambda_n} (\theta'_\lambda - \theta_\lambda)\phi_\lambda \right\|_\infty \leq c_{1,\Phi}\sqrt{l_n}\|\theta' - \theta\|_{\ell_2} \leq c_{1,\Phi}\tilde{c}_1 j \epsilon_n \leq \sqrt{2} c_{1,\Phi}\tilde{c}_1 \leq 1.$$

Now, straightforward computations lead to

$$\begin{aligned} |c(\theta') - c(\theta)| &= \left| \log\left(\int_0^1 f_\theta(x) \exp\left(\sum_{\lambda \in \Lambda_n} (\theta'_\lambda - \theta_\lambda)\phi_\lambda(x) \right) dx \right) \right| \\ &\leq \left| \log\left(\int_0^1 f_\theta(x) \exp\left(\| \sum_{\lambda \in \Lambda_n} (\theta'_\lambda - \theta_\lambda)\phi_\lambda \|_\infty \right) dx \right) \right| \\ &= \| \sum_{\lambda \in \Lambda_n} (\theta'_\lambda - \theta_\lambda)\phi_\lambda \|_\infty. \end{aligned}$$

Then,

$$\begin{aligned} h^2(f_\theta, f_{\theta'}) &= \int_0^1 f_\theta(x) \left(\exp\left(\frac{1}{2} \sum_{\lambda \in \Lambda_n} (\theta'_\lambda - \theta_\lambda)\phi_\lambda(x) + \frac{1}{2}\left(c(\theta) - c(\theta')\right) \right) - 1 \right)^2 dx \\ &\leq \int_0^1 f_\theta(x) \left(\exp\left(\| \sum_{\lambda \in \Lambda_n} (\theta'_\lambda - \theta_\lambda)\phi_\lambda \|_\infty \right) - 1 \right)^2 dx \\ &\leq 4\| \sum_{\lambda \in \Lambda_n} (\theta_\lambda - \theta'_\lambda)\phi_\lambda \|_\infty^2 \\ &\leq 4c_{1,\Phi}^2 l_n \|\theta - \theta'\|_{\ell_2}^2 \leq 4c_{1,\Phi}^2 \tilde{c}_1^2 j^2 \epsilon_n^2 \leq \frac{j^2\epsilon_n^2}{4}, \end{aligned} \tag{8}$$

where we have used the (rough) bound : $\exp(x) - 1 \leq 2x$ for any $x \in [0,1]$. This proves (7). By identifying θ and f_θ, it means that every covering of $S_{n,j}$ by ℓ_2-balls of radius $\tilde{c}_1 j \epsilon_n l_n^{-1/2}$ provides a covering of $S_{n,j}$ by Hellinger-balls of radius $j\epsilon_n/2$. Then, we use the following lemma proved in Section 5.3.

Lemma 3.1. *We assume that* $\log(f_0) \in \mathcal{B}^\gamma_{p,q}(R)$ *with* $p \geq 2$, $1 \leq q \leq \infty$ *and* $1/2 < \gamma < r + 1$. *We set* $c_0 = \inf_{x \in [0,1]} f_0(x) > 0$. *There exists a positive constant* $c \leq 1/2$ *depending on* β, γ, R *and* Φ *such that, for* $j \geq 1$, *if*

$$(j+1)^2 \epsilon_n^2 l_n \leq c$$

then for $f_\theta \in S_{n,j}$

$$\|\theta_0 - \theta\|_{\ell_2}^2 \leq \frac{1}{c}(\log n)^2 h^2(f_0, f_\theta).$$

The lemma shows that if $(j+1)^2\epsilon_n^2 l_n \leq c$ then, by identifying θ and f_θ, $S_{n,j}$ is included into the ℓ_2-ball centered at θ_0 with radius $c^{-1/2}(j+1)\epsilon_n \log n$. Therefore, in this case, combining (8) and Lemma 3.1, we obtain

$$H_{n,j} \leq \log\left(\left(C\frac{(j+1)\epsilon_n \log n}{j\epsilon_n l_n^{-1/2}}\right)^{l_n}\right) \leq l_n \log n,$$

for n large enough. Then, we have $H_{n,j} \leq Knj^2\epsilon_n^2$ as soon as $j \geq J_{0,n} = \sqrt{j_0 \log n / L(n)}$, where j_0 is a constant and condition (B) is satisfied for such j's.

When $(j+1)^2\epsilon_n^2 l_n > c$, then since for $f_\theta \in \mathcal{F}_n^*$, $\|\theta\|_{\ell_2} \leq w_n$,

$$H_{n,j} \leq \log\left(\left(C\frac{w_n}{j\epsilon_n l_n^{-1/2}}\right)^{l_n}\right) \leq 2l_n \log(w_n) \leq 2w_0 l_n n^\rho (\log n)^q,$$

for n large enough. Choosing w_0, q and ρ small enough such that $l_n^2(\log n)^q \leq n^{1-\rho}$, implies $H_{n,j} \leq Knj^2\epsilon_n^2$ and condition (B) is satisfied for such j's.

Condition (C) Let $k_n \in \mathbb{N}^*$ increasing to ∞ and $K_n = \{1, ..., k_n\}$, define

$$A(u_n) = \left\{\theta \text{ s.t. } \theta_\lambda = 0 \text{ for every } \lambda \notin K_n \text{ and } \sum_{\lambda \in K_n} (\theta_{0\lambda} - \theta_\lambda)^2 \leq u_n^2\right\},$$

where u_n goes to 0 such that

$$\sqrt{k_n} u_n \to 0. \tag{9}$$

We define for any λ,

$$\beta_\lambda(f_0) = \int_0^1 \phi_\lambda(x) f_0(x) dx.$$

Denote

$$f_{0K_n} = \exp\left(\sum_{\lambda \in K_n} \theta_{0\lambda}\phi_\lambda(x) - c(\theta_{0K_n})\right), \quad f_{0\bar{K_n}} = \exp\left(\sum_{\lambda \notin K_n} \theta_{0\lambda}\phi_\lambda(x) - c(\theta_{0\bar{K_n}})\right).$$

