Submitted to Bernoulli

Heavy-tailed and Horseshoe priors for regression and sparse Besov rates

SERGIOS AGAPIOU[1] ISMAËL CASTILLO[2] PAUL EGELS[3]

[1]*Department of Mathematics and Statistics, University of Cyprus, Nicosia, Cyprus.*
E-mail: agapiou.sergios@ucy.ac.cy
[2]*Sorbonne Université, LPSM; 4, place Jussieu, 75005 Paris, France.*
E-mail: ismael.castillo@sorbonne-universite.fr
[3]*Sorbonne Université, LPSM; 4, place Jussieu, 75005 Paris, France.*
E-mail: paul.egels@sorbonne-universite.fr

The large variety of functions encountered in nonparametric statistics, calls for methods that are flexible enough to achieve optimal or near-optimal performance over a wide variety of functional classes, such as Besov balls, as well as over a large array of loss functions. In this work, we show that a class of heavy-tailed prior distributions on basis function coefficients introduced in [1] and called Oversmoothed heavy-Tailed (OT) priors, leads to Bayesian posterior distributions that satisfy these requirements; the case of horseshoe distributions is also investigated, for the first time in the context of nonparametrics, and we show that they fit into this framework. Posterior contraction rates are derived in two settings. The case of Sobolev–smooth signals and L_2–risk is considered first, along with a lower bound result showing that the imposed form of the scalings on prior coefficients by the OT prior is necessary to get full adaptation to smoothness. Second, the broader case of Besov-smooth signals with $L_{p'}$–risks, $p' \geq 1$, is considered, and minimax posterior contraction rates, adaptive to the underlying smoothness, and including rates in the so-called *sparse* zone, are derived. We provide an implementation of the proposed method and illustrate our results through a simulation study.

Keywords: Bayesian nonparametrics, Besov spaces, frequentist analysis of posterior distributions, heavy-tailed prior distributions, horseshoe prior, sparse Besov rates.

1. Introduction

The ability of estimators to be flexible, in the sense that their properties remain excellent in a variety of contexts, is a particularly sought-after characteristic. In nonparametric statistics, where the quantity of interest is typically an unknown function (which can represent a signal, an image etc. in practical applications), wavelet thresholding methods are a fundamental class that arguably satisfies the previous flexibility desideratum. Simple non-linear thresholding rules such as hard or soft thresholding are very broadly used in many areas of statistics and signal processing; from the mathematical perspective, they achieve asymptotic near-minimaxity over a very broad variety of functional classes and loss functions. While they are simple to implement and overall display very good empirical behaviour, wavelet thresholding methods still face a number of challenges numerically, in particular for low or moderate sample sizes, as one needs to find good values for thresholding constants, which may require some tuning.

Bayesian nonparametric methods, on the other hand, have experienced rapid growth since the early 2000's, when the development of sampling algorithms to simulate from posterior distributions or approximations thereof (MCMC, ABC, Variational Bayes to name just a few) has been paralleled by a progressive mathematical understanding of the properties needed for the prior distribution to achieve optimal or near-optimal convergence properties for the corresponding posterior distributions. We refer to [29, 12] for recent overviews on the field. Among the most popular classes of priors on functions in

statistics and machine learning are Gaussian processes [42] (henceforth GPs). For a number of function classes, in particular ones for which the signal's smoothness is 'homogeneous' over the considered input space, one can show that GPs achieve optimal posterior contraction rates, provided their parameters are well chosen [48, 10], and adaptation to smoothness can be achieved by, for example, drawing a scaling parameter at random [49]. However, for more 'inhomogeneous' signals, Gaussian processes can be shown to lead to suboptimal rates [5] for certain losses at least (e.g. quadratic) and, more generally, properties of posterior distributions for functions with 'heterogeneous' regularity across the input space are currently less understood; we review below a few recent results in this direction.

Regarding optimality of rates, historically the first systematic study of non-matched situations where the parameter of the loss function does not match the norm index defining the functional class, is due to Nemirovskii [40] and to Nemirovskii, Polyak and Tsybakov [38, 39], where in particular it was noticed that linear estimators can be suboptimal. Donoho and Johnstone [23] studied this type of questions in sequence space, while Donoho, Johnstone, Kerkyacharian and Picard [26] investigated *adaptive* minimax rates in density estimation for Besov classes. For such classes, it turns out that the information-theoretic boundaries can be described by three 'zones', depending on the loss function and the characteristics of the Besov space – we refer to [33] for an in-depth discussion –: a *regular* zone, where best linear estimators achieve the minimax rate and the rate is the classical $n^{-\beta/(2\beta+1)}$ rate in terms of regularity β and number of observations n, an *intermediate* zone, where the rate is still the same but linear estimators are suboptimal, and a *sparse* zone, where the rate changes and can be expressed as a combination of the regularity and loss function parameters. This illustrates the delicate interplay between measures of loss and regularity, which in particular leads to significant, polynomial–in–n differences in convergence rates. Aside of wavelet methods, an adaptive local bandwidth selector based on Lepski's method was developed for kernel–based estimators in [36]. While working on this paper, we learned of the very recent work [35], where the authors derive oracle inequalities and convergence rates in non-matched situations for penalisation methods.

Coming back to Bayesian methods, as noted above, Gaussian Processes can be suboptimal when estimating functions with inhomogeneous smoothness (a fact formally proved in [5] in the white noise model, with mean square loss). Recently, it has been noted that taking heavier tails than Gaussian in defining priors can help in obtaining more diverse behaviour. In [3], p-exponential priors on coefficients of a basis expansion are considered; while the posterior convergence rates in mean square loss improve in a range of aspects compared to the Gaussian case, they are still not adaptive to smoothness. A main advantage is that a wider range of Besov spaces is accessible with optimal contraction rates with these priors, for well chosen parameters. For example, unlike GP priors, priors with exponential tails can be tuned to attain the minimax rate in mean square loss, over both classes of homogeneously smooth functions as well as classes permitting spatial inhomogeneities, see [3] and the recent preprint [21]. To make these priors adaptive, it is still necessary, however, and similar to GPs as discussed above, to draw at least one extra parameter at random; such adaptive counterparts have been obtained in [4] for the white noise model, while [32] considered density estimation. In [1], the following (surprising at first) result is obtained: putting heavy-tailed priors (in the sense that they have polynomially decreasing tails) on coefficients can lead to *automatic* adaptation to smoothness. By taking well-chosen (but, crucially, deterministic and universal) scalings on heavy-tailed coefficients – defining so-called Oversmoothed heavy-Tailed (OT) priors –, the authors in [1] derive adaptation, up to a logarithmic factor, to homogeneous and inhomogeneous smoothness in the L_2–norm, and in the L_∞ (supremum)–norm over Hölder classes. Results cover Gaussian white noise regression using standard posterior distributions and priors with at least two moments (excluding e.g. Cauchy or horseshoe priors), while extensions are provided to other models such as density estimation and binary classification using fractional posteriors. The work [13] uses related heavy-tailed priors on coefficients of deep neural networks to derive automatic adaptation to smoothness and compositional structures or to smoothness and geometric structures, as well as adaptation to some anisotropic Besov classes, but results therein are confined to the L_2–loss.

We now review a few of the existing results for Bayesian posteriors for 'non-canonical distances' (in the loose sense of distances for which one cannot directly apply a generic contraction rate theorem as in [28]). For the Bayesian approach, which is likelihood-based, the study of convergence in terms of certain loss functions, such as L_p ($p > 2$) or L_∞–losses, is notoriously difficult. The first systematic study of posterior contraction in such norms is due to Giné and Nickl [30], where consistency (in general sub-minimax) rates are obtained under generic conditions. An approach to get posterior contraction at minimax rate in the supremum norm is introduced in [11]. Adaptive rates have been obtained for specific priors and models [34, 14, 16, 37, 1]. The work [17] shows contraction under the (global) supremum norm for tree-type BCART priors, while [43] proves local rates for spike-and-slab priors and BCART priors. Yet, to the best of our knowledge, there is no systematic study of Bayesian posterior distributions for Besov regularities and loss functions other than L_2 and L_∞ and one of the paper's aims is to fill this gap in a regression setting. To reduce technicalities, we will do so in the Gaussian white noise model, that can be seen as a prototypical nonparametric model [31].

We also underline that, while 'adaptive' priors (in the sense that, for example, they adapt to the unknown underlying smoothness) can sometimes be obtained by drawing certain parameters in the prior at random, in general – especially for distances for which one cannot apply a generic result as in [28] – it is unclear how to get adaptation to smoothness in terms of a specific loss, such as an L_p–loss. To exemplify this, let us consider the class of sieve priors with random truncation, namely priors defined as functions expanded on a finite number K of coefficients over a basis, with random coefficients, and where the cut-off K is itself drawn at random. The work [6], for instance, shows that such priors are typically adaptive in terms of the L_2–loss in regression, while a certain sub-optimality is noted for the posterior mean in pointwise loss. In their study of Bayesian trees, the authors in [17], prove a lower bound result (Theorem 5 in that source) showing that the posterior contraction rate of a fully-grown tree with random depth K (which is a regular histogram prior and can be viewed as a sieve with cut-off K), while optimal for the L_2-loss (up to a logarithmic term), is *polynomially* sub-optimal in terms of a supremum-type loss. This illustrates that the choice of the prior to target adaptation with respect to a certain loss must be done with particular care.

Regarding specific priors, it has been reported in the literature that *horseshoe* priors [9, 41], that are a very popular choice in the context of sparse high-dimensional models (and can be shown to display near-optimal behaviour in these sparse settings [47, 46]), can also be deployed in nonparametrics with empirically very good behaviour, but up to now there is no theory backing up these empirical findings.

The contributions of this paper are threefold

1. we consider the question of obtaining minimax convergence rates for posterior distributions over a wide variety of Besov classes and losses. We show that the oversmoothed heavy-tailed (OT) priors recently introduced in [1] (and featuring scaling parameters with a specific deterministic decrease, slightly faster than polynomial in n) achieve adaptive (near)-minimax rates in the three aforementioned zones. This includes the 'sparse' zone with convergence rates featuring indices relative to both the functional class and the loss function. We also derive a sharpness result showing that scalings with exactly polynomial decrease can lead to suboptimal rates;
2. for the first time, we obtain posterior contraction rates for horseshoe prior distributions in the context of nonparametric estimation; doing so is nontrivial, since this prior both features very heavy (Cauchy) tails and a density going to infinity at zero. We also obtain rates for OT priors without the two moment assumption imposed in [1].
3. we implement our method and show that in terms of statistical risk it is at least comparable to, while sometimes improving upon, state-of-the-art software.

Gaussian white noise model. For a given true regression function $f_0 \in L_2 := L_2([0,1])$ and $n \geq 1$, the data $Y^{(n)}$ will be generated from the following model

$$dY^{(n)} = f_0(t)\,dt + \frac{1}{\sqrt{n}} dW(t), \qquad t \in [0,1], \tag{1}$$

where W is the standard Brownian motion on $\mathbb{R}$. Given $\{\varphi_k : k \geq 1\}$ an orthonormal basis of L_2 for the canonical inner product $\langle .\,, .\rangle$, we denote by $f_{0,k} := \langle f_0, \varphi_k\rangle$ the coefficients of f_0 (which are square summable, $(f_{0,k}) \in \ell_2$), so that

$$f_0 = \sum_{k\geq 1} f_{0,k}\varphi_k \qquad \text{holds in the } L_2\text{–sense.}$$

Projecting (1) onto $\{\varphi_k\}$ and denoting $X_k := \int_0^1 \varphi_k(t)\,dY^{(n)}(t)$, one gets the (infinite) normal sequence model

$$X_k = f_{0,k} + \frac{1}{\sqrt{n}}\xi_k, \quad k \geq 1, \tag{2}$$

where $\{\xi_k\}_{k\geq 1}$ are independent $\mathcal{N}(0,1)$ variables. In the following, we write $X = X^{(n)} = (X_1, X_2, \dots)$ for the corresponding sequence of observations.

Frequentist analysis of posterior distributions. Consider data X generated from the sequence model (2) with true regression function $f_0 \in L_2$ identified to its coefficients $(f_{0,k}) \in \ell_2$. In the following, this will be written as $X = X^{(n)} \sim P_{f_0}^{(n)}$, and one denotes by E_{f_0} the expectation under this distribution. The reconstruction of f_0 from the data X will be conducted through a Bayesian analysis: from a prior distribution Π on ℓ_2 to be chosen below, one constructs a data-dependent probability measure on ℓ_2 called the posterior distribution and given, for any measurable $B \subset \ell_2$, by

$$\Pi[B \,|\, X] := \frac{\int_B \exp\{-\frac{n}{2}\sum_{k\geq 1}(X_k - f_k)^2\}\,d\Pi(f)}{\int \exp\{-\frac{n}{2}\sum_{k\geq 1}(X_k - f_k)^2\}\,d\Pi(f)}. \tag{3}$$

For a given distance d, we say that the posterior contracts around f_0 at the rate $\varepsilon_n \to 0$ in d-loss, if

$$E_{f_0}\Pi(d(f, f_0) \leq M\varepsilon_n \,|X) \to 1 \quad \text{as } n \to \infty, \tag{4}$$

with $M > 0$ a sufficiently large constant.

Heavy tailed series priors. For a function $f \in L_2$, with $f = \sum_{k\geq 1} f_k\varphi_k$, we define a prior Π on f by drawing the coefficients f_k independently using heavy-tailed density functions. Here we consider heavy-tailed priors as in [1] that have the form

$$f_k = \sigma_k \zeta_k, \tag{5}$$

where σ_k are deterministic decaying coefficients and ζ_k are independent identically distributed (i.i.d.) random variables with a given heavy-tailed density h on $\mathbb{R}$. In particular, we assume that h satisfies the following conditions: there exist constants $c_1, c_2 > 0$ and $\kappa \geq 0$ such that

(H1) h is symmetric, positive, and decreasing on $[0, +\infty)$,
(H2) for all $x \geq 0$,

$$\log\left(1/h(x)\right) \leq c_1(1 + \log^{1+\kappa}(1 + x)),$$

(H3) for all $x \geq 1$,

$$\overline{H}(x) := \int_x^{+\infty} h(u)du \leq \frac{c_2}{x}.$$

Note that, compared to [1], the density h is not required to be bounded and must only satisfy the tail condition (H3). The latter condition is very mild and includes distributions without integer moments. For instance, $\kappa = 0$ in (H2) allows for densities with polynomial tails such as the Cauchy density.

Oversmoothed heavy-Tailed (OT) priors. Following [1], the choice of scalings (σ_k) in (5) we argue in favour, is a deterministic decay which is (slightly) faster than any polynomial in k: for some $\nu > 0$ and all $k \geq 1$, we set

$$\sigma_k = e^{-(\log k)^{1+\nu}}. \tag{6}$$

For any specific choice of the density h satisfying assumptions (H1)-(H2)-(H3), and any fixed $\nu > 0$, the resulting prior is an instance of what we call the OT-prior. The idea behind this prior is to shrink small values of the observed noisy coefficients by the deterministic small scaling factors σ_k, while the heavy-tails of the ζ_k will enable the prior to capture substantial signals with high probability, see [1, Section 2] for a more detailed intuition. Although the results below hold for any $\nu > 0$, for simplicity we restrict the proofs to the case $\nu = 1$ (the proofs stay the same for different values of ν). The value of the hyper-parameter ν has relatively little effect on the non-asymptotic behavior of the posterior: across all simulations presented in Section 4, we take the same value $\nu = 1/2$.