We have

$$\begin{aligned} K(f_0, f_{0K_n}) &= \sum_{\lambda \notin K_n} \theta_{0\lambda}\beta_\lambda(f_0) + c(\theta_{0K_n}) - c(\theta_0) \\ &= \sum_{\lambda \notin K_n} \theta_{0\lambda}\beta_\lambda(f_0) + \log\left(\int_0^1 f_0(x) e^{-\sum_{\lambda \notin K_n} \theta_{0\lambda}\phi_\lambda(x)} dx\right). \end{aligned}$$

Using inequality (15) of Lemma 5.1, we obtain

$$\begin{aligned} & \int_0^1 f_0(x) e^{-\sum_{\lambda \notin K_n} \theta_{0\lambda}\phi_\lambda(x)} dx \\ = \quad & 1 - \sum_{\lambda \notin K_n} \theta_{0\lambda}\beta_\lambda(f_0) + \frac{1}{2}\int_0^1 f_0(x)\left(\sum_{\lambda \notin K_n} \theta_{0\lambda}\phi_\lambda(x)\right)^2 dx \times (1 + o(1)) . \end{aligned}$$

We have

$$\left|\sum_{\lambda \notin K_n} \theta_{0\lambda}\beta_\lambda(f_0)\right| \quad \leq \quad \|f_0\|_2 \left(\sum_{\lambda \notin K_n} \theta_{0\lambda}^2\right)^{\frac{1}{2}}$$

and

$$\int_0^1 f_0(x)\left(\sum_{\lambda \notin K_n} \theta_{0\lambda}\phi_\lambda(x)\right)^2 dx \quad \leq \quad \|f_0\|_\infty \sum_{\lambda \notin K_n} \theta_{0\lambda}^2 .$$

So,

$$\begin{aligned} \log\left(\int_0^1 f_0(x) e^{-\sum_{\lambda \notin K_n} \theta_{0\lambda}\phi_\lambda(x)} dx\right) \quad = \quad & -\sum_{\lambda \notin K_n} \theta_{0\lambda}\beta_\lambda(f_0) - \frac{1}{2}\left(\sum_{\lambda \notin K_n} \theta_{0\lambda}\beta_\lambda(f_0)\right)^2 \\ & + \frac{1}{2}\int_0^1 f_0(x)\left(\sum_{\lambda \notin K_n} \theta_{0\lambda}\phi_\lambda(x)\right)^2 dx + o\left(\sum_{\lambda \notin K_n} \theta_{0\lambda}^2\right), \end{aligned}$$

and

$$\begin{aligned} K(f_0, f_{0K_n}) \quad = \quad & \frac{1}{2}\int_0^1 f_0(x)\left(\sum_{\lambda \notin K_n} \theta_{0\lambda}\phi_\lambda(x)\right)^2 dx \\ & -\frac{1}{2}\left(\sum_{\lambda \notin K_n} \theta_{0\lambda}\beta_\lambda(f_0)\right)^2 + o\left(\sum_{\lambda \notin K_n} \theta_{0\lambda}^2\right). \end{aligned}$$

This implies that for n large enough,

$$K(f_0, f_{0K_n}) \leq \|f_0\|_\infty \sum_{\lambda \notin K_n} \theta_{0\lambda}^2 \leq C k_n^{-2\gamma}.$$

Now, if $\theta \in A(u_n)$ we have

$$\begin{aligned} K(f_0, f_\theta) \quad & = \quad K(f_0, f_{0K_n}) + \sum_{\lambda \in K_n} (\theta_{0\lambda} - \theta_\lambda)\beta_\lambda(f_0) - c(\theta_{0K_n}) + c(\theta) \\ & \leq \quad C k_n^{-2\gamma} + \sum_{\lambda \in K_n} (\theta_{0\lambda} - \theta_\lambda)\beta_\lambda(f_0) - c(\theta_{0K_n}) + c(\theta). \end{aligned}$$

We set for any x, $T(x)=\sum_{\lambda\in K_n}(\theta_\lambda-\theta_{0\lambda})\phi_\lambda(x)$. Using (13), $\|T\|_\infty\le C\sqrt{k_n}u_n\to 0$. So,

$$\int_0^1 f_{0K_n}(x)\exp(T(x))dx = 1+\int_0^1 f_{0K_n}(x)T(x)dx+\int_0^1 f_{0K_n}(x)T^2(x)v(n,x)dx,$$

where v is a bounded function. Since $\log(1+u)\le u$ for any $u>-1$, for $\theta\in A(u_n)$ and n large enough,

$$\begin{aligned} -c(\theta_{0K_n})+c(\theta) &= \log\left(\int_0^1 f_{0K_n}(x)e^{T(x)}dx\right)\\ &\le \int_0^1 f_{0K_n}(x)T(x)dx+\int_0^1 f_{0K_n}(x)T^2(x)v(n,x)dx\\ &\le \sum_{\lambda\in K_n}(\theta_\lambda-\theta_{0\lambda})\beta_\lambda(f_{0K_n})+Ck_nu_n^2. \end{aligned}$$

So,

$$\begin{aligned} K(f_0,f_\theta) &\le Ck_n^{-2\gamma}+\sum_{\lambda\in K_n}(\theta_{0\lambda}-\theta_\lambda)\left(\beta_\lambda(f_0)-\beta_\lambda(f_{0K_n})\right)\\ &\le Ck_n^{-2\gamma}+u_n\|f_0-f_{0K_n}\|_2. \end{aligned}$$

Besides, (15) implies

$$\|f_0-f_{0K_n}\|_2^2 \le \|f_0\|_\infty^2\int_0^1\left(1-\exp\left(-\sum_{\lambda\notin K_n}\theta_{0\lambda}\phi_\lambda(x)-c(\theta_{0K_n})+c(\theta_0)\right)\right)^2dx$$

and

$$|c(\theta_{0K_n})-c(\theta_0)| \le \|\sum_{\lambda\notin K_n}\theta_{0\lambda}\phi_\lambda\|_\infty.$$