Student and Horseshoe priors. We consider two specific classes of distributions for ζ_k in (5). The first is the class of Student distributions with at least one degree of freedom, that verify (H1)-(H2)-(H3). The second is the Horseshoe prior with parameter $\sigma_k > 0$,

$$f_k \sim HS(\sigma_k) \qquad \text{independently across } k \geq 1; \tag{7}$$

for any $\tau > 0$, the distribution of $f \sim HS(\tau)$ is a mixture arising from $\lambda \sim C^+(0,1)$ and $f|\lambda \sim \mathcal{N}(0, \tau\lambda^2)$, where $C^+(0,1)$ is the half-Cauchy distribution with location parameter 0 and scale 1. The Horseshoe distribution $HS(\tau)$ has a density h_τ satisfying, for all $t \neq 0$ (see [9], Theorem 1.1),

$$\frac{1}{(2\pi)^{3/2}\tau} \log(1 + \frac{4\tau^2}{t^2}) \leq h_\tau(t) \leq \frac{1}{(2\pi)^{3/2}\tau} \log(1 + \frac{2\tau^2}{t^2}). \tag{8}$$

Horseshoe distributions are of the form (5), indeed if $f_k \sim HS(\sigma_k)$ one can write $f_k = \sigma_k \zeta_k^1$ where $\zeta_k^1 \sim HS(1)$. A random variable $\zeta^1 \sim HS(1)$ is symmetric and has Cauchy-like tails (this readily follows from (8)), its density h_1 satisfies conditions (H1), (H2) with $\kappa = 0$, and $(H3)$. A variation of the Horseshoe prior that is of common use in high-dimensional models is a truncated version with uniform rescaling (with respect to k), that is for some $\tau > 0$,

$$f_k \overset{i.i.d}{\sim} HS(\tau) \;\text{ for }\; k \leq n, \quad \text{and} \quad f_k = 0 \;\text{ for }\; k > n. \tag{9}$$

This prior matches the definition of the heavy-tailed prior (5) with $\sigma_k = \tau$ for all k up to the truncation point $k = n$, after which it assigns all coefficients to zero. The uniform rescaling τ is typically taken to be of the order of a negative power of n, see the discussion in Section 5 and Theorem 1 below.

Outline. In Section 2, we establish posterior contraction rates in the L_2-norm for Hilbert–Sobolev truths. Specifically, in Section 2.1 we show that OT priors on basis coefficients, as well as truncated

horseshoe priors, achieve (nearly) minimax-optimal rates. In Section 2.2, we derive lower bounds (both in probability and in expectation) for heavy-tailed priors with the slightly more common (so far in the literature of series priors) choice of polynomially decaying scalings of the form $\sigma_k = k^{-\alpha-1/2}$. In particular, for β-regular Sobolev truths, while this choice was shown in [1] to lead to (nearly) minimax rates in the over-smoothing regime $\alpha \geq \beta$, here we show that they give rise to polynomially slower–than–minimax rates in the under-smoothing regime $\alpha < \beta$, hence establishing that they cannot be fully adaptive to smoothness, which gives another rationale for choosing OT-scalings as in (6).

Section 3 considers a much broader setting of functional classes allowing for non-homogeneous smoothness and non-matched losses, showing that OT heavy-tailed priors adapt to the Besov regularity of the truth and achieve (nearly) minimax rates for general L_p losses, in all three zones, "regular", "intermediate" and "sparse".

Section 4 provides numerical experiments along with implementation details, illustrating a range of rate behaviors across different signal classes. Section 5 provides a brief discussion putting the results of the paper into perspective. Sections 6.1 and 6.2 contain the proofs of Theorem 1 and 2 respectively. All other proofs and additional technical material are contained in the supplement [2].

2. Contraction in L_2–loss for Hilbert-Sobolev spaces

The rate at which the posterior distributions (3) concentrate around the truth f_0 as in (4), crucially depends on both the regularity of the truth and the loss d. As a warm-up to the more elaborated results considered in the next section with double-indexed wavelet bases, we consider in this section the simplest case of Hilbert-Sobolev-type balls and the L_2–loss, with expansions into a simple single-index orthonormal basis (e.g. the Fourier basis). Recalling that $f_k = \langle f, \varphi_k \rangle$, for $\beta, F > 0$, denote

$$\mathcal{S}^\beta(F) = \Big\{ f \in L_2([0,1]) : \quad \sum_{k \geq 1} k^{2\beta} f_k^2 \leq F^2 \Big\}. \tag{10}$$

For a suitable choice of orthonormal basis $\{\varphi_k\}$, the set $\mathcal{S}^\beta(F)$ corresponds to a ball of L_2–functions with β derivatives in the L_2–sense. A more general study of posterior contraction under L_p–losses and Besov smoothness is conducted in Section 3.

2.1. Upper bounds for OT and truncated Horseshoe priors

Our first result provides an upper bound on the contraction rate for OT priors, where the decay sequence is given by (6). Up to a logarithmic factor (a logarithmic loss is in fact unavoidable for separable estimators, as shown in [8]) such priors achieve the minimax convergence rate in terms of the L_2–loss over the above class.

In all the results to follow, when we refer to priors given by (5), it will be understood that the prior density h of the variables ζ_k satisfies conditions (H1)–(H2)–(H3).

Theorem 1. *Suppose, in the sequence model (2), that $f_0 \in \mathcal{S}^\beta(F)$ and let Π be the OT series prior with independent coefficients (f_k) given by (5) (in particular, the prior (7) is admissible) and scalings given by (6). As $n \to \infty$,*

$$E_{f_0} \Pi \Big[\{ f : ||f - f_0||_2^2 > \mathcal{L}_n n^{-2\beta/(2\beta+1)} \} \,|\, X \Big] \to 0,$$

where $\mathcal{L}_n = \log^\delta n$, for some $\delta > 0$. The same conclusion holds if the prior Π is the truncated Horseshoe prior (9) with uniform scaling $\tau = n^{-4}$, up to setting $\mathcal{L}_n = \log^{2+\delta} n$.

The proof is given in Section 6.1. Theorem 1 establishes that, for the OT priors, the posterior contraction rate is nearly minimax over Sobolev balls $\mathcal{S}^\beta(F)$. Among special case of OT priors, one can choose for instance a Cauchy density for h, as well as the horseshoe density h_τ. Further, the truncated version of the horseshoe prior (9) also leads to an optimal rate (up to slightly worse log-factors). This result strengthens Theorem 1 in [1], where a similar upper bound was obtained for OT priors, but with bounded densities and finite second-order moment. The key difference arises from our proof technique: rather than applying Markov's inequality working under the posterior expectation, we work directly at the level of the posterior probability. This approach allows us to replace the finite second moment condition by the weaker tail condition (H3), enabling the use of heavier-tailed priors, which may have unbounded expectation, such as the Cauchy and Horseshoe priors. A simulation study for these heavy-tailed priors is provided in Section 4.1. We also note that in [1] a result for Cauchy priors was provided, but for fractional posteriors only, and not the standard posterior as considered here.

***Remark* 1.** When using the truncated Horseshoe in Theorem 1, we take $\tau = n^{-b}$ with $b = 4$. This is for technical reasons, and the proof similarly goes through for any power $b \geq 4$. In practice, unless n is very large, one may prefer to take a less conservative choice e.g. $\tau = 1/n$ (as we do in the simulations section). We conjecture that the conclusion of Theorem 1 still holds true for this choice.

2.2. Lower bound for polynomially decaying scalings

Another possible choice of prior to recover $f_0 \in \mathcal{S}^\beta(F)$ is the HT(α) prior, which is defined as follows. It is a heavy-tailed prior (5) defined in a similar way as the OT prior, but with *polynomially* decaying scaling parameters, for some $\alpha > 0$,

$$\sigma_k = k^{-1/2-\alpha}, \qquad \text{for all } k \geq 1. \tag{11}$$

In [1], it was shown that this prior leads to adaptation in the over-smoothing case, that is, if $\alpha \geq \beta$, the posterior contraction rate is the minimax rate over $\mathcal{S}^\beta(F)$ (up to log factors). If on the contrary, it turns out that $\alpha < \beta$, then since parameter classes such as $\mathcal{S}^\beta(F)$ are nested (i.e. $\mathcal{S}^\beta(F) \subset \mathcal{S}^\alpha(F)$ if $\alpha < \beta$), we have that $f_0 \in \mathcal{S}^\alpha(F)$, so by applying the over-smoothing result with α in place of β (we can since we are now in the case of 'matched' regularity of prior and truth), one gets that the posterior contracts at rate (at least) $n^{-\alpha/(2\alpha+1)}$ up to a log factor. However, this does not rule out that the rate is in fact faster. Our next result shows that this is not the case by providing a matching lower bound for the posterior contraction rate in the under-smoothing case $\alpha < \beta$.

Theorem 2. *Let $f_0 \in \mathcal{S}^\beta(F)$ and Π be the HT(α) series prior with $1/2 < \alpha < \beta$. For $\delta > 0$, define*

$$\varphi_n := n^{-\alpha/(2\alpha+1)} \log^{-\delta} n.$$

Then, for δ large enough, as $n \to \infty$,

$$E_{f_0} \Pi\left[\{f : \|f - f_0\|_2 \leq \varphi_n \,|\, X\right] \to 0.$$

Under the weaker smoothness assumption $0 < \alpha < \beta$, the following in-expectation bound holds, as $n \to \infty$,

$$E_{f_0} \int \|f - f_0\|_2^2 \, d\Pi(f \,|\, X) \gtrsim n^{-2\alpha/(2\alpha+1)}.$$

The proof is provided in Section 6.2 and Section C in the supplement [2]. Theorem 2 shows that in the under-smoothing case $\alpha < \beta$, the posterior cannot concentrate at a rate faster than $n^{-\alpha/(2\alpha+1)}$, which is slower than the minimax rate over $\mathcal{S}^\beta(L)$. This is a sharpness (i.e. lower-bound) result that demonstrates that the adaptive property of the HT(α) prior is fundamentally one-sided, meaning that α must be chosen large enough to ensure adaptation. This fact provides a strong incentive to use the OT prior (6), which, as shown in Theorem 1, always guarantees adaptation by ensuring that one is 'always in the over-smoothing case' (hence the name). Also, the lower bound result holds for any given function in $\mathcal{S}^\beta(F)$, and not just for some 'least-favourable' ones.

The first part of Theorem 2 provides a lower bound for the in-probability contraction rate, which is expressed in the same form as the in-probability upper bound in Theorem 1. By an application of a Markov-like inequality, it is not hard to check that this result leads to a lower bound for the contraction rate 'in expectation' $E_{f_0} \int \|f - f_0\|_2^2 \, d\Pi(f \mid X)$, with a rate φ_n^2 (including a logarithmic factor). The second part of the statement provides a more precise lower bound, without logarithmic factor. Simulations regarding contraction in the under-smoothing case are provided in Section A in the supplement [2].

***Remark* 2.** The additional prior condition $\alpha > 1/2$ appearing in the in-probability lower bound arises from our proof technique, based on Lemma 5 which assumes that the density h in our definition of the heavy-tailed series priors possesses certain fractional moments. An inspection of the proof of Theorem 2, reveals that this condition can be relaxed to $\alpha > (1/\rho - 1/2) \vee 0$ if we slightly strengthen assumption (H3) to $\overline{H}(x) \lesssim x^{-\rho}$ with $\rho > 1$. For instance, for a Student distribution with at least two degrees of freedom, the prior regularity condition becomes empty.

3. Sparse Besov rates using heavy-tailed priors

In this section we consider the problem of estimating, from the Bayesian nonparametrics point of view, a Besov function in $L_{p'}$–losses for $p' \geq 1$ in the white noise model, again understood as projected into sequence space; we wish to achieve this through a method (prior) that leads to a posterior distribution *adaptive* to the smoothness level of the unknown regression function. To the best of our knowledge, this question has not been addressed for a Bayesian method (or, more generally, for any likelihood-based approach). We first recall the definition of Besov spaces in terms of wavelet coefficients and then introduce the statistical problem and core result of this paper.

3.1. Besov classes, notation

Let $S > 0$, and J be the smallest integer satisfying $2^J \geq 2S$. Consider $\{\phi_{Jk} : k \in \{0, \dots, 2^J - 1\}\} \cup \{\psi_{jk} : j \geq J, k \in \{0, \dots, 2^j - 1\}\}$ an orthonormal, boundary corrected, S-regular wavelet basis of $L_2([0,1])$ (see [18] for a construction). For any $f \in L_p$ (if $p < \infty$) or f continuous (if $p = \infty$), we have

$$f = \sum_{k=0}^{2^J-1} \langle f, \phi_{Jk} \rangle \phi_{Jk} + \sum_{j \geq J} \sum_{k=0}^{2^j-1} \langle f, \psi_{jk} \rangle \psi_{jk} \quad \text{in } L_p, \quad 1 \leq p \leq \infty.$$

To simplify the notation and to avoid splitting the study of the scaling and wavelet coefficients in the following statistical results, whenever $j \geq J$, we write $f_{jk} := \langle f, \psi_{jk} \rangle$ and whenever $j < J$, we write

$$f_{jk} := \langle f, \phi_{J2^j+k} \rangle, \quad 0 \leq j < J, \quad 0 \leq k < 2^j,$$
$$f_{-1,0} := \langle f, \phi_{J0} \rangle.$$

The L_p–norm of the sequence $(f_{jk})_k$ with $k \in \{0, \dots, 2^j - 1\}$ is denoted as

$$||f_{j\cdot}||_p := \left(\sum_{k=0}^{2^j-1} |f_{jk}|^p \right)^{1/p}.$$

Under conditions on the wavelet basis (all satisfied with a boundary corrected S-Daubechies system), we have a Parseval-like quasi-equality for L_p–norms, with $p \geq 1$, (see Proposition 4.2.8 in [31] and a modification thereof for boundary corrected bases p.357)

$$\left\| \sum_{k=0}^{2^j-1} f_{jk}\psi_{jk} \right\|_p \simeq 2^{j(1/2-1/p)} ||f_{j\cdot}||_p. \tag{12}$$

Here '$\simeq$' means equality up to a constant (depending only on p and the wavelet system). To describe the regularity of a function f we use the decay of its coefficients in (12), for $p, q \in [1, \infty)$ and $0 < s < S$, and define the norm

$$||f||_{B^s_{pq}} := \left(\sum_{j \geq -1} 2^{qj(s+1/2-1/p)} ||f_{j\cdot}||_p^q \right)^{1/q}, \tag{13}$$

with the usual adaptation whenever $p = \infty$ or $q = \infty$. A Besov-type ball is defined for any $p, q \in [1, \infty]$, $F > 0$ and $0 < s < S$, as

$$B^s_{pq}(F) = B^s_{pq}([0,1], F) := \{f \in L_p([0,1]) : ||f||_{B^s_{pq}} \leq F\}. \tag{14}$$

The corresponding Besov space B^s_{pq} is defined as

$$B^s_{pq} := \bigcup_{F>0} B^s_{pq}(F). \tag{15}$$

If the wavelet system is chosen smooth enough, the space B^s_{pq} corresponds to the usual Besov space defined in terms of moduli of continuity (see 4.3 in [31] for a proof). From the definition of Besov spaces in terms of wavelet coefficients as in (13), one can deduce the following embedding properties

$$B^s_{pq} \subset B^{s'}_{p'q} \quad \text{whenever } p' > p \text{ and } s' - 1/p' = s - 1/p, \tag{16}$$

$$B^s_{pq} \subset B^{s'}_{pq'} \quad \text{whenever } s' < s \text{ or when } s' = s \text{ and } q' \geq q. \tag{17}$$

In particular, if $s - 1/p > 0$, $B^s_{pq} \subset B^{s'}_{\infty\infty}$ is included in the space of continuous functions.

3.2. Regression in Besov spaces, sparse rates, contraction for OT decay

Following the same approach as in the Hilbert-Sobolev case, if the signal belongs to a Besov space B^s_{pq}, from the definition of the space in terms of wavelet coefficients it is natural to estimate it from the projections of the white noise model into the (now, double-indexed wavelet) basis, that is,

$$X_{jk} = f_{0,jk} + \frac{1}{\sqrt{n}}\xi_{jk}, \quad j \geq -1, \quad 0 \leq k < 2^j, \tag{18}$$

where $\{\xi_{jk}\}$ are independent $\mathcal{N}(0,1)$ variables. The next condition ensures that $B^s_{pq} \subset L_{p'} \cap L_2$ so that the problem of estimating the true unknown function $f_0 \in B^s_{pq}$ in $L_{p'}$–norm is well-defined. (also, recall that we only consider positive smoothness indices $s > 0$)

$$s > (1/p - 1/(p' \vee 2))_+. \tag{19}$$

Minimax rates in this sequence setting and Besov smoothness were derived by Donoho and Johnstone [27] (announced in [25]), building up from the seminal work of Nemirovskii [40]; the case of the density estimation model was treated in [26]. We refer to the monograph [33], Section 10.4, for a detailed discussion. As it turns out, the minimax rate for this estimation problem crucially depends on the parameter p' of the loss function. To describe this rate, we now introduce some notation – we note that for simplicity we will not keep track on the precise power in the logarithmic factors appearing in the rates; this slightly simplifies the discussion compared to [27] (who also treat the even more general case of Besov losses)–.

We particularize two regions where the statistical behavior differs

$$\mathcal{R} := \{(p,p') : p' < (2s+1)p\} \qquad \text{and} \qquad \mathcal{S} := \{(p,p') : p' \geq (2s+1)p\}.$$

The following index η specifies in which region the indices (p,p') are

$$\eta := sp - \frac{p'-p}{2}, \tag{20}$$

as indeed $\eta > 0$ if and only if $(p,p') \in \mathcal{R}$, and $\eta \leq 0$ if and only if $(p,p') \in \mathcal{S}$. Consider

$$s' := s - 1/p + 1/p', \tag{21}$$

which is positive by (19), and define the rate $\varepsilon_n := n^{-r}$, where, for η as in (20),

$$r := \begin{cases} s/(1+2s), & \text{if } \eta > 0 \\ s'/(1+2(s-1/p)), & \text{if } \eta \leq 0 \end{cases}. \tag{22}$$

In the following lines, we provide some insight into the appearance of the elbow in the rate (22) and the distinction between the two regions. Note that this rate is known to be minimax (up to logarithmic factors) over B^s_{pq} in the continuous case $s > 1/p$. We remark that functions in B^s_{pq} can be "non-homogeneously" smooth; that is, as p' increases, it becomes possible to induce increasingly large perturbations of their $L_{p'}$–norm using only a small number of wavelet coefficients. Consequently, the statistical problem of reconstructing $f_0 \in B^s_{pq}$ from noisy observations of its wavelet coefficients becomes more difficult as p' increases. The region $\mathcal{R}$ is referred to as the *regular* region: for (p,p') in $\mathcal{R}$, the rate (22) corresponds to the standard minimax rate encountered in nonparametric statistics. It is worth noting also that $\mathcal{R}$ can further be divided into two zones. The first one is the 'homogeneous' zone $p' \leq p$, where the aforementioned perturbative effect does not occur; therein, *linear* estimators, understood as linear functionals of the empirical measure $n^{-1}\sum \delta_{X_i}$, are minimax-optimal. The second is the 'non-homogeneous' regular zone $p' > p$, where the perturbative effect does take place; therein, linear estimators are known to be suboptimal; in fact, the linear-minimax rate here is polynomially slower than the global minimax rate, and is of order $n^{-s'/\{2s'+1\}}$, for s' as defined in (21). When $p' < (2s+1)p$, the perturbative effect is not strong enough to cause a discrepancy in the global minimax rate. In contrast, in the so-called *sparse* region $\mathcal{S}$, defined by $p' \geq (2s+1)p$, the minimax rate becomes polynomially slower than the usual nonparametric rate from the regular region (although still

faster than the corresponding linear-minimax rate for $p' < \infty$). In $\mathcal{S}$, the most difficult functions in B^s_{pq} to estimate in the $L_{p'}$-norm exhibit highly localized irregularities, corresponding to only a few wavelet coefficients of large magnitude (hence the term "sparse"), which makes the statistical problem significantly harder.

Theorem 3. *Let $0 < s < S$, $p, q \in [1, \infty]$, $1 \le p' < \infty$, $F > 0$ and suppose (19). Consider observations from model (18) with an unknown function $f_0 \in B^s_{pq}(F)$ for some $F > 0$. Let Π be the OT wavelet series prior sampling coefficients f_{jk} independently as*

$$f_{jk} = 2^{-j^2}\zeta_{jk}, \tag{23}$$

where (ζ_{jk}) are i.i.d. copies of a heavy-tailed random variable ζ satisfying conditions (H1)–(H2)–(H3). Then, for r given by (22),

$$E_{f_0}\Pi\left[\{f : \|f - f_0\|_{p'} \ge \mathcal{L}_n n^{-r}\} \,|\, X\right] \to 0,$$

as $n \to \infty$, where $\mathcal{L}_n = \log^\delta n$ for some $\delta > 0$.

The proof of this result is provided in Section D in the supplement [2]. Theorem 3 shows that OT series priors achieve complete minimax adaptation, up to logarithmic factors, over Besov-type balls *simultaneously* over the regular ($\mathcal{R}$) and sparse ($\mathcal{S}$) regions. To the best of our knowledge, this is the first Bayesian procedure proven to achieve such rates. In contrast, random Gaussian series, similar to linear estimates, are known to be suboptimal in this setting (see [3], Theorem 4.1). Our results also significantly extend previous work [1] and [13] on heavy-tailed priors, which beyond classical Sobolev/Hölder spaces considered only the case of Besov spaces B^s_{pp} and anisotropic Besov spaces $B^{\mathbf{s}}_{pp}([0,1]^d)$ (this time in d dimension) with $p < p' = 2$, in particular being restricted to the (homogeneous part of the) regular region $\mathcal{R}$, where linear and global minimax rates coincide and are equal to the usual rate.

***Remark* 3.** The scaling coefficient 2^{-j^2} in (23) is the analogue, in double-index notation, of the scaling $e^{-(\log k)^{1+\nu}}$, with $\nu = 1$ therein, and up to a constant factor in the exponent. As anticipated below (6), and similarly to the single-index OT-prior considered in Section 2, the results of Theorem 3 still hold, with a similar proof, with scaling coefficients 2^{-j^2} in the two-index version (23) replaced by $2^{-j^{1+\nu}}$ for some $\nu > 0$, although for simplicity to keep the notation minimal, we provide the proof only for the former.

***Remark* 4.** For simplicity in the present paper we do not consider the case of the supremum norm ($p' = \infty$). Let us briefly sketch how one can extend our results to this case, following the approach for the supremum norm introduced in [1] for Hölder-type spaces. For any (p, q), the embedding $B^s_{p,q} \subset B^{s'}_{\infty,\infty}$, given by (16) and (17) shows that estimating a B^s_{pq} function in L_∞-norm is not harder than estimating a $B^{s'}_{\infty\infty}$ function. Noting that $B^{s'}_{\infty\infty}$ corresponds to the usual Hölder-Zygmund space whenever $s' \notin \mathbb{N}$, one can use techniques of [1] Theorem 4 (contraction in sup-norm for Hölder-smooth functions) to obtain the desired rate. This rate corresponds, *up to a logarithmic factor*, to $n^{-s'/(2s'+1)}$, regardless of p and q (since $p' = \infty$, we have $s' = s - 1/p$, which is positive by (19), and $\eta < 0$ for all $p \ge 1$, thus the target rate - up to log terms - defined in (22) is $n^{-s'/(2s'+1)}$). Note that following the argument in [1] will require the additional moment assumption $E|\zeta| < \infty$; we believe that this condition could possibly be omitted albeit under a slightly more technical argument.

4. A simulation study

In this section, we present numerical simulations that corroborate and illustrate our theory. Additional simulations illustrating the lower bound obtained in Theorem 2 are presented in Appendix A [2].

4.1. White noise model regression under OT and Horseshoe priors

We consider the white noise regression model (1), expanded in the orthonormal basis $\varphi_k(t) = \sqrt{2}\cos(\pi(k-1/2)t)$ leading to the normal sequence model (2). As underlying truth, we use a function with coefficients with respect to (φ_k) given by $f_{0,k} = k^{-3/2}\sin(k)$. In particular, this true function can be thought of as having Sobolev regularity (almost) $\beta = 1$.

We consider four priors on the coefficients of the unknown. The first three are of the form $f_k = \sigma_k\zeta_k$ for i.i.d. ζ_k, for the same choice of scalings $\sigma_k = e^{-(\log k)^{3/2}}$, but different distributions of ζ_1:

- Student distribution with 3 degrees of freedom, leading to a Student OT prior;
- Cauchy distribution, leading to the Cauchy OT prior;
- Horseshoe distribution, leading to the Horseshoe OT prior.

The fourth prior is the truncated Horseshoe prior as in (9), with $\sigma_k = \tau = 1/n$, where n is the noise precision parameter in (2).

Note that, for the OT priors we use $3/2$ instead of 2 in the exponent of the logarithm in the definition (6) used in our analysis. As noted in the discussion following (6), the contraction rates are in fact identical for any exponent strictly larger than 1, however we found 'the finite' n behaviour to be slightly better for 3/2 compared to 2 (although the difference seemed relatively small in all conducted experiments), so we kept this choice throughout the simulations.

Due to independence, the posterior decomposes into an infinite product of univariate posteriors. For all considered priors, we use Stan, with random initialization uniformly on the interval $(-2, 2)$, to sample each of the univariate posteriors [45]. One could also use simple random walk type algorithms, however, Stan is particularly convenient due to its simple implementation and adaptive tuning. In all four cases, we truncate at $K = 200$, which, for the considered regularities of the truth and the priors, suffices for the truncation error to be of lower order compared to the estimation error, for the considered noise levels, which range from $n = 10^3$ to $n = 10^5$ (in the specific case of the truncated Horseshoe prior case with $\tau_k = 1/n$, this truncation point in fact appears somewhat earlier than the truncation point $K = n$ allowed in the last part of Theorem 1).

In Figure 1, we present the posterior means as well as 95% credible regions for various noise levels, computed by taking the 95% out of the 4000 draws (after burn-in/warm up) which are closest to the mean in L_2–sense. The three OT priors appear to perform very well at all noise levels, with the two heavier tailed priors (Cauchy and Horseshoe) leading to slightly broader credible regions. The Horseshoe prior with scaling $\tau = 1/n$, appears to be slightly overconfident in all but the lowest noise levels. This was in fact expected based on the intuition on scaled heavy-tailed priors outlined in [1, Section 2] in the univariate normal mean model: for small scalings, posterior means tend to strongly shrink towards 0, while for large scalings they tend to preserve the data. Since for this prior, all coefficients (even in low frequencies) correspond to a small scaling $\tau = 1/n$, all observations in frequencies with small signal (how small only depends on n, and is independent of the frequency), are set very close to 0. As a result there is very little variance in the posterior. On the contrary, the OT-type priors' scalings for the first few frequencies remain large, before the faster than polynomial decay in the scaling parameters eventually kicks in, leading to important frequencies getting significant values under the posterior (see also [1, Section 4] for more discussion on this). This in turn leads to posteriors exhibiting more variability in the OT priors case.

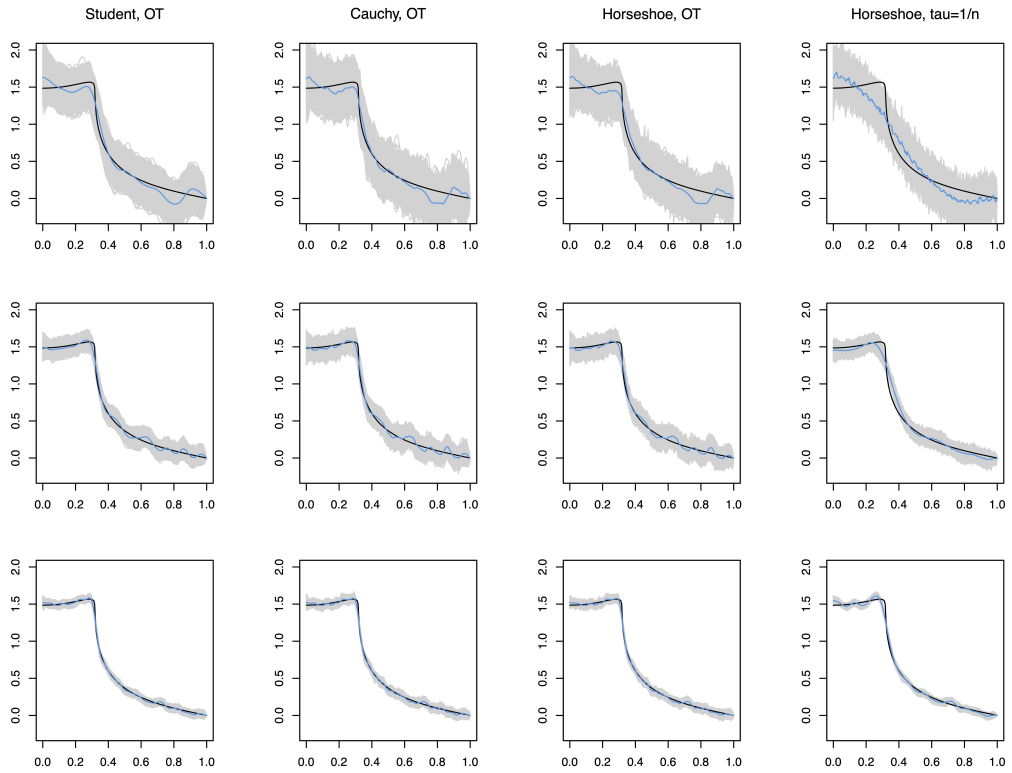

Figure 1. White noise model: true function (black), posterior means (blue), 95% credible regions (grey), for $n = 10^3, 10^4, 10^5$ top to bottom and for the four considered priors left to right.

4.2. Performance for spatially inhomogeneous truths when varying the loss

In this section we consider four spatially inhomogeneous true functions introduced in [23] for testing wavelet thresholding algorithms. We consider the full range of $L_{p'}$–losses, thereby considerably extending the simulation setting for OT priors in [1, Section E.2], where only L_2–loss was considered.

The four functions can be seen in Figure 2. We expand the functions in the Daubechies-8 maximally symmetric wavelet basis [20] (Symmlet-8) and add standard normal noise on each wavelet coefficient. We use 2048 coefficients and coarse level 5, while each function has been appropriately rescaled to get a signal-to-noise ratio (as captured by the ratio of the L_2–norms of the function to the noise) approximately equal to 7, as in [23]. Hence, we have a normal sequence model as in (18) with $n = 1$. Analysis and synthesis of the wavelet expansions is performed in Wavelab850 [22].

Identically to [1, Section E.2], we consider priors on the wavelet coefficients of the form $f_{jk} = \sigma_j \zeta_{jk}$ for i.i.d. ζ_{jk}, with the following choices of the scalings σ_j and the distribution of ζ_{00}:

- Gaussian hierarchical prior: $\sigma_j = \tau 2^{-j(1/2+\alpha)}$ with $\tau \sim$Inv-Gamma$(1,1)$, $\alpha \sim \text{Exp}(1)$, ζ_{00} standard normal;
- Cauchy OT prior: $\sigma_j = 2^{-j^{1+\nu}}$, with $\nu = 1/2$, ζ_{00} distributed according to the standard Cauchy distribution.

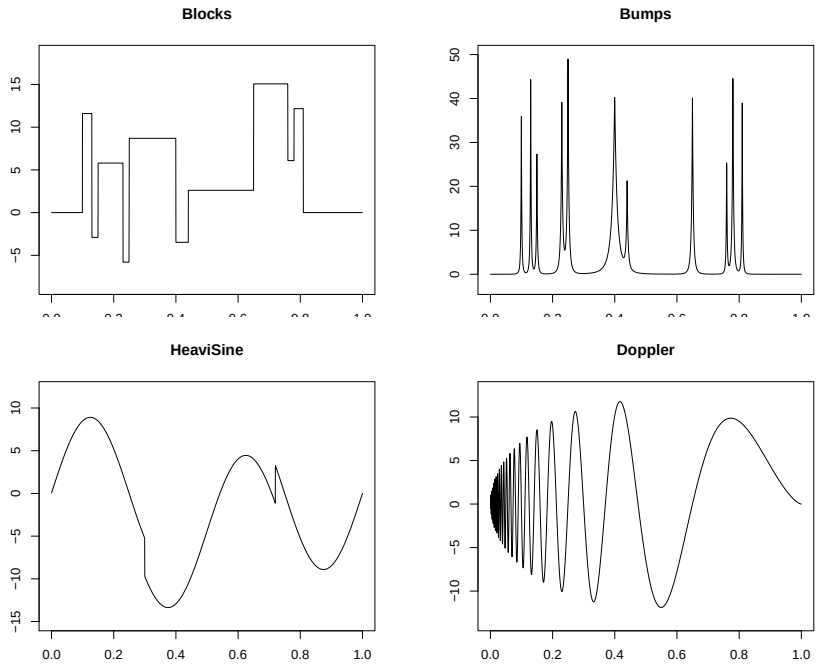

Figure 2. Spatially inhomogeneous true functions.