Finally,

$$\|f_0-f_{0K_n}\|_2\le C\|\sum_{\lambda\notin K_n}\theta_{0\lambda}\phi_\lambda\|_\infty\le Ck_n^{\frac{1}{2}-\gamma}$$

and

$$K(f_0,f_\theta)\le Ck_n^{-2\gamma}+Cu_nk_n^{\frac{1}{2}-\gamma}. \tag{10}$$

We now bound $V(f_0,f_\theta)$. For this purpose, we refine the control of $|c(\theta_{0K_n})-c(\theta_0)|$:

$$\begin{aligned} |c(\theta_{0K_n})-c(\theta_0)| &= \left|\log\left(\int_0^1 f_0(x)\exp\left(-\sum_{\lambda\notin K_n}\theta_{0\lambda}\phi_\lambda(x)\right)dx\right)\right|\\ &= \left|\log\int_0^1 f_0(x)\left(1-\sum_{\lambda\notin K_n}\theta_{0\lambda}\phi_\lambda(x)+w(n,x)\left(\sum_{\lambda\notin K_n}\theta_{0\lambda}\phi_\lambda(x)\right)^2\right)dx\right|, \end{aligned}$$

where w is a bounded function. So,

$$\begin{aligned} |c(\theta_{0K_n}) - c(\theta_0)| &\leq C\left(\sum_{\lambda \notin K_n} |\theta_{0\lambda}\beta_\lambda(f_0)| + \int_0^1 \left(\sum_{\lambda \notin K_n} \theta_{0\lambda}\phi_\lambda(x)\right)^2 dx\right) \\ &\leq C\left(\sum_{\lambda \notin K_n} \theta_{0\lambda}^2\right)^{\frac{1}{2}} \leq C k_n^{-\gamma}. \end{aligned}$$

In addition,

$$\begin{aligned} |c(\theta_{0K_n}) - c(\theta)| &\leq \sum_{\lambda \in K_n} |\theta_\lambda - \theta_{0\lambda}|\,|\beta_\lambda(f_{0K_n})| + Ck_n u_n^2 \\ &\leq u_n\left(\|f_0 - f_{0K_n}\|_2 + \|f_0\|_2\right) + Ck_n u_n^2 \\ &\leq Cu_n + Ck_n u_n^2. \end{aligned}$$

Finally,

$$\begin{aligned} V(f_0, f_\theta) &= \int_0^1 f_0(x)\left(\sum_{\lambda \in K_n} (\theta_{0\lambda} - \theta_\lambda)\phi_\lambda(x) + \sum_{\lambda \notin K_n} \theta_{0\lambda}\phi_\lambda(x) + c(\theta) - c(\theta_0)\right)^2 \\ &\leq 2\|f_0\|_\infty\left(\sum_{\lambda \in K_n} (\theta_{0\lambda} - \theta_\lambda)^2 + \sum_{\lambda \notin K_n} \theta_{0\lambda}^2\right) + 2(c(\theta) - c(\theta_0))^2 \\ &\leq Cu_n^2 + Ck_n^{-2\gamma} + Ck_n^2 u_n^4. \end{aligned} \tag{11}$$

Now, let us consider the case (PH). We take k_n and u_n such that

$$k_n = k_0 \epsilon_n^{-1/\gamma} \quad \text{and } u_n = u_0 \epsilon_n k_n^{-\frac{1}{2}},$$

where k_0 and u_0 are constants depending on $\|f_0\|_\infty$, γ, R and $\varPhi$. Note that (9) is then satisfied. If ϵ_0 is large enough and u_0 is small enough, then, by using (10) and (11),

$$K(f_0, f_\theta) \leq \epsilon_n^2 \quad \text{and} \quad V(f_0, f_\theta) \leq \epsilon_n^2.$$

So, Condition (C) is satisfied if

$$\mathbb{P}^\pi\left\{A(u_n)\right\} \geq e^{-cn\epsilon_n^2}.$$

We have:

$$\mathbb{P}^\pi\left\{A(u_n)\right\} \geq \mathbb{P}^\pi\left\{\theta \text{ s.t. } \sum_{\lambda \in K_n} (\theta_\lambda - \theta_{0\lambda})^2 \leq u_n^2 \,\middle|\, k_n\right\} \times \exp\left(-c_1 k_n L(k_n)\right).$$

The prior on θ implies that, with $G_\lambda = \lambda^\beta \theta_\lambda \tau_0^{-1/2}$,

$$\begin{aligned}
P_1 &:= \mathbb{P}^\pi \left\{ \theta \text{ s.t. } \sum_{\lambda \in K_n} (\theta_\lambda - \theta_{0\lambda})^2 \leq u_n^2 \middle| k_n \right\} \\
&\geq \mathbb{P}^\pi \left\{ \theta \text{ s.t. } \sum_{\lambda \in K_n} \left| \sqrt{\tau_0} \lambda^{-\beta} G_\lambda - \theta_{0\lambda} \right| \leq u_n \middle| k_n \right\} \\
&= \mathbb{P}^\pi \left\{ \theta \text{ s.t. } \sum_{\lambda \in K_n} \lambda^{-\beta} \left| G_\lambda - \tau_0^{-\frac{1}{2}} \lambda^\beta \theta_{0\lambda} \right| \leq \tau_0^{-\frac{1}{2}} u_n \middle| k_n \right\} \\
&= \int ... \int 1_{\left\{ \sum_{\lambda \in K_n} \lambda^{-\beta} \left| x_\lambda - \tau_0^{-\frac{1}{2}} \lambda^\beta \theta_{0\lambda} \right| \leq \tau_0^{-\frac{1}{2}} u_n \right\}} \prod_{\lambda \in K_n} g(x_\lambda) dx_\lambda \\
&\geq \int ... \int 1_{\left\{ \sum_{\lambda \in K_n} \lambda^{-\beta} |y_\lambda| \leq \tau_0^{-\frac{1}{2}} u_n \right\}} \prod_{\lambda \in K_n} g\left(y_\lambda + \tau_0^{-\frac{1}{2}} \lambda^\beta \theta_{0\lambda} \right) dy_\lambda.
\end{aligned}$$

When $\gamma \geq \beta$, we have $\sup_{\lambda \in K_n} \left| \tau_0^{-\frac{1}{2}} \lambda^\beta \theta_{0\lambda} \right| < \infty$ and $\sup_n \left\{ \tau_0^{-\frac{1}{2}} k_n^\beta u_n \right\} < \infty$. Using assumptions on the prior, there exists a constant D such that

$$\begin{aligned}
P_1 &\geq D^{k_n} \int ... \int 1_{\left\{ \sum_{\lambda \in K_n} \lambda^{-\beta} |y_\lambda| \leq \tau_0^{-\frac{1}{2}} u_n \right\}} \prod_{\lambda \in K_n} dy_\lambda \\
&\geq \exp\left(-C k_n \log n \right).
\end{aligned}$$