In addition, we consider frequentist estimation using the hybrid version of SureShrink, which is a soft wavelet thresholding algorithm developed by Donoho and Johnstone to be optimally smoothness-adaptive over Besov spaces, even for extremely sparse signals, see [24, 23].

We refer to [1, Section E.2] for details on the Markov chain Monte Carlo algorithms employed to sample the posteriors, as well as for the resulting visualizations. The implementation of hybrid SureShrink was done using the WaveShrink function of Wavelab850 [22].

To estimate the errors of the three considered methods, we averaged errors over 100 realizations of the data. In particular, for the two priors, we consider two types of errors. The first one is the $L_{p'}$–error of the posterior means, hence after averaging we estimate the error

$$E_{f_0}||\hat{f} - f_0||_{p'},$$

where $\hat{f}$ is the posterior mean. The same type error is computed for $\hat{f}$ being the thresholding estimator. The second type of error, computed only in the two Bayesian settings, estimates

$$E_{f_0}E_{\Pi[\cdot|X]}||f - f_0||_{p'},$$

where the inner expectation is estimated by taking the average of the $L_{p'}$–errors of the Markov chain samples after burn-in, and the outer by averaging over the 100 data realizations. The latter error captures the contraction of the whole posterior around the truth.

In Figure 3 we show estimation errors in $L_{p'}$–loss on the left and contraction-type errors on the right, for $p' = 1, 2, 3, 4, 6$. Errors in supremum-loss are displayed in Table 1. The Cauchy OT prior, convincingly outperforms the Gaussian hierarchical prior in all losses for the more 'jumpy' truths (Blocks and

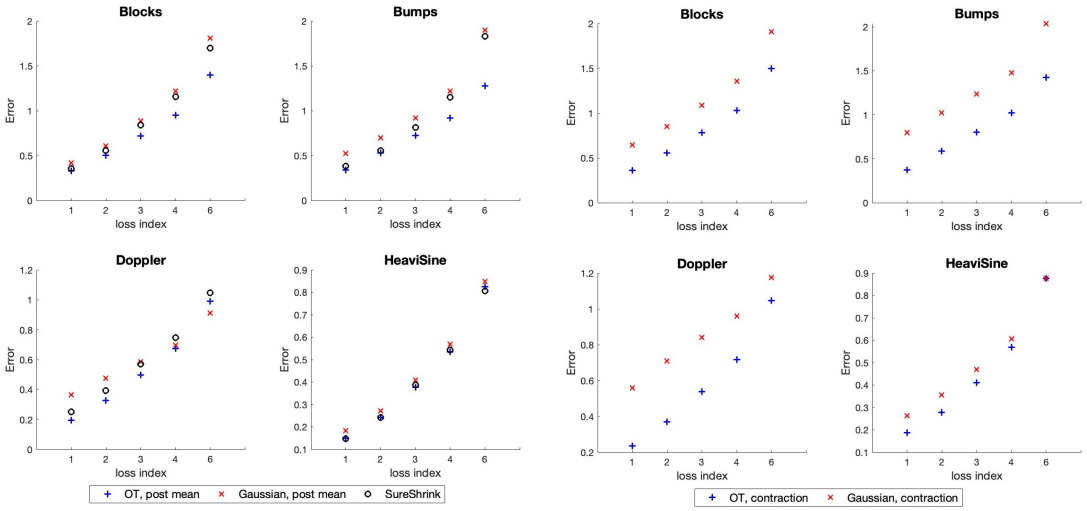

Figure 3. Average errors in $L_{p'}$ for $p' = 1, 2, 3, 4, 6$, for four model spatially inhomogeneous truths. Signal-to-noise ratio approximately 7 for all truths, errors averaged over 100 realizations of the noise. Errors for posterior means on left, contraction-type errors on right, for Cauchy OT prior (blue plus sign markers) and Gaussian hierarchical prior (red cross markers). In the left plot, the black circles are Hybrid SureShrink estimation errors

Bumps). For the smoother truths (Doppler and HeaviSine), the Cauchy prior has significantly better errors for $L_{p'}$-losses with smaller p' (especially contraction-type errors), while as p' increases the errors become more even, with the Gaussian prior slightly outperforming Cauchy OT in supremum-loss (this is, we believe, only true for some specific signals, and we conjecture that the hierarchical Gaussian procedure is in fact suboptimal for minimax adaptation in supremum norm, based on the negative result from [17] mentioned in the introduction). This is despite the fact that, as can be seen in [1, Figures E.6 and E.7], the Gaussian prior fails to denoise the signals and the Cauchy-based posteriors are visually significantly better. The performance of the Cauchy-OT prior matches and often surpasses that of the hybrid SureShrink algorithm.

Table 1. White noise model with spatially inhomogeneous truths: L_∞ average errors of posterior means (contraction-type errors in parentheses), under Cauchy OT and Gaussian hierarchical priors. Hybrid SureShrink estimation error also displayed.

	Cauchy OT	**Gaussian hierarchical**	**SureShrink**
Blocks	4.20 (4.46)	5.08 (5.41)	4.52
Bumps	3.65 (4.08)	5.92 (6.21)	5.64
Doppler	2.85 (3.02)	2.41 (2.91)	2.86
HeaviSine	2.51 (2.69)	2.36 (2.46)	2.18

4.3. Comparison of the performance of OT priors to wavelet thresholding algorithms, in sparse Besov spaces

In this subsection, we construct a sequence of true functions belonging in $B^{3/2}_{1\infty}$, which are close to being 'least favourable', in the sense that they are the most difficult to estimate among that Besov class. Our construction follows the strategy for constructing the functions used to establish lower bounds on the estimation rate over (sparse) Besov spaces in density estimation, as outlined e.g. in [33].

We construct four functions $f_0^{(i)} : [0,1] \to \mathbb{R}, i = 1, \ldots, 4$, defined via their Symmlet-8 wavelet coefficients. Each function, has non-zero wavelet coefficients only at one level: the function $f_0^{(i)}$ has non-zero wavelet coefficients only at the level $j = 2i$. Recalling the definition of the $B^{3/2}_{1\infty}$–norm from Section 3, this means that each function has $B^{3/2}_{1\infty}$–norm equal to

$$2^{2i} \sum_{k=0}^{2^{2i}-1} |f_{2i,k}|, \quad i = 1, \ldots, 4.$$

For $k = 0, \ldots, 2^{2i} - 1$, we choose $f_{2i,k} = 20 \cdot 2^{-2i} w_{2i,k}$, where $|w_{2i,k}|$ sum to 1. In this way, we ensure that each $||f_0^{(i)}||_{B^{3/2}_{1\infty}} = 20$, for $i = 1, \ldots, 4$. The $w_{2i,k}$, are drawn via a 'stick-breaking' construction based on uniform random variables, see [29, Section 3.3.2]; to avoid all significant signal being on the left of the unit interval, we randomly permute the sticks. The resulting functions can be seen in Figure 4. Finally, we add standard normal noise scaled by $1/n$ for $n = 10^{i+1}$ to the coefficients of each truth, to define the noisy observations. For each truth, we generate 100 realizations of the observation.

We consider Cauchy OT prior and the hybrid SureShrink estimator. The implementation of the inference is performed in the same way as in the previous subsection. In Figure 5, we show the logarithms of $L_{p'}$ errors of the posterior means and the thresholding estimators as a function of $\log n$. Here, for each $p' = 1, 2, 3, 4, 6, \infty$, we have four data points, one for each $f_0^{(i)}$, with corresponding noise precision level $n = 10^{i+1}$. The errors for each i are averaged over the 100 realizations of the data, before we apply the logarithm.

According to the minimax rates discussed in Section 3, over $B^{3/2}_{1\infty}$, for $p' \leq 4$ the minimax rate is the usual rate (here $n^{-3/8}$), while for $p' > 4$ we are in the sparse zone and the minimax rate is slower (here $n^{-1/4-1/(2p')}$). Indeed, in Figure 5 we observe that for stronger norms, the errors for both the Cauchy OT prior and SureShrink, appear to decay at a slower rate with n. Furthermore, again the performance of the Cauchy OT prior appears to be on par with (hybrid) SureShrink.

5. Discussion

In this work we introduce a number of nonparametric priors that are flexible enough to reach (near–) minimax optimality over Besov spaces and over a whole spectrum of loss functions. The prior we recommend is the OT (Oversmoothed heavy-Tailed) prior with scale parameters $\sigma_k = \exp\{-(\log k)^{1+\nu}\}$, where we recommend the universal choice $\nu = 1/2$ yielding uniformly excellent results in practice, although close choices such as $\nu = 1$ remain quite close in terms of performance as well; and in terms of type of heavy-tailed prior, taking a Cauchy distribution or a Student(p) with e.g. $p = 3$ degrees of freedom (where the difference of performance between the two is, again, quite mild). The new lower

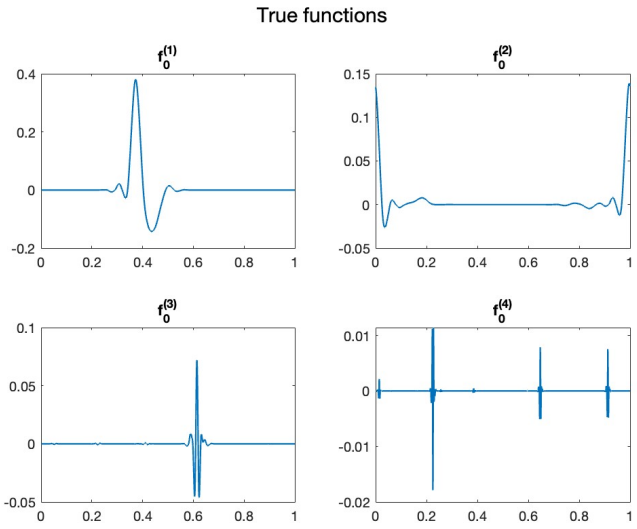

Figure 4. 'Least favourable' truths $f_0^{(i)}$, $i = 1, \ldots, 4$ of unit $B_{1\infty}^{3/2}$-norm, constructed to have non-zero wavelet coefficients only at level $j = 2i$, and with nonzero coefficients constructed via randomly permuted strick-breaking, scaled by 2^{-2i}.

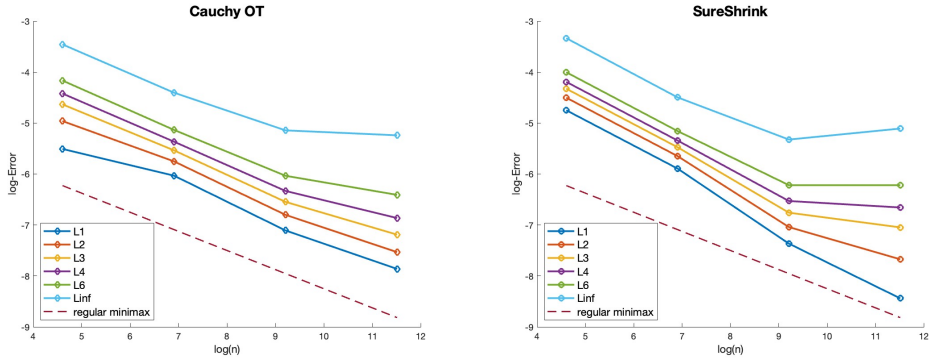

Figure 5. Log average errors for four 'least favourable' truths from Figure 4, with corresponding $n = 10^{i+1}$. $L_{p'}$-errors for the posterior mean for Cauchy OT priors (left) and for the hybrid SureShrink estimator (right), for $p' = 1, 2, 3, 4, 6, \infty$.

bounds results of the present paper (Theorem 2) also confirm that the special type of decrease as above for σ_k (slightly faster than any polynomial in k) is really needed in order to achieve full adaptation, since we prove that the HT(α) prior only achieves one-sided adaptation in the oversmoothing case $\alpha \geq \beta$, as also seen in our simulation study.

The present paper also opens the door to the investigation of (very–) heavy tailed priors for nonparametric inference, including the classical Cauchy prior, or the more elaborate horseshoe distribution. The simulations in Section 4 reveal a number of differences between the OT with Cauchy or Student

prior and the horseshoe prior. From the theoretical perspective, the horseshoe prior fits the main Theorem in a similar way as Cauchy. Empirically, if we take a series prior with scalings σ_k as in the OT prior with horseshoe distributions, we obtain roughly similar behaviour as for Cauchy. However, algorithmically, although in the present setting the difference in computing time is not very significant, in general the fact that the horseshoe models a near-spike at zero – in order to mimic sparse vectors – is in principle not necessary in the present purely nonparametric context. For more complex models (such as density estimation or classification, not considered here, but studied in terms of certain losses in [1]), one may think that modelling a near-spike will be more costly computationally.

Another interesting point of comparison is the *truncated* horseshoe prior. Such a prior is generally used in the context of sparsity – for instance for sparse sequences or high-dimensional linear regression [9] –; similarly, a Cauchy prior with common scale parameter was proposed for high dimensional linear regression in [19]. While such a prior can be used in the present nonparametrics context as well, it behaves somewhat similarly to a hard-thresholding estimator and so seems less flexible than e.g. the SureShrink algorithm that we compare to in Section 4, and indeed its performance is not as favourable for nonparametric function estimation, as one would expect (we did not try to optimise over the parameter τ, which in principle could be done e.g. via empirical Bayes, but the corresponding procedure seems overall more suited to purely sparse classes rather than nonparametric ones). We see that part of the problem with this type of prior is the choice of τ; we expect something similar to occur with spike-and-slab priors as considered in [34]. Indeed, although these priors can be conjectured to achieve similar optimality results for Besov classes as established in the present work, one may think that good performance in practice requires some tuning of the weight of the spike (which plays a similar role as τ for the horseshoe distribution). Also, as discussed in [1], deploying spike-and-slab priors in more complex models is expected to lead to computational difficulties in terms of exploration of sets of submodels (i.e. the ones corresponding to where zeros are distributed from the prior/posterior); and, on the other hand, OT priors retain a form of computational tractability beyond regression models, as their built-in deterministic shrinkage (through non-random scales σ_k) enables to use MCMC without having to sample posteriors on submodels, as would be the case for model-selection type priors such as sieve priors with random cut-off or spike-and-slab priors.

6. Proofs of the results of Section 2

6.1. Proof of Theorem 1

For simplicity we focus on the case $\nu = 1$ for the OT scalings (6), the proof being similar for any other $\nu > 0$. Also part of the statement involving the truncated horseshoe as in (9) is proven in the supplementary material [2]. Let $K_n := n^{1/(2\beta+1)}$ and $v_n = K_n^{-2\beta} \log^\delta n$ with $\delta > 0$ to be chosen large enough below. For $f \in L_2$, let $f^{[K_n]}$ denote its orthogonal projection onto the linear span of the first K_n basis vectors and set $f^{[K_n^c]} := f - f^{[K_n]}$. A union bound leads to

$$\Pi\left[\{f : ||f - f_0||_2^2 > v_n\} \,|\, X\right] \leq \Pi\left[\{f : ||f^{[K_n^c]} - f_0^{[K_n^c]}||_2^2 > v_n/2\} \,|\, X\right] + \Pi\left[\{f : ||f^{[K_n]} - f_0^{[K_n]}||_2^2 > v_n/2\} \,|\, X\right].$$

Let us first deal with the coefficients $k \leq K_n$. Using Markov's inequality and next splitting the sum with $(a+b)^2 \leq 2a^2 + 2b^2$, one can bound the second term in the last display by $(2/v_n)$ times

$$\int ||f^{[K_n]} - f_0^{[K_n]}||_2^2 \, d\Pi(f \,|\, X) \leq 2 \sum_{k \leq K_n} \int (f_k - X_k)^2 \, d\Pi(f \,|\, X) + 2 \sum_{k \leq K_n} (X_k - f_{0,k})^2.$$

Under E_{f_0}, using the definition of the model (2), the second sum on the right hand side is smaller than $K_n/n = o(v_n)$. It now suffices to show that

$$\sum_{k \le K_n} E_{f_0} \left(\int (f_k - X_k)^2 \, d\Pi(f \,|\, X) \right) = o(v_n).$$

Combining Lemma 2 with $p = 2$ and Lemma 1 on coordinate k we get, noting that $|f_{0,k}| \le F$ for all k follows from (10), for any $t \in \mathbb{R}$

$$nE_{f_0} \int (f_k - X_k)^2 \, d\Pi(f \,|\, X) \lesssim t^{-2} \left[1 + \log^2(\sigma_k \sqrt{n}) + t^4 + \log^{2(1+\kappa)} \left(1 + \frac{F + 1/\sqrt{n}}{\sigma_k} \right) \right].$$

Since $\sigma_k = e^{-\log^2 k}$ and $k \le K_n$, we have $\log \sigma_k^{-1} \lesssim \log^2 n$. By taking $t^2 \asymp \log^{2(1+\kappa)} n$, the bound of the last display is of order $\log^{2(1+\kappa)} n =: \log^{\delta'} n$. Therefore the sum in the last but one display is of order $(\log^{\delta'} n) K_n/n = o(v_n)$ if δ is chosen larger than δ'. For the remaining terms $k > K_n$, since $f_0 \in \mathcal{S}^\beta(F)$, when n is large enough $\sum_{k > K_n} |f_{0,k}|^2 \le v_n/8$, so

$$E_{f_0} \Pi \left[\{ f : \sum_{k > K_n} |f_k - f_{0,k}|^2 > v_n/2 \} \,|\, X \right] \le E_{f_0} \Pi \left[\{ f : \sum_{k > K_n} |f_k|^2 > v_n/8 \} \,|\, X \right].$$

Let us introduce the sequence $z_k := k^{-1} \log^{-2} k$. Using in order the summability of (z_k) and a union bound, we have, for a suitable constant $M > 0$,

$$E_{f_0} \Pi \left[\{ f : \sum_{k > K_n} |f_k|^2 > v_n/8 \} \,|\, X \right] \le E_{f_0} \Pi \left[\{ f : \max_{k > K_n} z_k^{-1} f_k^2 \ge M v_n \} \,|\, X \right].$$

For any event A_n, the upper-bound of the last display is further bounded by

$$\sum_{k > K_n} E_{f_0} \left(\Pi \left[\{ f : f_k^2 \ge M z_k v_n \} \,|\, X \right] \mathbf{1}_{A_n} \right) + P_{f_0}(A_n^c). \tag{24}$$

Let us define the event

$$A_n := \bigcap_{K_n < k \le n, k \in N} A_{k,0} \cap \bigcap_{\ell \ge 1} \bigcap_{\ell n < k \le (\ell+1) n, k \in N} A_{k,\ell}, \tag{25}$$

where we have set

$$A_{k,\ell} := \left\{ |X_k| \le \underbrace{\sqrt{\frac{4 \log(n(\ell+1)^2)}{n}}}_{y_{k,\ell}} \right\}; \qquad N := \{ k : |f_{0,k}| \le 1/\sqrt{n} \}.$$

For any ℓ, k such that $\ell n < k \le (\ell+1) n$ we set $x_k = y_{k,\ell}$. With such choice of (x_k) the conditions of Lemma 3 are satisfied on the event A_n. Writing ϕ for the density of the standard Gaussian distribution,

using $\phi \lesssim 1$ for the numerator and for the denominator Lemma 3 with $m = 0$, on coordinate k, we have

$$E_{f_0}\left(\Pi\left[\{f : f_k^2 \geq M z_k v_n\} \,|\, X\right] \mathbf{1}_{A_n}\right) \leq E_{f_0}\left(\frac{\int_{|\theta|>\sqrt{M z_k v_n}} \phi(\sqrt{n}(X_k-\theta))h(\theta/\sigma_k)\,d\theta}{\int \phi(\sqrt{n}(X_k-\theta))h(\theta/\sigma_k)\,d\theta}\mathbf{1}_{A_n}\right)$$
$$\lesssim E_{f_0}\left(\frac{1}{\phi(\sqrt{n}x_k)} 2\overline{H}\left(\sigma_k^{-1}\sqrt{M v_n z_k}\right)\mathbf{1}_{A_n}\right).$$

Now for any $k > K_n$ and n large enough, we have $\sigma_k^{-1}\sqrt{M v_n z_k} \geq 1$, therefore using assumption (H3)

$$\overline{H}\left(\sigma_k^{-1}\sqrt{M v_n z_k}\right) \lesssim \frac{\sigma_k}{\sqrt{M v_n z_k}}.$$

Noting that for $y_{k,\ell}$ as defined above we have $\phi(\sqrt{n}y_{k,\ell}) \gtrsim (n(\ell+1)^2)^{-2}$ the sum in (24) is bounded as

$$\sum_{k>K_n} E_{f_0}\left(\Pi\left[\{f : f_k^2 \geq M z_k v_n\} \,|\, X\right] \mathbf{1}_{A_n}\right)$$
$$\lesssim \sum_{K_n<k\leq n} n^2 \frac{\sigma_k}{\sqrt{M v_n z_k}} + \frac{1}{\sqrt{M v_n}} \sum_{\ell\geq 1} \sum_{\ell n<k\leq(\ell+1)n} (n(\ell+1)^2)^2 \sigma_k z_k^{-1/2}$$
$$\lesssim \frac{n^{7/2}\log n}{\sqrt{v_n}} e^{-C\log^2 n} + \frac{n^{7/2}}{\sqrt{v_n}} \sum_{\ell\geq 1} (\ell+1)^4 \sqrt{\ell+1}\log((\ell+1)n) e^{-\log^2(\ell n)}$$

Taking $v_n = (\log^\delta n) K_n^{-2\beta}$ with $\delta > \delta'$ large enough, the sums in the last display all go to 0, when $n \to \infty$. Finally from Lemma 4 we have $P_{f_0}(A_n^c) \to 0$, which concludes the proof.

6.2. Proof of Theorem 2

Here we only show the proof of the in-probability lower bound, the proof for the in-expectation version can be found in Section C of the supplementary material [2]. Let $K_\alpha := n^{1/(2\alpha+1)}$ and $\varepsilon_n := K_\alpha^{-\alpha}\log^\gamma n$, where $\gamma > 0$ is to be chosen below. It is enough to show (see for instance [10], in particular Lemma 1 therein) that, as $n \to \infty$,

$$\frac{\Pi\left(||f - f_0||_2 \leq \varphi_n\right)}{e^{-2n\varepsilon_n^2}\Pi\left(||f - f_0||_2 \leq \varepsilon_n\right)} \to 0.$$

Since $\alpha < \beta$, we have $f_0 \in S^\beta(F) \subset S^\alpha(F)$ and thanks to [1], Theorem 6, as soon as $\gamma > 0$ is taken large enough, the $HT(\alpha)$ prior satisfies the following prior mass condition, for n large enough,

$$\Pi\left(||f - f_0||_2 \leq \varepsilon_n\right) \geq e^{-n\varepsilon_n^2}. \tag{26}$$

For the numerator, looking at the orthogonal projection on coordinates $k > K_\alpha$, we have

$$\Pi\left(||f - f_0||_2 \leq \varphi_n\right) \leq \Pi\left(||f^{[K_\alpha^c]} - f_0^{[K_\alpha^c]}||_2 \leq \varphi_n\right).$$

Since $f_0 \in S^\beta(F)$ and $\alpha < \beta$, for n large enough, we have $||f_0^{[K_\alpha^c]}||_2 \leq FK_\alpha^{-\beta} \leq \varphi_n$ by definition of φ_n. Thus the triangle inequality leads to

$$\Pi\left(||f^{[K_\alpha^c]} - f_0^{[K_\alpha^c]}||_2 \leq \varphi_n\right) \leq \Pi\left(||f^{[K_\alpha^c]}||_2 \leq 2\varphi_n\right).$$

To bound the last term, apply Lemma 5 with $\mu := 1/2 + \alpha$, $K_n = K_\alpha$ and $\lambda := n(\log n)^{(2\gamma+1)(2\alpha+1)}$ such that $\mu > 1$ and $K_n\lambda^{-1/(2\mu)} \leq 1$ when n is large enough. Thanks to hypothesis (H3), the condition $E(|\zeta|^{1/\mu}) < \infty$ is satisfied for $\mu > 1$ (hence the prior regularity assumption $\alpha > 1/2$, see also Remark 2), therefore we obtain

$$\log\Pi\left(||f^{[K_\alpha^c]}||_2 \leq 2\varphi_n\right) \leq 4\lambda\varphi_n^2 - C\lambda^{1/(2\alpha+1)}.$$

If $\delta > 0$ is large enough, we have $4\lambda\varphi_n^2 \leq (C/2)\lambda^{1/(2\alpha+1)}$ and the previous bound is smaller than

$$-(C/2)n^{1/(2\alpha+1)}(\log n)^{2\gamma+1} \lesssim -n\varepsilon_n^2 \log n.$$

Combining this bound and the upper bound (26) leads to

$$\frac{\Pi\left(||f - f_0||_2 \leq \varphi_n\right)}{e^{-2n\varepsilon_n^2}\Pi\left(||f - f_0||_2 \leq \varepsilon_n\right)} \leq e^{-C'n\varepsilon_n^2 \log n} \to 0,$$

as $n \to \infty$, which concludes the proof for the in-probability bound.

6.3. Technical lemmas

Lemma 1. *Let $\theta \sim \pi$ and $X|\theta \sim \mathcal{N}(\theta, 1/n)$. Suppose π is the law of $\sigma \cdot \zeta$ where $\sigma > 0$ and ζ has an heavy-tailed density satisfying (H1)–(H2). Then, for all $t \in \mathbb{R}$, $\theta_0 \in \mathbb{R}$, $\sigma > 0$,*

$$E_{\theta_0}E^\pi\left[e^{t\sqrt{n}(\theta - X)} \,|\, X\right] \lesssim \sigma\sqrt{n}e^{t^2/2}e^{c_1\log^{1+\kappa}(1+\frac{|\theta_0|+1/\sqrt{n}}{\sigma})}$$

Proof. For any $t \in \mathbb{R}$, denoting h the density of ζ, we have

$$\begin{aligned} E_{\theta_0}E^\pi\left[e^{t\sqrt{n}(\theta - X)} \,|\, X\right] &= E_{\theta_0}\frac{\int \exp(t\sqrt{n}(\theta - X))\phi(\sqrt{n}(X - \theta))h(\theta/\sigma)\,d\theta}{\int \phi(\sqrt{n}(X - \theta))h(\theta/\sigma)\,d\theta} \\ &= E_{\xi\sim\mathcal{N}(0,1)}\frac{\int e^{t(v-\xi) - \frac{(v-\xi)^2}{2}}h(\frac{\theta_0 + v/\sqrt{n}}{\sigma})\,dv}{\int e^{-\frac{(v-\xi)^2}{2}}h(\frac{\theta_0 + v/\sqrt{n}}{\sigma})\,dv}. \end{aligned}$$

The integral on the numerator is bounded as follows, using that h is a density function

$$\int e^{t(v-\xi) - \frac{(v-\xi)^2}{2}}h(\frac{\theta_0 + v/\sqrt{n}}{\sigma})\,dv = e^{t^2/2}\int e^{-\frac{(t-(v-\xi))^2}{2}}h(\frac{\theta_0 + v/\sqrt{n}}{\sigma})\,dv \leq \sigma\sqrt{n}e^{t^2/2}$$

For the denominator using that h is symmetric and decreasing on $(0, \infty)$, along with condition (H2), one gets

$$\int e^{-\frac{(v-\xi)^2}{2}} h(\frac{\theta_0 + v/\sqrt{n}}{\sigma})\, dv \gtrsim \int_{-1}^{1} e^{-\frac{(v-\xi)^2}{2}} e^{-c_1 \log^{1+\kappa}(1+\frac{\theta_0+v/\sqrt{n}}{\sigma})}\, dv$$
$$\gtrsim e^{-c_1 \log^{1+\kappa}(1+\frac{|\theta_0|+1/\sqrt{n}}{\sigma})} \int_{-1}^{1} e^{-\frac{(v-\xi)^2}{2}}\, dv.$$

Using [15] pages 2015-2016, there exists a universal constant $K > 0$, such that

$$E_{\xi\sim\mathcal{N}(0,1)}\left[\left(\int_{-1}^{1} e^{-\frac{(v-\xi)^2}{2}}\, dv.\right)^{-1}\right] \leq K.$$

Combining the previous inequalities leads to the desired result. □

Lemma 2. *Let Y be a real valued random variable. Then for $t > 0$, $p \geq 1$ and $\mathcal{L}(t) := E(\exp(t|Y|))$,*

$$E(|Y|^p) \leq \frac{2^{p-1}}{t^p}\,(p^p + \log^p \mathcal{L}(t))$$

Proof. Write $E(|Y|^p) = t^{-p}E[\log^p(\exp(t|Y|))]$. Using concavity of the map $x \mapsto \log^p x$ for $x > e^{p-1}$, by Jensen's inequality, one may write

$$E[\log^p(\exp(t|Y|))] \leq E[\log^p(e^{p-1} + \exp(t|Y|))] \leq \log^p(e^{p-1} + \mathcal{L}(t))$$
$$\leq (p + \log \mathcal{L}(t))^p \leq 2^{p-1}(p^p + \log^p \mathcal{L}(t)),$$

where the second to last inequality uses $\log(e^{p-1} + x) \leq p + \log x$ valid for $x \geq 1$. □

Lemma 3 (Lower bounds for posterior integrals)**.** *Let ϕ be the density function of the standard Gaussian distribution. Let $\sigma > 0$, $m \geq 0$ and h be a density function satisfying (H1). Let X be a random variable and x a deterministic non-negative real number such that $|X| \leq x$ and $\sigma \lesssim x$, we have*

$$\int |\theta|^m \phi(\sqrt{n}(X-\theta))h(\theta/\sigma)\, d\theta \gtrsim \sigma^{m+1}\phi(\sqrt{n}x).$$

Proof. By symmetry of both h and ϕ, it is enough to focus on the case $X \geq 0$. By restricting the domain of integration to the set $[X - x, X + x]$, the integral of interest is greater than

$$\phi(\sqrt{n}x) \int_{X-x}^{X+x} |\theta|^m h(\theta/\sigma)\, d\theta.$$

By assumption, $X \leq x$, the integral in the last display can be further bounded from below by $\int_0^x |\theta|^m h(\theta/\sigma)\, d\theta = \sigma^{m+1} \int_0^{x/\sigma} |u|^m h(u)\, du$, recalling $X \geq 0$. Using $\sigma \lesssim x$ as well as the positivity and monotonicity of h, the later integral is further bounded below, for some $c > 0$, by $\int_0^c |u|^m h(u)\, du \gtrsim 1$. Putting everything together and using symmetry, one gets the desired result. □

Lemma 4. *Let A_n be the event defined in (25) and assume $f_0 \in \mathcal{S}^\beta(F)$, with $\beta, F > 0$. Then as $n \to \infty$, we have*

$$P_{f_0}(A_n^c) \to 0.$$

Proof. For any $k \in N$, by definition we have $|f_{0,k}| \leq 1/\sqrt{n}$, thus for any $\ell \geq 0$,

$$P_{f_0}(A_{k,\ell}^c) = P_{f_0}\left\{\left|f_{0,k} + \frac{\xi_k}{\sqrt{n}}\right| > \sqrt{\frac{4\log(n(\ell+1)^2)}{n}}\right\} \leq P\left\{|\mathcal{N}(0,1)| > \sqrt{4\log(n(\ell+1)^2)} - 1\right\}.$$

As n gets large enough, this Gaussian tail probability is further upper-bounded, for any $\ell \geq 0$, by

$$P\left\{|\mathcal{N}(0,1)| > \sqrt{3\log(n(\ell+1)^2)}\right\} \leq (n(\ell+1)^2)^{-3/2}.$$

Combined with (25) the definition of the event A_n and a union bound, we obtain

$$P_{f_0}(A_n^c) \leq \sum_{\ell \geq 0}(\ell+1)n \times (n(\ell+1)^2)^{-3/2} \leq n^{-1/2}\sum_{\ell \geq 0}(\ell+1)^{-2} \lesssim 1/\sqrt{n},$$

which ensures $P_{f_0}(A_n^c) \to 0$ as $n \to \infty$. □

Lemma 5. *Consider a random sum of the form*

$$S_n := \sum_{k > K_n} \sigma_k^2 \zeta_k^2,$$

where $\sigma_k = k^{-\mu}$ for some $\mu > 1/2$ and ζ_k are independent and identically distributed copies of the real random variable ζ, satisfying $E[|\zeta|^{1/\mu}] < \infty$. Then for any $\varepsilon > 0$, it holds

$$\log P\big(S_n \leq \varepsilon^2\big) \leq \lambda\varepsilon^2 - C\lambda^{\frac{1}{2\mu}}, \tag{27}$$

for any $\lambda > 0$ such that $K_n\lambda^{-\frac{1}{2\mu}} \leq 1$, where C is a positive constant (depending on μ and the distribution of ζ).