When $\gamma < \beta$, there exist a and $b > 0$ such that $\forall |y| \leq M$ for some positive constant M

$$g(y+u) \geq a \exp(-b|u|^{p_*}).$$

Using the above calculations we obtain if $p_* \leq 2$,

$$\begin{aligned}
P_1 &\geq D^{k_n} \exp\{-C \sum_{\lambda \in K_n} \lambda^{p_* \beta} |\theta_{0\lambda}|^{p_*}\} \int ... \int 1_{\left\{ \sum_{\lambda \in K_n} \lambda^{-\beta} |y_\lambda| \leq \tau_0^{-\frac{1}{2}} u_n \right\}} \prod_{\lambda \in K_n} dy_\lambda \\
&\geq \exp\left[-C k_n^{1 - p_*/2 + p_*(\beta - \gamma)} \right] \exp\left(-C k_n \log n \right) \\
&\geq \exp\left(-C k_n \log n \right) \quad \text{ if } \beta \leq 1/2 + \gamma
\end{aligned}$$

and if $p_* > 2$, $\sum_{\lambda \in K_n} \lambda^{p_* \beta} |\theta_{0\lambda}|^{p_*} \leq k_n^{p_* \beta - p_* \gamma}$ so that

$$\begin{aligned}
P_1 &\geq D^{k_n} \exp\{-C \sum_{\lambda \in K_n} \lambda^{p_* \beta} |\theta_{0\lambda}|^{p_*}\} \exp\left(-C k_n \log n \right) \\
&\geq \exp\left(-C k_n \log n \right) \quad \text{ if } \beta \leq \gamma + 1/p_*.
\end{aligned}$$

Condition (C) is established by choosing k_0 small enough. Similar computations lead to the result in the case (D). The result for the norm $\|\theta - \theta_0\|_{\ell_2}$ is proved using (19).

4 Conclusions

This paper has investigated posterior concentration rates for Besov balls $\mathcal{B}^{\gamma}_{p,q}(R)$ in the case $p \geq 2$. General prior models allow us to obtain adaptive minimax rates up to logarthmic terms on the class

$$\mathcal{C}(\gamma_0, R_0) = \left\{\mathcal{B}^{\gamma}_{p,q}(R) : \gamma_0 \leq \gamma < r+1,\, p \geq 2,\, 1 \leq q \leq \infty,\, 0 < R \leq R_0\right\},$$

where $R_0 > 0$ and $\gamma_0 > 1/2$. When considering rates of convergence where the loss function is the ℓ_2-norm or the Hellinger distance, Fourier and wavelets bases lead to the same results. But they can potentially differ if we consider different types of loss functions such as $\mathbb{L}_s$-norms, with $s \neq 2$.

We mention that posterior concentration rates for wavelet density estimation on Besov balls have also been studied in Section 4.5 of van der Vaart and van Zanten (2008). In this particular framework, priors proposed by van der Vaart and van Zanten (2008) are based on (truncated) Gaussian distributions and are special cases of our prior model (D). Furthermore, the rates obtained in Theorem 2.1 are similar to rates obtained in their Theorem 4.5 up to a logarithmic term. So, our results show that when truncated series are considered, the choice of the Gaussian prior is not critical. Also, randomizing as in the prior model (PH) leads to adaptation, which is not possible with purely Gaussian priors (see Theorem 1 of Castillo (2008)). Note also that non-Gaussian priors have been proved to be particularly useful in the context of sparsity: see for instance Dalalyan and Tsybakov (2007) who established sparse oracle inequalities for aggregation in the PAC-Bayesian setting, Rivoirard (2006) who studied minimax rates on maxisets of classical procedures and Park and Casella (2008) who provided a Bayesian interpretation of the Lasso procedure.

Note that we have only focused on the case $p \geq 2$. Indeed, Besov spaces $\mathcal{B}^{\gamma}_{p,q}$ with $p < 2$ model very different functions under the $\mathbb{L}_2$-loss, so the value $p = 2$ constitutes an elbow, that clearly appears in Inequality (12). Such phenomena have been investigated using various approaches: in the minimax approach by Donoho et al. (1995) or Reynaud-Bouret et al. (2011) and in the Bayesian context where least favorable priors for such spaces are built (see Johnstone (1994) or Rivoirard (2006)). From these studies, we can draw the following conclusions: when $p < 2$, Besov spaces $\mathcal{B}^{\gamma}_{p,q}$ model sparse signals where at each resolution level, very few of the wavelet coefficients are non-negligible. But these coefficients can be very large. When $p \geq 2$, $\mathcal{B}^{\gamma}_{p,q}$-spaces typically model dense signals where the wavelet coefficients are not large but most of them can be non-negligible. Of course, the study of posterior concentration rates in the case $p < 2$ is an exciting topic, we wish to investigate in future work. The results obtained in this paper illustrate the type of behaviour to be expected using wavelets or Fourier bases, outside the simple case of density estimation on $[0, 1]$. Theoretical extensions of our results to more intricate problems such as estimating intensities of multivariate counting processes or multivariate regression functions are also challenging problems of interest.

5 Appendix

5.1 A result for convergence rates of posterior distributions

To prove Theorem 2.1, we use the following version of theorems on posterior convergence rates.

Theorem 5.1. *Let f_0 be the true density and let π be a prior on $\mathcal{F}$ satisfying the following conditions: There exist $(\epsilon_n)_n$ a positive sequence decreasing to zero with $n\epsilon_n^2 \to +\infty$ and a constant $c > 0$ such that for any n, there exists $\mathcal{F}_n^* \subset \mathcal{F}$ satisfying*

- *(A)*

$$\mathbb{P}^\pi \left\{ \mathcal{F}_n^{*c} \right\} = o(e^{-(c+2)n\epsilon_n^2}).$$

- *(B) For any $j \in \mathbb{N}^*$, let*

$$S_{n,j} = \{ f \in \mathcal{F}_n^* \ \text{ s.t. } \ j\epsilon_n < h(f_0, f) \leq (j+1)\epsilon_n \},$$

  and $H_{n,j}$ the Hellinger metric entropy of $S_{n,j}$, i.e. the logarithm of the smallest number of balls of radius $j\epsilon_n/2$ needed to cover $S_{n,j}$. There exists $J_{0,n}$ (that may depend on n) such that for all $j \geq J_{0,n}$,

$$H_{n,j} \leq K n j^2 \epsilon_n^2,$$

  where K is an absolute constant.