Proof. The proof follows the ideas of [7]. First, notice that for any $\lambda > 0$, by the (exponential) Markov inequality it holds

$$\log P(S_n \leq \varepsilon^2) \leq \lambda\varepsilon^2 + \log Ee^{-\lambda S_n}.$$

By independence, the second term is equal to

$$\log \prod_{k > K_n} Ee^{-\lambda\sigma_k^2\zeta_k^2} = \sum_{k > K_n} \log Ee^{-\lambda\sigma_k^2\zeta_k^2} \leq \int_{K_n}^{\infty} \log Ee^{-\lambda x^{-2\mu}\zeta^2} dx,$$

where in the last bound we have used comparison of sum with an integral (taking advantage of the fact that the expectation in the integrand is increasing with x). Changing variables twice, first setting $\sqrt{\lambda}x^{-\mu} = y$ and then $y = z^{-\mu}$, we get that the last integral is equal to

$$\frac{1}{\mu}\lambda^{\frac{1}{2\mu}}\int_0^{\sqrt{\lambda}K_n^{-\mu}} \big(\log E^{-y^2\zeta^2}\big)y^{-\frac{1}{\mu}-1}dy = \lambda^{\frac{1}{2\mu}}\int_{K_n\lambda^{-\frac{1}{2\mu}}}^{\infty} \log Ee^{-z^{-2\mu}\zeta^2}dz.$$

Noting that the integrand in the right hand side is non-positive, under the assumption $K_n\lambda^{-\frac{1}{2\mu}} \leq 1$, we have that the right hand side is upper bounded by

$$\lambda^{\frac{1}{2\mu}}\int_1^{\infty} \log Ee^{-z^{-2\mu}\zeta^2}dz,$$

where [7, Lemma 4.3] shows that the last integral is finite if and only if $E[|\zeta|^{1/\mu}] < \infty$, as assumed. The claimed bound (27) follows by combining the above considerations. □

References

[1] AGAPIOU, S. and CASTILLO, I. (2024). Heavy-tailed Bayesian nonparametric adaptation. *Ann. Statist.* **52** 1433–1459. MR4804815

[2] AGAPIOU, S., CASTILLO, I. and EGELS, P. Supplement to "Heavy-tailed and Horseshoe priors for regression and sparse Besov rates".

[3] AGAPIOU, S., DASHTI, M. and HELIN, T. (2021). Rates of contraction of posterior distributions based on p-exponential priors. *Bernoulli* **27** 1616–1642. MR4278794

[4] AGAPIOU, S. and SAVVA, A. (2024). Adaptive inference over Besov spaces in the white noise model using p-exponential priors. *Bernoulli* **30** 2275–2300. MR4746608

[5] AGAPIOU, S. and WANG, S. (2024). Laplace priors and spatial inhomogeneity in Bayesian inverse problems. *Bernoulli* **30** 878–910. MR4699538

[6] ARBEL, J., GAYRAUD, G. and ROUSSEAU, J. (2013). Bayesian Optimal Adaptive Estimation Using a Sieve Prior. *Scandinavian Journal of Statistics* **40** 549–570.

[7] AURZADA, F. (2007). On the lower tail probabilities of some random sequences in l_p. *Journal of Theoretical Probability* **20** 843–858.

[8] CAI, T. T. (2008). On information pooling, adaptability and superefficiency in nonparametric function estimation. *Journal of Multivariate Analysis* **99** 421–436.

[9] CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751

[10] CASTILLO, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* **2** 1281–1299. MR2471287

[11] CASTILLO, I. (2014). On Bayesian supremum norm contraction rates. *Ann. Statist.* **42** 2058–2091. MR3262477

[12] CASTILLO, I. (2024). *Bayesian nonparametric statistics, St Flour lecture notes* **LI**. Springer.

[13] CASTILLO, I. and EGELS, P. (2024). Posterior and variational inference for deep neural networks with heavy-tailed weights. arXiv eprint 2406.03369.

[14] CASTILLO, I. and MISMER, R. (2021). Spike and slab Pólya tree posterior densities: adaptive inference. *Ann. Inst. Henri Poincaré Probab. Stat.* **57** 1521–1548. MR4291462

[15] CASTILLO, I. and NICKL, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *The Annals of Statistics* **41** 1999 – 2028.

[16] CASTILLO, I. and RANDRIANARISOA, T. (2022). Optional Pólya trees: posterior rates and uncertainty quantification. *Electron. J. Stat.* **16** 6267–6312. MR4515717

[17] CASTILLO, I. and ROČKOVÁ, V. (2021). Uncertainty quantification for Bayesian CART. *Ann. Statist.* **49** 3482–3509. MR4352538

[18] COHEN, A., DAUBECHIES, I. and VIAL, P. (1993). Wavelets on the Interval and Fast Wavelet Transforms. *Applied and Computational Harmonic Analysis* **1** 54-81.

[19] DALALYAN, A. S. and TSYBAKOV, A. B. (2012). Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.* **78** 1423–1443. MR2926142

[20] DAUBECHIES, I. (1992). *Ten lectures on wavelets*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **61**. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. MR1162107

[21] DOLERA, E., FAVARO, S. and GIORDANO, M. (2024). On strong posterior contraction rates for Besov-Laplace priors in the white noise model. arXiv preprint arxiv:2411.06981.

[22] DONOHO, D., MALEKI, A., SHAHRAM, M. et al. (2006). Wavelab 850. *Software toolkit for time-frequency analysis*.
[23] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089
[24] DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464
[25] DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B* **57** 301–369. With discussion and a reply by the authors. MR1323344
[26] DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539. MR1394974
[27] DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1997). Universal near minimaxity of wavelet shrinkage. In *Festschrift for Lucien Le Cam* 183–218. Springer, New York. MR1462946
[28] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 (2001m:62065)
[29] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of nonparametric Bayesian inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge University Press, Cambridge. MR3587782
[30] GINÉ, E. and NICKL, R. (2011). Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.* **39** 2883–2911. MR3012395
[31] GINÉ, E. and NICKL, R. (2015). *Mathematical foundations of infinite-dimensional statistical models* **40**. Cambridge university press.
[32] GIORDANO, M. (2023). Besov-Laplace priors in density estimation: optimal posterior contraction rates and adaptation. *Electron. J. Stat.* **17** 2210–2249. MR4649387
[33] HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (1998). *Wavelets, approximation, and statistical applications*. *Lecture Notes in Statistics* **129**. Springer-Verlag, New York. MR1618204
[34] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **43** 2259–2295. MR3396985
[35] LACOUR, C., MASSART, P. and RIVOIRARD, V. (2025). Is model selection possible for the ℓ_p-loss? PCO estimation for regression models. *arXiv preprint arXiv:2504.11217*.
[36] LEPSKI, O. V., MAMMEN, E. and SPOKOINY, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* **25** 929–947. MR1447734
[37] NAULET, Z. (2022). Adaptive Bayesian density estimation in sup-norm. *Bernoulli* **28** 1284–1308. MR4388939
[38] NEMIROVSKIĬ, A. S., POLYAK, B. T. and TSYBAKOV, A. B. (1983). Estimates of the maximum likelihood type for a nonparametric regression. *Dokl. Akad. Nauk SSSR* **273** 1310–1314. MR731296
[39] NEMIROVSKIĬ, A. S., POLYAK, B. T. and TSYBAKOV, A. B. (1985). The rate of convergence of nonparametric estimates of maximum likelihood type. *Problemy Peredachi Informatsii* **21** 17–33. MR820705
[40] NEMIROVSKIY, A. S. (1985). Nonparametric estimation of smooth regression functions. *Soviet J. Comput. Systems Sci.* **23** 1–11. MR844292
[41] POLSON, N. G. and SCOTT, J. G. (2011). Shrink globally, act locally: sparse Bayesian regularization and prediction. In *Bayesian statistics 9* 501–538. Oxford Univ. Press, Oxford With discussions by Bertrand Clark, C. Severinski, Merlise A. Clyde, Robert L. Wolpert, Jim e. Griffin, Philiip J. Brown, Chris Hans, Luis R. Pericchi, Christian P. Robert and Julyan Arbel. MR3204017

[42] RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian processes for machine learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435
[43] ROČKOVÁ, V. and ROUSSEAU, J. (2024). Ideal Bayesian spatial adaptation. *J. Amer. Statist. Assoc.* **119** 2078–2091. MR4797924
[44] SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2013). Empirical Bayes scaling of Gaussian priors in the white noise model. *Electron. J. Stat.* **7** 991–1018. MR3044507
[45] STAN DEVELOPMENT TEAM (2024). Stan Modelling Language Users Guide and Reference Manual v. 2.34.
[46] VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* **12** 1221–1274. With a rejoinder by the authors. MR3724985
[47] VAN DER PAS, S. L., KLEIJN, B. J. K. and VAN DER VAART, A. W. (2014). The horseshoe estimator: posterior concentration around nearly black vectors. *Electron. J. Stat.* **8** 2585–2618. MR3285877
[48] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663
[49] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442

SUPPLEMENTARY MATERIAL

This supplement contains additional proofs and simulations. Appendix A presents simulations regarding the lower bound results. The proof of Theorem 1 for the truncated Horseshoe prior distribution is presented in Appendix B, the proof of the in-expectation lower bound of Theorem 2 in Appendix C. The proof of the key main result Theorem 3 providing rates under Besov smoothness in $L_{p'}$-loss can be found in Appendix D. Finally, Appendix E contains a number of additional technical Lemmas.

A. Additional simulations: suboptimality of HT(α) priors in the undersmoothing case

We again consider the white noise regression model (1), this time expanded in the orthonormal basis $\varphi_k(t) = \sqrt{2}\sin(\pi k t)$. leading to the normal sequence model (2). As underlying truth, we use a function with coefficients with respect to (φ_k) given by $f_{0,k} = k^{-2.25}\sin(10k)$. This is the setting studied in [44, Section 3]. In particular, this true function can be thought of as having Sobolev regularity (almost) $\beta = 1.75$.

We consider HT(α) priors (given by (11)) based on the Student distribution with 3 degrees of freedom, for four choices of regularity α, 0.75 and 1.25 (undersmoothing), 1.75 (match) and 2.75 (oversmoothing). Similarly to the previous subsection, we again exploit the independence structure of the model, and employ Stan to sample the one dimensional posteriors. In all cases we again truncate at $K = 200$.

In Figure A.6, we present posterior sample means as well as 95% credible regions for noise precision parameters $n = 2 \times 10^2$ (top row) and $n = 2 \times 10^4$ (bottom row). As expected by the theory, the matched and oversmoothed priors perform very well (cf. [1, Theorem 1]), while the two undersmoothing priors lead to too rough posterior means and perform very poorly (see Theorem 2).

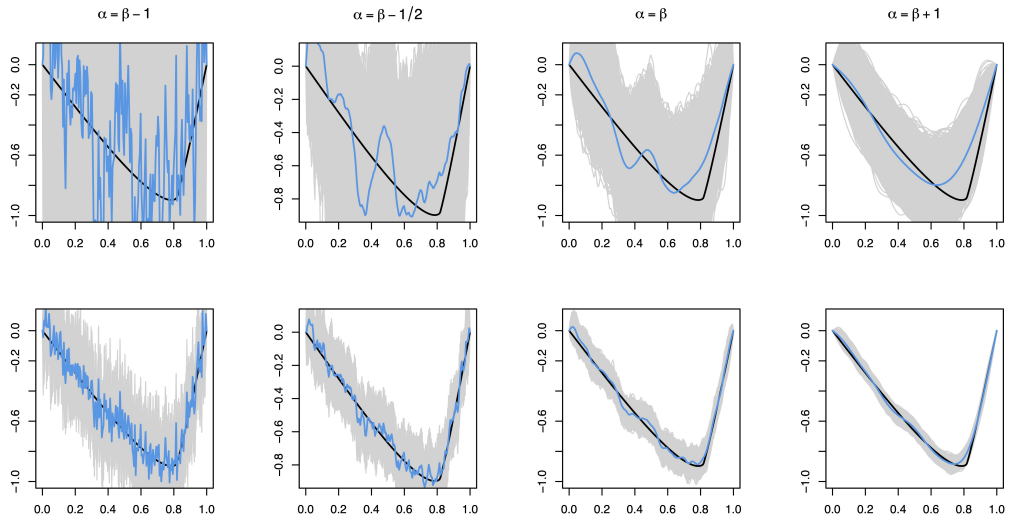

Figure A.6. White noise model: true function (black), posterior means (blue), 95% credible regions (grey), for $n = 2 \times 10^2$ (top row) and $n = 2 \times 10^4$ (bottom row). Student HT(α) prior with $\nu = 3$ degrees of freedom, for $\alpha = 0.75, 1.25, 1.75, 2.75$ left to right, where the Sobolev regularity of the truth β is (almost) 1.75.

B. Proof of Theorem 1 for the truncated Horseshoe

Following the proof of Theorem 1, recall $K_n := n^{1/(2\beta+1)}$ and $v_n = K_n^{-2\beta} \log^{\delta} n$ with $\delta > 0$ to be chosen large enough below. Taking care first of indices $k \leq K_n$ we only need to show

$$E_{f_0} \sum_{k \leq K_n} \int (f_k - X_k)^2 \, d\Pi(f \,|\, X) = o(v_n).$$

Recalling that F is a bound on f_0, applying Lemma 2 with $p = 2$ and Lemma E.1 on coordinate k we get, for any $t \in \mathbb{R}$

$$nE_{f_0} \int (f_k - X_k)^2 \, d\Pi(f \,|\, X) \lesssim t^{-2}(1 + t^4 + \log^2 \left[\frac{\tau\sqrt{n}}{\log\left(1 + \frac{4\tau^2}{(F+1)^2}\right)} \right]).$$

Since $\tau = n^{-a}$ with $a > 0$, whenever n is large enough such that $4\tau^2 < (F+1)^2$, the right hand side of the last display is bounded (up to constant) by

$$t^{-2}(1 + t^4 + \log^2(\tau^{-1}\sqrt{n})).$$

This bound is optimized by taking t of the order of $\sqrt{\log n}$, which leads to

$$E_{f_0} \sum_{k \leq K_n} \int (f_k - X_k)^2 \, d\Pi(f \,|\, X) \lesssim \sum_{k \leq K_n} \frac{\log n}{n} \lesssim \frac{K_n}{n} \log n,$$

this last bound is $o(v_n)$ for $\delta > 1$. Now for the terms $k > K_n$, since the Horseshoe prior is truncated after $k = n$, this term reduces to

$$E_{f_0}\Pi\left[\{f : \sum_{K_n<k\leq n} |f_k|^2 > v_n/8\} \,|\, X\right].$$

We can follow the rest of the proof of Theorem 1 with σ_k replaced by τ, noting that the Horseshoe HS(1) satisfies condition (H3) thanks to the inequality (8). Therefore, as long as $\tau^{-1}\sqrt{Mv_nz_k} \geq 1$ for $n \geq k > K_n$, we have

$$E_{f_0}\Pi\left[\{f : \sum_{K_n<k\leq n} |f_k|^2 > v_n/8\} \,|\, X\right] \lesssim \sum_{K_n<k\leq n} n^2 \frac{\tau}{\sqrt{Mv_nz_k}} + P_{f_0}(A_n^c),$$

Where $A_n = \bigcap_{K_n<k\leq n, k\in N} A_{k,0}$ is the event as in (25). The sum in the last display goes to 0 when $n \to \infty$ for $\tau = n^{-4}$ and $\delta > 2$ or for $\tau = n^{-a}$ and $a > 4$.

C. Proof of the in-expectation lower bound in Theorem 2

To show the in–expectation bound, first define the set

$$\mathcal{N}_n = \{k : |f_{0,k}| > 1/\sqrt{n}\}.$$

Denoting the cardinality of $\mathcal{N}_n$ as N_n, we have

$$F^2 \geq \sum_{k\in\mathcal{N}_n} k^{2\beta}|f_{0,k}|^2 \geq n^{-1}\sum_{k\in\mathcal{N}_n} k^{2\beta} \geq n^{-1}\sum_{k=1}^{N_n} k^{2\beta} \geq n^{-1}(2\beta+1)^{-1}(N_n)^{2\beta+1}.$$

Combining this with the assumption $\alpha < \beta$, we get

$$N_n \leq (nF^2(2\beta+1))^{1/(2\beta+1)} \leq ((2\beta+1)F^2)^{1/(2\beta+1)}K_\alpha =: \widetilde{K}_\alpha.$$

For any k, define the event,

$$A_k := \left\{|X_k| \leq 2/\sqrt{n}\right\}.$$

To establish the in–expectation lower bound, we first restrict the set of indices to

$$\{k > \widetilde{K}_\alpha\} \cap \mathcal{N}_n^c = \{k > \widetilde{K}_\alpha : |f_{0,k}| \leq 1/\sqrt{n}\}.$$

Using for all k, $|f_k - f_{0,k}|^2 \geq |f_k|^2/2 - |f_{0,k}|^2$, we get

$$\begin{aligned} E_{f_0}\int \|f - f_0\|_2^2 \, d\Pi(f \,|\, X) &\geq \sum_{\{k>\widetilde{K}_\alpha\}\cap\mathcal{N}_n^c} E_{f_0}\left(\int |f_k - f_{0,k}|^2 \, d\Pi(f \,|\, X)\right) \\ &\geq \sum_{\{k>\widetilde{K}_\alpha\}\cap\mathcal{N}_n^c} E_{f_0}\left(\int \frac{1}{2}|f_k|^2 \, d\Pi(f \,|\, X)\mathbf{1}_{A_k}\right) - \sum_{\{k>\widetilde{K}_\alpha\}\cap\mathcal{N}_n^c} |f_{0,k}|^2. \end{aligned}$$

First taking care of the non-stochastic sum, since $f_0 \in S^\beta(F)$ and $\widetilde{K}_\alpha \asymp K_\alpha$, we have

$$\sum_{\{k>\widetilde{K}_\alpha\}\cap\mathcal{N}_n^c} |f_{0,k}|^2 \leq \sum_{k>\widetilde{K}_\alpha} |f_{0,k}|^2 \lesssim K_\alpha^{-2\beta}.$$

For the remaining sum, recall that ϕ denotes the standard Gaussian density function, we now study

$$E_{f_0}\left(\int |f_k|^2 \, d\Pi(f \,|\, X)\mathbf{1}_{A_k}\right) = E_{f_0}\left(\frac{\int \theta^2\phi(\sqrt{n}(X_k-\theta))h(\theta/\sigma_k)\,d\theta}{\int \phi(\sqrt{n}(X_k-\theta))h(\theta/\sigma_k)\,d\theta}\mathbf{1}_{A_k}\right) =: E_{f_0}\left(\frac{T}{B}\mathbf{1}_{A_k}\right).$$

Using $\phi \lesssim 1$, we can bound the denominator as $B \lesssim \sigma_k$. To bound the numerator from below we set $x_k := 2/\sqrt{n}$ such that for all $k > \widetilde{K}_\alpha$ we have $\sigma_k \leq \widetilde{K}_\alpha^{-\alpha-1/2} \lesssim x_k$. We can then use Lemma 3 with $m = 2$ on the event $A_k = \{|X_k| \leq x_k\}$, we get

$$T \gtrsim \phi(\sqrt{n}x_k)\sigma_k^3 \gtrsim \sigma_k^3.$$

Therefore, we obtain

$$\sum_{\{k>\widetilde{K}_\alpha\}\cap\mathcal{N}_n^c} E_{f_0}\left(\int \frac{1}{2}|f_k|^2 \, d\Pi(f \,|\, X)\mathbf{1}_{A_k}\right) \gtrsim \sum_{\{k>\widetilde{K}_\alpha\}\cap\mathcal{N}_n^c} \sigma_k^2 P_{f_0}(A_k).$$

For any $k \in \mathcal{N}_n^c$, we have $|f_{0,k}| \leq 1/\sqrt{n}$, for such k's, one has

$$P_{f_0}(A_k^c) \leq Pr(|N(0,1)| > 1) \leq 1/2.$$

This allows to bound the right hand side of the second to last display as

$$\sum_{\{k>\widetilde{K}_\alpha\}\cap\mathcal{N}_n^c} \sigma_k^2 P_{f_0}(A_k) \gtrsim \sum_{\{k>\widetilde{K}_\alpha\}\cap\mathcal{N}_n^c} \sigma_k^2.$$

Recalling $|\mathcal{N}_n| = N_n \leq \widetilde{K}_\alpha$ and using the fact that (σ_k) is non-increasing, we get

$$\sum_{\{k>\widetilde{K}_\alpha\}\cap\mathcal{N}_n^c} \sigma_k^2 \geq \sum_{k>2\widetilde{K}_\alpha} \sigma_k^2 \gtrsim K_\alpha^{-2\alpha},$$

where for the last inequality we use $\widetilde{K}_\alpha \asymp K_\alpha$. Finally, since $\alpha < \beta$, as n is large enough,

$$E_{f_0}\int \|f - f_0\|_2^2 \, d\Pi(f \,|\, X) \gtrsim K_\alpha^{-2\alpha} - K_\alpha^{-2\beta} \gtrsim K_\alpha^{-2\alpha}.$$

D. Proof for sparse Besov rate Theorem 3

Throughout this proof we might write $\mathcal{L}_n$ for logarithmic factors of the form $\log^D n$ and some constant $D > 0$, which may change from one line to another. Because we look at functions on $[0,1]$, the $L_{p'}$ spaces get smaller with p'. Since the minimax rate (given by (22)) is independent of p' in the regular region $\{p' \leq p\}$ we can reduce the case $p' \leq p$ to only $p' = p$. In the following of the proof we focus on the case $p \leq p' < \infty$. Recalling the definitions (20) of η, (21) of s' and (22) of r, we consider

$$2^{J_0} := n^{1-2r} \qquad \text{and} \qquad 2^{J_1} := (nF^2)^{\frac{1}{2(s-1/p)+1}}, \tag{D.1}$$

Note that since $s - 1/p + 1/2 > 0$, we have $J_0 \leq J_1$ for n large enough. Also note that in the case $\eta \leq 0$ we have $2^{J_1} = (nF^2)^{r/s'}$. Simple algebraic manipulations give

$$(p' - p)/2 + \eta(1 - 2r) = rp' \quad \text{if } \eta > 0, \tag{D.2}$$

$$(p' - p)/2 + \eta r/s' = rp' \quad \text{if } \eta \leq 0. \tag{D.3}$$

For a function $f \in L_2$ denote by $f^{[J_1]}$ and $f^{[J_1^c]}$ the projections of f onto the span of the wavelets $\{\psi_{jk}\}_{j<J_1,k}$ and $\{\psi_{jk}\}_{j\geq J_1,k}$ respectively. Consider $v_n := n^{-r}\log^{\delta} n$ the targeted rate, with $\delta > 0$ to be chosen large enough below. Using the triangle inequality for the $L_{p'}$-norm and a union bound yields

$$\begin{aligned}\Pi\left[\{f : ||f - f_0||_{p'} > v_n\} \,|\, X\right] \leq{}& \Pi\left[\{f : ||f^{[J_1^c]} - f_0^{[J_1^c]}||_{p'} > v_n/2\} \,|\, X\right] \\ &+ \Pi\left[\{f : ||f^{[J_1]} - f_0^{[J_1]}||_{p'} > v_n/2\} \,|\, X\right].\end{aligned}$$

We show that under E_{f_0} both of the previous displayed terms go to 0 with $n \to \infty$. Let us first take care of the indices $j \geq J_1$. Using Lemma E.2 for f_0 with $J^- = J_1$ and $J^+ = \infty$, we get

$$||f_0^{[J_1^c]}||_{p'} \lesssim \sum_{j\geq J_1} 2^{j(1/2-1/p')}||f_{0,j\cdot}||_{p'}.$$

Using $s - 1/p = s' - 1/p'$ and the embedding $\ell_p \subset \ell_{p'}$, available for $p' \geq p$

$$||f_0^{[J_1^c]}||_{p'} \lesssim \sum_{j\geq J_1} 2^{-js'}2^{j(1/2+s-1/p)}||f_{0,j\cdot}||_p \lesssim \sup_{j\geq J_1}\{2^{j(1/2+s-1/p)}||f_{0,j\cdot}||_p\} \sum_{j\geq J_1} 2^{-js'}.$$

The embedding $B^s_{pq}(F) \subset B^s_{p\infty}(F)$ shows that the supremum in the last display is bounded by F and

$$||f_0^{[J_1^c]}||_{p'} \lesssim F2^{-J_1 s'}.$$

When $\eta \leq 0$ the previous bound is equal to $F(nF^2)^{-r} \lesssim n^{-r}$. When $\eta > 0$, noting that $1/(1 + 2(s - 1/p)) \geq s/(s'(2s + 1))$ leads to

$$2^{-J_1 s'} = (nF^2)^{-\frac{s'}{1+2(s-1/p)}} \lesssim n^{-\frac{s}{2s+1}} = n^{-r}.$$

Therefore, it holds for n large enough that $||f_0^{[J_1^c]}||_{p'} \leq v_n/4$. So,

$$\Pi\left[\{f : ||f^{[J_1^c]} - f_0^{[J_1^c]}||_{p'} > v_n/2\} \,|\, X\right] \leq \Pi\left[\{f : ||f^{[J_1^c]}||_{p'} > v_n/4\} \,|\, X\right].$$

Once again, applying Lemma E.2 to $f^{[J_1^c]}$ with p' shows that it is sufficient to control, for a large enough constant $M > 0$,

$$E_{f_0}\Pi\left[\{f : \sum_{j\geq J_1} 2^{j(1/2-1/p')}||f_{j\cdot}||_{p'} > Mv_n\} \,|\, X\right]. \tag{D.4}$$

Using the summability of $(2^{-j/2})_j$ (note that this sequence plays the role of (z_k) in the proof of Theorem 1), we get the following bound

$$\sum_{j\geq J_1} 2^{j(1/2-1/p')}||f_{j\cdot}||_{p'} = \sum_{j\geq J_1} 2^{j(1/2-1/p')}(\sum_k |f_{jk}|^{p'})^{1/p'} \leq \sum_{j\geq J_1} 2^{j/2}\sup_k |f_{jk}| \lesssim \sup_{j\geq J_1,k}\{2^j|f_{jk}|\}.$$

Plugging the previous inequality in (D.4) and applying a union bound, we are left to control, for a large enough constant $M'>0$, and any event A_n,

$$E_{f_0}\Pi\left[\{f : \sup_{j\geq J_1,k}\{2^j|f_{jk}| > M'v_n\}\,|\,X\right] \leq \sum_{j\geq J_1}\sum_k E_{f_0}\left(\Pi\left[\{f : |f_{jk}| > M'2^{-j}v_n\}\,|\,X\right]\mathbf{1}_{A_n}\right) + P_{f_0}(A_n^c).$$

Writing ϕ for the standard Gaussian density and following the proof of Theorem 1, we can employ Lemma 3 with $m=0$ on coordinates (j,k) for x_{jk} to be chosen below, and obtain

$$E_{f_0}\left(\Pi\left[\{f : |f_{jk}| > M'2^{-j}v_n\}\,|\,X\right]\mathbf{1}_{A_n}\right) \lesssim E_{f_0}\left(\frac{1}{\phi(\sqrt{n}x_{jk})}2\overline{H}\left(\sigma_j^{-1}M'v_n2^{-j}\right)\mathbf{1}_{A_n}\right). \quad \text{(D.5)}$$

Now for any $j\geq J_1 \gtrsim \log_2 n$, since $\sigma_j = 2^{-j^2}$, for n large enough we have $\sigma_j^{-1}M'v_n2^{-j}\geq 1$ so we use assumption (H3) and obtain

$$\overline{H}\left(\sigma_j^{-1}M'v_n2^{-j}\right) \lesssim \frac{2^j\sigma_j}{M'v_n}.$$

To define the event A_n, first consider the sets of indices (the set Λ_n will be usefull further down the proof for indices $j<J_1$)

$$\Lambda_n := \left\{(j,k) : J_0 < j < J_1, |f_{0,jk}| \leq 1/\sqrt{n}\right\} \qquad \text{and} \qquad \mathcal{I} := \Lambda_n \cup \{(j,k) : j\geq J_1\}. \quad \text{(D.6)}$$

Now consider for any $l\geq 0$

$$\mathcal{A}_{jk,l} := \left\{|X_{jk}| \leq \sqrt{\frac{4(l+1)\log n}{n}}\right\},$$

and define the event

$$A_n := \bigcap_{J_0<j\leq\log_2 n\,,\,(j,k)\in\mathcal{I}} \mathcal{A}_{jk,0} \cap \bigcap_{l\geq 1}\bigcap_{l\log_2 n<j\leq(l+1)\log_2 n\,,\,(j,k)\in\mathcal{I}} \mathcal{A}_{jk,l}. \quad \text{(D.7)}$$

For $l\geq 0$, let us set $x_{jk} := \sqrt{(4(l+1)\log n)/n}$ whenever $l\log_2(n) < j \leq (l+1)\log_2 n$. For every $j > J_0$ we have $\sigma_j \lesssim 1/\sqrt{n}$ when n is large enough, therefore the constrains $|X_{jk}|\leq x_{jk}$ and $\sigma_j \lesssim x_{jk}$ are satisfied on the event A_n. We can use Lemma 3 with $m=0$ on the event A_n to obtain the bound in (D.5). Noting that $\phi(\sqrt{n}x_{jk}) \gtrsim n^{-2(l+1)}$ when $l\log_2(n) < j \leq (l+1)\log_2 n$ leads to

$$\sum_{j\geq J_1}\sum_k E_{f_0}\left(\Pi\left[\{f : |f_{jk}| > M'2^{-j}v_n\}\,|\,X\right]\mathbf{1}_{A_n}\right)$$

$$\lesssim \sum_{J_0<j\leq\log_2 n}\sum_k n^2\frac{2^j\sigma_j}{M'v_n} + \sum_{l\geq 1}\sum_{j=l\log_2 n-1}^{(l+1)\log_2 n}\sum_k n^{2(l+1)}\frac{2^j\sigma_j}{M'v_n}.$$