- *(C) If $B_n(\epsilon_n) = \{ f \in \mathcal{F} \ \text{ s.t. } \ K(f_0, f) \leq \epsilon_n^2, \ V(f_0, f) \leq \epsilon_n^2 \}$, we have $\mathbb{P}^\pi \left\{ B_n(\epsilon_n) \right\} \geq e^{-cn\epsilon_n^2}$.*

Then, we have:

$$\mathbb{P}^\pi \left\{ f \ \text{ s.t. } \ h(f_0, f) \leq J_{0,n}\epsilon_n | X^n \right\} = 1 + o_{\mathbb{P}_0}(1).$$

Proof. The proof of Theorem 5.1 is a slight modification of Theorem 2.4 of Ghosal et al. (2000). We introduce $G_n = \{ f \ \text{ s.t. } \ h(f_0, f) > J_{0,n}\epsilon_n \}$. So,

$$\mathbb{P}^\pi \left\{ f \ \text{ s.t. } \ h(f_0, f) > J_{0,n}\epsilon_n | X^n \right\} \leq \mathbb{P}^\pi \left[G_n \cap \mathcal{F}_n^* | X^n \right] + \mathbb{P}^\pi \left\{ (\mathcal{F}_n^*)^c | X^n \right\},$$

and using the same arguments as Ghosal et al. (2000), condition (C) combined with condition (A) implies that

$$\mathbb{P}^\pi \left\{ (\mathcal{F}_n^*)^c | X^n \right\} = o_{\mathbb{P}_0}(1).$$

We now study

$$\mathbb{P}^\pi \left\{ G_n \cap \mathcal{F}_n^* | X^n \right\} = \frac{\int_{G_n \cap \mathcal{F}_n^*} e^{\ell_n(f) - \ell_n(f_0)} d\pi(f)}{\int_{\mathcal{F}} e^{\ell_n(f) - \ell_n(f_0)} d\pi(f)} := \frac{N_n}{D_n}.$$

From the proof of Theorem 2.4 of Ghosal et al. (2000), we obtain that

$$D_n \geq e^{-(c+2)n\epsilon_n^2}$$

with $\mathbb{P}_0$-probability going to 1. We set $L_{n,j} = \exp(H_{n,j})$. Let us a consider a set of densities $(f_l)_{1\leq l\leq L_{n,j}}$ such that $\cup_{1\leq l\leq L_{n,j}} B_h(f_l, j\epsilon_n/2)$, the union of the balls of center f_l and radius $j\epsilon_n/2$ for the Hellinger distance, constitutes a covering of $S_{n,j}$. Following Section 7 of Ghosal et al. (2000), for any l, there exists a test $\phi_{(l)}$ such that

$$\mathbb{E}_{f_0}[\phi_{(l)}] \leq e^{-8Knh^2(f_0,f_l)}, \qquad \sup_{f:\, h(f,f_l)\leq h(f_0,f_l)/2} \mathbb{E}_f[1-\phi_{(l)}] \leq e^{-8Knh^2(f_0,f_l)},$$

where K is an absolute constant. For any $f \in S_{n,j}$, there exists l such that $h(f_l, f) \leq j\epsilon_n/2 \leq h(f_0, f)/2$. Let

$$\phi_n = \max_{j\geq J_{0,n}} \max_{1\leq l\leq L_{n,j}} \phi_{(l)}.$$

By definition of $H_{n,j}$, for any l, $B_h(f_l, j\epsilon_n/2) \cap S_{n,j} \neq \emptyset$. So, there exists $\tilde{f}_l$ such that $\tilde{f}_l \in B_h(f_l, j\epsilon_n/2) \cap S_{n,j}$ and

$$h(f_0, f_l) \geq h(f_0, \tilde{f}_l) - h(f_l, \tilde{f}_l) \geq j\epsilon_n - j\epsilon_n/2 = j\epsilon_n/2.$$

We obtain

$$\mathbb{E}_{f_0}[\phi_n] \leq \sum_{j\geq J_{0,n}} \exp(H_{n,j}) e^{-2Kj^2n\epsilon_n^2} \leq \sum_{j\geq J_{0,n}} e^{-Kj^2n\epsilon_n^2} = o(1)$$

and

$$\sup_{f\in S_{n,j}} \mathbb{E}_f[1-\phi_n] \leq e^{-2Kj^2n\epsilon_n^2}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{f_0}[N_n(1-\phi_n)] &\leq \int d\pi(f) \sum_{j\geq J_{0,n}} 1_{\{f\in S_{n,j}\}} \mathbb{E}_f[1-\phi_n] \\
&\leq \sum_{j\geq J_{0,n}} e^{-2Kj^2n\epsilon_n^2}.
\end{aligned}$$

Finally, by taking

$$J_{0,n} > \sqrt{\frac{c+2}{2K}},$$

we have:

$$\begin{aligned}
&\mathbb{E}_{f_0}[\mathbb{P}^\pi\{G_n \cap \mathcal{F}_n^* | X^n\}] \\
&\quad\leq \mathbb{P}_0\left\{D_n < e^{-(c+2)n\epsilon_n^2}\right\} + \mathbb{E}_{f_0}[\phi_n] + \mathbb{E}_{f_0}\left[\frac{N_n}{D_n}(1-\phi_n)1_{\{D_n\geq e^{-(c+2)n\epsilon_n^2}\}}\right] \\
&\quad\leq \mathbb{P}_0\left\{D_n < e^{-(c+2)n\epsilon_n^2}\right\} + \mathbb{E}_{f_0}[\phi_n] + \sum_{j\geq J_{0,n}} e^{-(2Kj^2-c-2)n\epsilon_n^2} = o(1),
\end{aligned}$$

which ends the proof of Theorem 5.1. ■

5.2 Function approximation spaces

This section is devoted to function approximation spaces and to a technical lemma useful to establish our main result.

We first give a brief description of Sobolev and Besov spaces. We recall that a function $h \in \mathbb{L}_2$ belongs to the Sobolev space W^γ ($\gamma \in \mathbb{N}^*$) if it is γ-times weakly differentiable and if $h^{(j)} \in \mathbb{L}_2$, $j = 1, \dots, \gamma$. The parameter γ measures the smoothness of underlying functions; larger γ means smoother functions. Periodized Sobolev spaces are characterized by Fourier and wavelet bases and using notations from the Introduction, we have

$$h := \sum_{\lambda \in \mathbb{N}} \theta_\lambda \phi_\lambda \in W^\gamma \iff \sum_{\lambda \in \mathbb{N}} |\lambda|^{2\gamma} \theta_\lambda^2 < \infty,$$

which allows us to extend the definition of periodized Sobolev spaces to the case $\gamma \in \mathbb{R}_+^*$. See Bergh and Löfström (1976), DeVore and Lorentz (1993) or Tsybakov (2009) for more details.

Besov spaces are classically defined by using modulus of continuity (see Definition 9.2 of Härdle et al. (1998)). In the framework introduced in the Introduction, periodized Besov spaces, denoted $\mathcal{B}^\gamma_{p,q}$, have the following characterization. We assume that the wavelet basis has standard regularity properties and r vanishing moments (see Härdle et al. (1998) for more details). Let $1 \le p, q \le \infty$ and $0 < \gamma < r + 1$, the $\mathcal{B}^\gamma_{p,q}$-norm of h is equivalent to the norm

$$\|h\|_{\gamma,p,q} = \begin{cases} |\theta_{-10}| + \left[\sum_{j\ge 0} 2^{jq(\gamma+\frac{1}{2}-\frac{1}{p})} \|(\theta_{jk})_k\|^q_{\ell_p}\right]^{1/q} & \text{if } q < \infty, \\ |\theta_{-10}| + \sup_{j \ge 0} 2^{j(\gamma+\frac{1}{2}-\frac{1}{p})} \|(\theta_{jk})_k\|_{\ell_p} & \text{if } q = \infty. \end{cases}$$

Using $\|\cdot\|_{\gamma,p,q}$, we say that h belongs to the Besov ball with radius $R > 0$ if $\|h\|_{\gamma,p,q} \le R$. For any $R > 0$, if $0 < \gamma' \le \gamma < r+1$, $1 \le p \le p' \le \infty$ and $1 \le q \le q' \le \infty$, we obviously have

$$\mathcal{B}^\gamma_{p,q}(R) \subset \mathcal{B}^\gamma_{p,q'}(R), \quad \mathcal{B}^\gamma_{p,q}(R) \subset \mathcal{B}^{\gamma'}_{p,q}(R).$$

Moreover

$$\mathcal{B}^\gamma_{p,q}(R) \subset \mathcal{B}^{\gamma'}_{p',q}(R) \text{ if } \gamma - \frac{1}{p} \ge \gamma' - \frac{1}{p'}.$$

Finally, using the Cauchy-Schwarz inequality when $p \ge 2$ and the inequality $\|(\theta_{jk})_k\|_{\ell_2} \le \|(\theta_{jk})_k\|_{\ell_p}$ when $p < 2$, we have for any $j \ge 0$,

$$2^{j\gamma} \|(\theta_{jk})_k\|_{\ell_2} \le 2^{j(\frac{1}{p}-\frac{1}{2})_+} \times 2^{j(\gamma+\frac{1}{2}-\frac{1}{p})} \|(\theta_{jk})_k\|_{\ell_p}, \tag{12}$$

which proves that for $p \ge 2$,

$$\mathcal{B}^\gamma_{p,q}(R) \subset \mathcal{B}^\gamma_{2,q}(R).$$

The class of Besov spaces provides a useful tool to classify wavelet decomposed signals with respect to their regularity and sparsity properties (see Johnstone (1994)). Roughly speaking, regularity increases when γ increases whereas sparsity increases when

p decreases. We finally recall that the scale of Besov spaces includes the class of the Sobolev spaces since we have $\mathcal{B}^{\gamma}_{2,2} = W^{\gamma}$.

Now, we prove the following result.

Lemma 5.1. *Set $K_n = \{1, 2, \dots, k_n\}$ with $k_n \in \mathbb{N}^*$. Assume one of the following two cases:*

- *$\gamma > 0$, $p = q = 2$ when Φ is the Fourier basis*
- *$0 < \gamma < r+1$, $2 \le p \le \infty$, $1 \le q \le \infty$ when Φ is the wavelet basis with r vanishing moments.*

Then the following results hold.

- *There exists a constant $c_{1,\Phi}$ depending only on Φ such that for any $\theta = (\theta_\lambda)_\lambda \in \mathbb{R}^{k_n}$,*
$$\left\| \sum_{\lambda \in K_n} \theta_\lambda \phi_\lambda \right\|_\infty \le c_{1,\Phi} \sqrt{k_n} \|\theta\|_{\ell_2}. \tag{13}$$
- *If $\log(f_0) \in \mathcal{B}^{\gamma}_{p,q}(R)$, then there exists $c_{2,\gamma}$ depending on γ only such that*
$$\sum_{\lambda \notin K_n} \theta_{0\lambda}^2 \le c_{2,\gamma}\, R^2 k_n^{-2\gamma}. \tag{14}$$
- *If $\log(f_0) \in \mathcal{B}^{\gamma}_{p,q}(R)$ with $\gamma > \frac{1}{2}$, then there exists $c_{3,\Phi,\gamma}$ depending on Φ and γ only such that:*
$$\left\| \sum_{\lambda \notin K_n} \theta_{0\lambda} \phi_\lambda \right\|_\infty \le c_{3,\Phi,\gamma}\, R\, k_n^{\frac{1}{2}-\gamma}. \tag{15}$$

Proof. Let us first consider the Fourier basis. We have:

$$\begin{aligned} \left\| \sum_{\lambda \in K_n} \theta_\lambda \phi_\lambda \right\|_\infty &\le \sum_{\lambda \in K_n} |\theta_\lambda| \times \|\phi_\lambda\|_\infty \\ &\le \sqrt{2} \sum_{\lambda \in K_n} |\theta_\lambda|, \end{aligned}$$

which proves (13). Inequality (14) follows from the definition of $\mathcal{B}^{\gamma}_{2,2} = W^{\gamma}$. To prove (15), we use the following inequality: for any x,

$$\begin{aligned} \left| \sum_{\lambda \notin K_n} \theta_{0\lambda} \phi_\lambda(x) \right| &\le \sqrt{2} \sum_{\lambda \notin K_n} |\theta_{0\lambda}| \\ &\le \sqrt{2} \left(\sum_{\lambda \notin K_n} |\lambda|^{2\gamma} \theta_{0\lambda}^2 \right)^{\frac{1}{2}} \left(\sum_{\lambda \notin K_n} |\lambda|^{-2\gamma} \right)^{\frac{1}{2}}. \end{aligned}$$