Recalling $\sigma_j = 2^{-j^2}$, $v_n = n^{-r}\log^\delta n$ and $J_0 \asymp \log_2 n$ the sums of the last display satisfy, as $n \to \infty$

$$\sum_{J_0 < j \le \log_2 n} \sum_k n^2 \frac{2^j \sigma_j}{M' v_n} = \sum_{J_0 < j \le \log_2 n} n^2 \frac{2^{2j} 2^{-j^2}}{M' v_n} \le \frac{n^4}{M' v_n} 2^{-J_0^2} \log_2 n = o(1),$$

$$\sum_{l \ge 1} \sum_{j=l\log_2 n - 1}^{(l+1)\log_2 n} \sum_k n^{2(l+1)} \frac{2^j \sigma_j}{M' v_n} = \sum_{l \ge 1} \sum_{j=l\log_2 n - 1}^{(l+1)\log_2 n} n^{2(l+1)} \frac{2^{2j} 2^{-j^2}}{M' v_n}$$
$$\le \frac{1}{M' v_n} \sum_{l \ge 1} n^{4(l+1)} 2^{-l^2 \log_2^2 n} = o(1).$$

Using Lemma 4 gives $P_{f_0}(A_n^c) \to 0$, as $n \to \infty$, we obtain

$$E_{f_0} \Pi \left[\{ f : ||f^{[J_1^c]} - f_0^{[J_1^c]}||_{p'} > v_n/2 \} \,|\, X \right] = o(1).$$

We are now left to control the indices $j < J_1$. Applying Lemma E.2 with the $L_{p'}$ norm, $J^- = -1$ and $J^+ = J_1$ and noting that $J_1 \lesssim \log n$, leads to

$$||f^{[J_1]} - f_0^{[J_1]}||_{p'} \lesssim (\log n)^{1-1/p'} \left(\sum_{j < J_1} \sum_k 2^{j(p'/2-1)} |f_{jk} - f_{0,jk}|^{p'} \right)^{1/p'}.$$

If $\delta > 0$ is large enough, it is then sufficient to control

$$E_{f_0} \Pi \left[\{ f : \sum_{j < J_1} \sum_k 2^{j(p'/2-1)} |f_{jk} - f_{0,jk}|^{p'} > M v_n^{p'} \} \,|\, X \right]. \tag{D.8}$$

Let $\mathcal{N}_n$ be the set of indices (j,k) with resolution level $j > J_0$ and high valued signal coefficients

$$\mathcal{N}_n := \left\{ (j,k) : j > J_0, \, |f_{0,jk}| > 1/\sqrt{n} \right\}. \tag{D.9}$$

For $f_0 = (f_{0,jk}) \in B^s_{pq}(F)$, and any (j,k) we have $|f_{0,jk}| \le ||f_{0,j\cdot}||_p \le F 2^{-j(s-1/p+1/2)}$. Therefore since $s - 1/p + 1/2 > 0$, we have

$$\mathcal{N}_n \subset \{ (j,k) : J_0 < j < J_1 \}. \tag{D.10}$$

Recall (D.6) the definition of the set Λ_n, we have the decomposition

$$\{ (j,k) : j < J_1 \} = \{ (j,k) : j \le J_0 \} \cup \mathcal{N}_n \cup \Lambda_n.$$

Using the triangle inequality and a union bound we split the sum in (D.8) into three terms according to the previous decomposition. The first of these three terms is

$$E_{f_0} \Pi \left[\{ f : \sum_{j < J_0} \sum_k 2^{j(p'/2-1)} |f_{jk} - f_{0,jk}|^{p'} > (M/3) v_n^{p'} \} \,|\, X \right].$$

Applying Markov's inequality reduces the study of this first term to obtain a bound on

$$\mathbf{A} := \sum_{j \le J_0} \sum_k 2^{j(p'/2-1)} E_{f_0} \int |f_{jk} - f_{0,jk}|^{p'} \, d\Pi(f \,|\, X). \tag{D.11}$$

From the convexity inequality $|x+y|^{p'} \lesssim |x|^{p'} + |y|^{p'}$, one gets

$$|f_{jk} - f_{0,jk}|^{p'} \lesssim |f_{jk} - X_{jk}|^{p'} + |X_{jk} - f_{0,jk}|^{p'}.$$

Under E_{f_0}, the observation (X_{jk}) in model (18) is distributed as $\mathcal{N}(f_{0,jk}, 1/n)$, therefore $E_{f_0}(|X_{jk} - f_{0,jk}|^{p'}) \lesssim n^{-p'/2}$. Along with $\mathbf{1}_{A_n} \le 1$, this leads to the bound,

$$\mathbf{A} \lesssim n^{-p'/2} \sum_{j \le J_0} 2^{jp'/2} + \sum_{j \le J_0} \sum_k 2^{j(p'/2-1)} E_{f_0} \left(\int |f_{jk} - X_{jk}|^{p'} \, d\Pi(f \,|\, X) \right). \tag{D.12}$$

Applying Lemma 2 with p' on coordinates $\{(j,k) : j \le J_0\}$, gives, for any $t > 0$

$$E_{f_0} \left(\int |f_{jk} - X_{jk}|^{p'} \, d\Pi(f \,|\, X) \right) \lesssim (t\sqrt{n})^{-p'} \left[1 + \log^{p'} E_{f_0} \left(\int e^{t\sqrt{n}|X_{jk} - f_{0,jk}|} \, d\Pi(f \,|\, X) \right) \right].$$

Note that since $f_0 \in B^s_{pq}(F)$, we have $|f_{0,jk}| \le F$, using Lemma 1 to bound the Laplace transform of the posterior with prior $f_{jk} = \sigma_j \zeta_{jk}$ and using $|x+y|^{p'} \lesssim |x|^{p'} + |y|^{p'}$,

$$E_{f_0} \left(\int |f_{jk} - X_{jk}|^{p'} \, d\Pi(f \,|\, X) \right) \lesssim (\sqrt{n}t)^{-p'} \left[1 + \log^{p'}(\sigma_j \sqrt{n}) + t^{2p'} + \log^{p'(1+\kappa)} \left(1 + \frac{F + 1/\sqrt{n}}{\sigma_j} \right) \right].$$

Now since $\sigma_j = 2^{-j^2}$ and $j \le J_0 \asymp \log n$ we have $\sigma_j^{-1} \le e^{C \log^2 n}$ for some $C > 0$. The bound then becomes, as n gets large enough,

$$E_{f_0} \left(\int |f_{jk} - X_{jk}|^{p'} \, d\Pi(f \,|\, X) \right) \lesssim (\sqrt{n}t)^{-p'} \left(1 + t^{2p'} + \log^{2p'(1+\kappa)} n \right).$$

This bound is optimized by $t^{2p'}$ of the order of the log factor in the last display which leads to

$$E_{f_0} \left(\int |f_{jk} - X_{jk}|^{p'} \, d\Pi(f \,|\, X) \right) \lesssim n^{-p'/2} \mathcal{L}_n. \tag{D.13}$$

Recall (D.1) the definition of J_0, plugging the previous bound in (D.12) leads to

$$\mathbf{A} \lesssim \mathcal{L}_n n^{-p'/2} \sum_{j \le J_0} 2^{jp'/2} \lesssim \mathcal{L}_n \left(\frac{2^{J_0}}{n} \right)^{p'/2} \lesssim \mathcal{L}_n n^{-rp'}.$$

This last bound shows that $\mathbf{A} = O(v_n^{p'})$ for $\delta > 0$ large enough, ensuring that, as $n \to \infty$,

$$E_{f_0} \Pi \left[\{ f : \sum_{j \le J_0} \sum_k 2^{j(p'/2-1)} |f_{jk} - f_{0,jk}|^{p'} > (M/3) v_n^{p'} \} \,|\, X \right] = o(1)$$

and thus taking care of the first term in the decomposition of (D.8). The second term we need to bound is the sum restricted on $\mathcal{N}_n$ (defined in (D.9)). Applying again Markov's inequality leave us to bound

$$\mathbf{B} := \sum_{(j,k)\in\mathcal{N}_n} 2^{j(p'/2-1)} E_{f_0} \int |f_{jk} - f_{0,jk}|^{p'} \, d\Pi(f \,|\, X). \tag{D.14}$$

On the set $\mathcal{N}_n \subset \{(j,k) : J_0 < j < J_1\}$, using the monotonicity of (σ_j) and recalling the definition (D.1) of J_1, we obtain $\log(\sigma_j^{-1}) \leq \log(\sigma_{J_1}^{-1}) \leq \log(2^{J_1^2}) \lesssim \log^2 n$. We now do the same split as in the term $\mathbf{A}$ and use Lemmas 2 and 1 to obtain the bound (D.13) in the term $\mathbf{B}$. Moreover, since $|f_{0,jk}| > 1/\sqrt{n}$ on $\mathcal{N}_n$ we have

$$\mathbf{B} \lesssim \mathcal{L}_n n^{-p'/2} \sum_{(j,k)\in\mathcal{N}_n} 2^{j(p'/2-1)} \lesssim \mathcal{L}_n n^{(p-p')/2} \sum_{(j,k)\in\mathcal{N}_n} 2^{j(p'/2-1)} |f_{0,jk}|^p.$$

Using first $\mathcal{N}_n \subset \{(j,k) : J_0 < j < J_1\}$ and then $||f_{0,j\cdot}||_p^p \leq F^p 2^{-jp(s-1/p+1/2)}$, we get

$$\mathbf{B} \lesssim \mathcal{L}_n n^{(p-p')/2} \sum_{J_0<j<J_1} 2^{j(p'/2-1)} ||f_{0,j\cdot}||_p^p \lesssim \mathcal{L}_n n^{(p-p')/2} \sum_{J_0<j<J_1} 2^{-j\eta},$$

where η is the index defined in (20). If $\eta < 0$, recalling that in this case $2^{J_1} = (nF^2)^{r/s'}$ (see (D.10) the definition of J_1) we employ equation (D.3) and obtain

$$\mathbf{B} \lesssim \mathcal{L}_n n^{(p-p')/2} 2^{-J_1\eta} \lesssim \mathcal{L}_n n^{(p-p')/2} n^{-\eta r/s'} \lesssim \mathcal{L}_n n^{-rp'}.$$

If $\eta > 0$, applying equation (D.2) we find

$$\mathbf{B} \lesssim \mathcal{L}_n n^{(p-p')/2} 2^{-J_0\eta} \lesssim \mathcal{L}_n n^{-rp'}.$$

Finally, when $\eta = 0$, we have $p' = (2s+1)p$ and therefore $(p'-p)/2 = sp = rp'$ such that $\mathbf{B} \lesssim \mathcal{L}_n n^{-rp'}$ (note that in the case $\eta = 0$ there is an extra log factor in $\mathcal{L}_n$). We have shown that $\mathbf{B} = O(v_n^{p'})$ and thus the second term in the decomposition of (D.8) satisfies, as $n \to \infty$,

$$E_{f_0}\Pi\left[\{f : \sum_{(j,k)\in\mathcal{N}_n} 2^{j(p'/2-1)} |f_{jk} - f_{0,jk}|^{p'} > (M/3) v_n^{p'}\} \,|\, X\right] = o(1).$$

The last term we need to control is the probability involving the sum restricted to Λ_n. Recalling that $|f_{0,jk}| \leq 1/\sqrt{n}$ on Λ_n and using $p' \geq p$, we can see that

$$\sum_{(j,k)\in\Lambda_n} 2^{j(p'/2-1)} |f_{0,jk}|^{p'} \leq n^{(p-p')/2} \sum_{J_0<j<J_1} \sum_k 2^{j(p'/2-1)} |f_{0,jk}|^p,$$

which is bounded as in the study of the $\mathbf{B}$ term by $\mathcal{L}_n n^{-rp'} \leq v_n^{p'}$ for $\delta > 0$ large enough. Using $|f_{jk} - f_{0,jk}|^{p'} \lesssim |f_{jk}|^{p'} + |f_{0,jk}|^{p'}$, it is then sufficient to control

$$E_{f_0}\Pi\left[\{f : \sum_{(j,k)\in\Lambda_n} 2^{j(p'/2-1)} |f_{jk}|^{p'} > (M/3) v_n^{p'}\} \,|\, X\right].$$

A union bound can be used as in the previous study of the indices $j > J_1$, this leaves us to control

$$\sum_{(j,k)\in\Lambda_n} E_{f_0}\Pi\left[\{f : |f_{jk}| > M'2^{-j}v_n\} \,|\, X\right].$$

Noting that $\Lambda_n \subset \{(j,k) : J_0 < j < J_1\}$ and that $J_0 \asymp \log n$, we follow the same steps as in the aforementioned study and show that the quantity in the last display goes to 0 with $n \to \infty$, achieving the control of the three terms and ensuring that, as $n \to \infty$,

$$E_{f_0}\Pi\left[\{f : ||f^{[J_1]} - f_0^{[J_1]}||_{p'} > v_n/2\} \,|\, X\right] = o(1),$$

which concludes the proof.

E. Additional Lemmas

Lemma E.1. *Let $\theta \sim HS(\tau)$, for $\tau > 0$ and $X|\theta \sim \mathcal{N}(\theta, 1/n)$. Then for some $C > 0$, it holds, for all $t \in \mathbb{R}$, $\theta_0 \in \mathbb{R}$, $\tau > 0$,*

$$E_{\theta_0} E_{\theta\sim HS(\tau)}\left[e^{t\sqrt{n}(\theta - X)} \,|\, X\right] \leq C \frac{\tau\sqrt{n}}{\log\left(1 + \frac{4\tau^2}{(|\theta_0|+1)^2}\right)} e^{t^2/2}$$

Proof. One can follow the proof of Lemma 1 and employ the sandwich inequality (8) for the Horseshoe to bound the denominator instead of the general heavy-tail lower bound (H2). □

Lemma E.2 (Bounding L_p–norms in terms of wavelet coefficients)**.** *Let $J^+ \geq J^- \geq -1$. Denote $\mathcal{J} := \{j : J^- \leq j \leq J^+\}$ and $\mathcal{K}_j := \{k : 0 \leq k < 2^j\}$ for any $j \geq -1$. Consider*

$$g := \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{K}_j} g_{jk}\psi_{jk}.$$

For any $p \geq 1$, we have

$$||g||_p \lesssim \sum_{j\in\mathcal{J}} 2^{j(1/2-1/p)} ||g_{j\cdot}||_p \qquad \text{and} \qquad ||g||_p^p \lesssim (J^+ - J^-)^{p-1} \sum_{j\in\mathcal{J}} 2^{j(p/2-1)} ||g_{j\cdot}||_p^p$$

Proof. Using the triangle inequality and the Parseval-like equality (12), we have

$$||g||_p = \|\sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{K}_j} g_{jk}\psi_{jk}\|_p \leq \sum_{j\in\mathcal{J}} \|\sum_{k\in\mathcal{K}_j} g_{jk}\psi_{jk}\|_p \lesssim \sum_{j\in\mathcal{J}} 2^{j(1/2-1/p)} ||g_{j\cdot}||_p.$$

Now simply apply Hölder's inequality with exponents p and $p/(p-1)$ to get the desired result. □

Lemma E.3. *Let A_n be the event defined in (D.7) and assume $f_0 \in B^s_{pq}(F)$, with $p, q \geq 1$ and $s, F > 0$. Then as $n \to \infty$, we have*

$$P_{f_0}(A_n^c) \to 0.$$

Proof. We notice that for any $(j,k) \in \mathcal{I}$, by definition, $|f_{0,jk}| \leq 1/\sqrt{n}$. Therefore the proof is similar to the single-index version of this lemma, Lemma 4, with l taking the role of $\log \ell$ therein, we have in particular, as n gets large enough, for all $(j,k) \in \mathcal{I}$ and $l \geq 0$

$$P_{f_0}(\mathcal{A}_{l,jk}^c) \leq P\left\{|\mathcal{N}(0,1)| > \sqrt{3(l+1)\log n}\right\} \leq n^{-\frac{3}{2}(l+1)}.$$

The last bound along with a union bound yield

$$P_{f_0}(A_n^c) \lesssim \sum_{l \geq 0} 2^{(l+1)\log_2 n} n^{-\frac{3}{2}(l+1)} \lesssim n^{-1/2} \sum_{l \geq 0} n^{-l/2} \lesssim 1/\sqrt{n},$$

ensuring $P_{f_0}(A_n^c) \to 0$ as $n \to \infty$. □