Now, we consider the wavelet basis. Without loss of generality, we assume that $\log_2(k_n+1) \in \mathbb{N}^*$. We have for any x,

$$\begin{aligned}\left|\sum_{\lambda\in K_n}\theta_\lambda\phi_\lambda(x)\right| &\leq \left(\sum_{\lambda\in K_n}\theta_\lambda^2\right)^{\frac{1}{2}}\left(\sum_{\lambda\in K_n}\phi_\lambda^2(x)\right)^{\frac{1}{2}} \\ &\leq \|\theta\|_{\ell_2}\left(\sum_{-1\leq j\leq\log_2(k_n)}\sum_{k<2^j}\varphi_{jk}^2(x)\right)^{\frac{1}{2}},\end{aligned}$$

with $\varphi_{-10} = \mathbb{1}_{[0,1]}$. Since, for some constant $A > 0$, $\varphi(x) = 0$ for $x \notin [-A, A]$, for $j \geq 0$,

$$\mathrm{card}\left\{k \in \{0, \ldots, 2^j - 1\} \text{ s.t. } \varphi_{jk}(x) \neq 0\right\} \leq 3(2A + 1).$$

(see Mallat (1998), p. 282 or Meyer (1992), p. 112). So, there exists c_φ depending only on φ such that

$$\left|\sum_{\lambda\in K_n}\theta_\lambda\phi_\lambda(x)\right| \leq \|\theta\|_{\ell_2}\left(\sum_{0\leq j\leq\log_2(k_n)}3(2A+1)2^jc_\varphi^2\right)^{\frac{1}{2}},$$

which proves (13). For the second point, we just use the inclusion $\mathcal{B}^\gamma_{p,q}(R) \subset \mathcal{B}^\gamma_{2,\infty}(R)$ and

$$\sum_{\lambda\notin K_n}\theta_{0\lambda}^2 = \sum_{j>\log_2(k_n)}\sum_{k=0}^{2^j-1}\theta_{0jk}^2 \leq R^2\sum_{j>\log_2(k_n)}2^{-2j\gamma} \leq \frac{R^2}{1-2^{-2\gamma}}k_n^{-2\gamma}.$$

Finally, for the last point, we have for any x:

$$\begin{aligned}\left|\sum_{\lambda\notin K_n}\theta_{0\lambda}\phi_\lambda(x)\right| &\leq \sum_{j>\log_2(k_n)}\left(\sum_{k=0}^{2^j-1}\theta_{0jk}^2\right)^{\frac{1}{2}}\left(\sum_{k=0}^{2^j-1}\varphi_{jk}^2(x)\right)^{\frac{1}{2}} \\ &\leq Ck_n^{\frac{1}{2}-\gamma},\end{aligned}$$

where $C \leq R(3(2A+1))^{\frac{1}{2}}c_\varphi(1-2^{\frac{1}{2}-\gamma})^{-1}$. ■

5.3 Proof of Lemma 3.1

This section is devoted to the proof of Lemma 3.1. We use definitions recalled in (2), (3) and (4). Using Theorem 5 of Wong and Shen (1995), with $M_1 = \left(\int_0^1\frac{f_0^2(x)}{f_\theta(x)}dx\right)^{\frac{1}{2}}$, if

$$h^2(f_0, f_\theta) \leq \frac{1}{2}(1-e^{-1})^2,$$

we have

$$V(f_0, f_\theta) \quad \leq \quad 5h^2(f_0, f_\theta)\left(|\log M_1| - \log(h(f_0, f_\theta))\right)^2. \tag{16}$$

But

$$M_1 = \int_0^1 f_0(x) \exp\left(\sum_{\lambda\in\Lambda_n}(\theta_{0\lambda} - \theta_\lambda)\phi_\lambda(x) + \sum_{\lambda\notin\Lambda_n}\theta_{0\lambda}\phi_\lambda(x) - c(\theta_0) + c(\theta)\right)dx$$

and

$$\log\left(\int_0^1 f_0(x)\exp\left(\sum_{\lambda\in\Lambda_n}(\theta_\lambda - \theta_{0\lambda})\phi_\lambda(x) - \sum_{\lambda\notin\Lambda_n}\theta_{0\lambda}\phi_\lambda(x)\right)dx\right) = c(\theta) - c(\theta_0).$$

So,

$$\begin{aligned} |c(\theta) - c(\theta_0)| \quad &\leq \quad \log\left(\int_0^1 f_0(x)\exp\left(\left\|\sum_{\lambda\in\Lambda_n}(\theta_\lambda - \theta_{0\lambda})\phi_\lambda - \sum_{\lambda\notin\Lambda_n}\theta_{0\lambda}\phi_\lambda\right\|_\infty\right)dx\right) \\ &\leq \quad \left\|\sum_{\lambda\in\Lambda_n}(\theta_\lambda - \theta_{0\lambda})\phi_\lambda - \sum_{\lambda\notin\Lambda_n}\theta_{0\lambda}\phi_\lambda\right\|_\infty \end{aligned}$$

and

$$\begin{aligned} M_1 \quad &\leq \quad \int_0^1 f_0(x)\exp\left(2\left\|\sum_{\lambda\in\Lambda_n}(\theta_\lambda - \theta_{0\lambda})\phi_\lambda - \sum_{\lambda\notin\Lambda_n}\theta_{0\lambda}\phi_\lambda\right\|_\infty\right)dx \\ &\leq \quad \exp\left(2\left\|\sum_{\lambda\in\Lambda_n}(\theta_\lambda - \theta_{0\lambda})\phi_\lambda - \sum_{\lambda\notin\Lambda_n}\theta_{0\lambda}\phi_\lambda\right\|_\infty\right) \\ &\leq \quad \exp\left(2c_{1,\Phi}\sqrt{l_n}\|\theta_0 - \theta\|_{\ell_2} + 2c_{3,\Phi,\gamma}Rl_n^{\frac{1}{2}-\gamma}\right) \end{aligned}$$

by using (13) and (15). Similarly,

$$M_1 \geq \exp\left(-2c_{1,\Phi}\sqrt{l_n}\|\theta_0 - \theta\|_{\ell_2} - 2c_{3,\Phi,\gamma}Rl_n^{\frac{1}{2}-\gamma}\right).$$

So,

$$|\log M_1| \leq 2c_{1,\Phi}\sqrt{l_n}\|\theta_0 - \theta\|_{\ell_2} + 2c_{3,\Phi,\gamma}Rl_n^{\frac{1}{2}-\gamma}.$$

Finally, since $f_\theta \in S_{n,j}$ for $j \geq 1$,

$$\begin{aligned} V(f_0, f_\theta) \quad &\leq \quad 5h^2(f_0, f_\theta)\left(2c_{1,\Phi}\sqrt{l_n}\|\theta_0 - \theta\|_{\ell_2} + 2c_{3,\Phi,\gamma}Rl_n^{\frac{1}{2}-\gamma} - \log(\epsilon_n)\right)^2 \\ &\leq \quad \tilde{C}h^2(f_0, f_\theta)\left(l_n\|\theta_0 - \theta\|_{\ell_2}^2 + (\log n)^2\right), \end{aligned} \tag{17}$$

where $\tilde{C}$ depends on Φ, R, β and γ. Since $f_0(x) \geq c_0$ for any x and $\int_0^1 \phi_\lambda(x)dx = 0$ for any $\lambda \in \Lambda_n$, we have by straightforward computations,

$$V(f_0, f_\theta) \quad \geq \quad c_0\|\theta_0 - \theta\|_{\ell_2}^2. \tag{18}$$

Combining (17) and (18), we conclude that

$$\|\theta_0 - \theta\|_{\ell_2}^2 \quad \leq \quad 2\tilde{C}c_0^{-1}(\log n)^2 h^2(f_0, f_\theta), \tag{19}$$

if $h^2(f_0, f_\theta)l_n \leq (j+1)^2\epsilon_n^2 l_n \leq c_0/(2\tilde{C})$. Lemma 3.1 is proved by taking

$$c = \min\left(\frac{c_0}{2\tilde{C}}, \frac{1}{2}(1 - e^{-1})^2\right).$$

References

Barron, A., Schervish, M., and Wasserman, L. (1999). "The consistency of posterior distributions in nonparametric problems." *Annals of Statistics*, 27: 536–561. 311

Bergh, J. and Löfström, J. (1976). *Interpolation spaces. An introduction*. Berlin: Springer-Verlag. Grundlehren der Mathematischen Wissenschaften, No. 223. 327

Castillo, I. (2008). "Lower bounds for posterior rates with Gaussian process priors." *Electronic Journal of Statistics*, 2: 1281–1299. 324

Dalalyan, A. S. and Tsybakov, A. B. (2007). "Aggregation by exponential weighting and sharp oracle inequalities." In *Learning theory*, volume 4539 of *Lecture Notes in Computer Science*, 97–111. Berlin: Springer.
URL http://dx.doi.org/10.1007/978-3-540-72927-3_9 324

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Berlin: Springer-Verlag. 327

Diaconis, P. and Freedman, D. (1986). "On the consistency of Bayes estimates." *Annals of Statistics*, 14: 1–26. 311

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). "Wavelet shrinkage: asymptopia?" *Journal of the Royal Statistical Society Series B*, 57(2): 301–369. With discussion and a reply by the authors.
URL http://links.jstor.org/sici?sici=0035-9246(1995)57:2<301:WSA>2.0.CO;2-S&origin=MSN 324

Ghosal, S., Ghosh, J. K., and van der Vaart, A. (2000). "Convergence rates of posterior distributions." *Annals of Statistics*, 28: 500–531. 311, 313, 325, 326

Ghosal, S., Lember, J., and van der Vaart, A. (2008). "Nonparametric Bayesian model selection and averaging." *Electronic Journal of Statistics*, 2: 63–89. 313

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). *Wavelets, approximation, and statistical applications*. Lecture Notes in Statistics, 129. New York: Springer-Verlag. 312, 327

Huang, T. (2004). "Convergence rates for posterior distributions and adaptive estimation." *Annals of Statistics*, 32: 1556–1593. 313

Johnstone, I. M. (1994). "Minimax Bayes, asymptotic minimax and sparse wavelet priors." In *Statistical decision theory and related topics, V (West Lafayette, IN, 1992)*, 303–326. New York: Springer. 324, 327

Mallat, S. (1998). *A wavelet tour of signal processing*. San Diego, CA: Academic Press Inc. 329

Meyer, Y. (1992). *Wavelets and operators*, volume 37 of *Cambridge Studies in Advanced Mathematics*. Cambridge: Cambridge University Press. Translated from the 1990 French original by D. H. Salinger. 329

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686.
URL `http://dx.doi.org/10.1198/016214508000000337` 324

Reynaud-Bouret, P., Rivoirard, V., and Tuleau-Malot, C. (2011). "Adaptive density estimation: a curse of support?" *Journal of Statistical Planning and Inference*, 141(1): 115–139.
URL `http://dx.doi.org/10.1016/j.jspi.2010.05.017` 324

Rivoirard, V. (2006). "Nonlinear estimation over weak Besov spaces and minimax Bayes method." *Bernoulli*, 12(4): 609–632.
URL `http://dx.doi.org/10.3150/bj/1155735929` 324

Scricciolo, C. (2006). "Convergence rates for Bayesian density estimation of infinite-dimensional exponential families." *Annals of Statistics*, 34: 2897–2920. 313

Shen, X. and Wasserman, L. (2001). "Rates of convergence of posterior distributions." *Annals of Statistics*, 29: 687–714. 313

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. New York: Springer. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. 327

van der Vaart, A. W. and van Zanten, J. H. (2008). "Rates of contraction of posterior distributions based on Gaussian process priors." *Annals of Statistics*, 36(3): 1435–1463. 313, 324

Verdinelli, I. and Wasserman, L. (1998). "Bayesian goodness-of-fit testing using infinite-dimensional exponential families." *Annals of Statistics*, 26: 1215–1241. 313

Wong, W. and Shen, X. (1995). "Probability inequalities for likelihood ratios and convergence rates of sieve MLEs." *Annals of Statistics*, 23: 339–362. 329

Zhao, L. (2000). "Bayesian aspects of some nonparametric problems." *Annals of Statistics*, 28: 532–552. 313

Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant BANHDITS. The authors would like to thank the anonymous Associate Editor and Referee for valuable comments and suggestions.