The Annals of Statistics
2024, Vol. 52, No. 4, 1433–1459
https://doi.org/10.1214/24-AOS2397

HEAVY-TAILED BAYESIAN NONPARAMETRIC ADAPTATION

BY SERGIOS AGAPIOU[1,a] AND ISMAËL CASTILLO[2,b]

[1] *Department of Mathematics and Statistics, University of Cyprus,* [a] *agapiou.sergios@ucy.ac.cy*
[2] *Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne University,* [b] *ismael.castillo@upmc.fr*

We propose a new Bayesian strategy for adaptation to smoothness in nonparametric models based on heavy-tailed series priors. We illustrate it in a variety of settings, showing in particular that the corresponding Bayesian posterior distributions achieve adaptive rates of contraction in the minimax sense (up to logarithmic factors) without the need to sample hyperparameters. Unlike many existing procedures, where a form of direct model (or estimator) selection is performed, the method can be seen as performing a *soft* selection through the prior tail. In Gaussian regression, such heavy-tailed priors are shown to lead to (near-)optimal simultaneous adaptation both in the L^2- and L^∞-sense. Results are also derived for linear inverse problems, for anisotropic Besov classes, and for certain losses in more general models through the use of tempered posterior distributions. We present numerical simulations corroborating the theory.

1. Introduction.

Adaptation to smoothness is a central topic in nonparametric statistics. In a regression setting to fix ideas, convergence rates of estimators of the unknown regression function generally depend on the assumed degree of smoothness and this raises the question of finding *adaptive* estimators, which can recover the unknown truth at (near-)optimal rate in the minimax sense, without assuming any prior knowledge of regularity. Popular non-Bayesian adaptation methods include Lepski's method [40], thresholding [25], and model selection [10]. Here we follow a Bayesian nonparametric approach and draw the unknown function randomly according to some prior distribution. In this setting, a possible way to derive 'adaptation' is by following a hierarchical Bayes principle: for instance, one first randomly draws a function of given regularity say $\alpha > 0$ and then draws α itself at random; this provides a hierarchical prior distribution which is then updated by conditioning on the observed data to form the posterior distribution. To give an example, initial prior draws can be, for instance, from an α-smooth Gaussian process (such as Brownian motion when $\alpha = 1/2$; and more generally, for example, Brownian motion integrated a fractional number of times); and α itself can be sampled according to a Gamma distribution.

In this work, we focus on prior distributions given in the form of a stochastic process characterised by a sequence of coefficients into an expansion basis. A popular related example in statistics and machine learning (e.g., [44]) is the one of Gaussian process priors, for which van der Vaart and van Zanten [57] proved the following generic result. Suppose we are in a simple one-dimensional nonparametric regression setting with Gaussian errors (e.g., a white noise model). If the true regression function is β-smooth in the Sobolev sense, and one considers an α-smooth Gaussian process as a prior distribution, then the posterior distribution contracts at rate

$$\varepsilon_n \lesssim \begin{cases} n^{-\beta/(1+2\alpha)} & \text{if } \alpha \geq \beta, \\ n^{-\alpha/(1+2\alpha)} & \text{if } \alpha \leq \beta, \end{cases} \tag{1}$$

and these rates cannot be improved [15].

Received August 2023; revised May 2024.
MSC2020 subject classifications. 62G05, 62G20.
Key words and phrases. Bayesian nonparametrics, frequentist analysis of posterior distributions, adaptation to smoothness, heavy tails, fractional posteriors.

In inverse problems and medical imaging, processes which feature tails heavier than Gaussian are increasingly used. The tails can be 'moderate', for instance, inbetween Gaussian and Laplace tails; see, for example, [2, 23, 39] and [12, 37] for numerical aspects, considering the case of so-called Besov priors; but also heavier than Laplace, including polynomially decreasing tails, such as recently considered, for example, in [48, 51, 52]. Yet overall there is up to now much less theoretical understanding of processes featuring a form of heavy tails.

The class of α-regular p-exponential measures for $p \in [1,2]$ is introduced in [4], where the authors model the coefficients onto a basis $\{\varphi_k\}$ of the function of interest with a prior with moderate tails (in-between Gaussian and Laplace) and variance decreasing as $k^{-1-2\alpha}$; therein the authors prove posterior contraction at rate

$$\varepsilon_n \lesssim \begin{cases} n^{-\beta/\{1+2\beta+p(\alpha-\beta)\}} & \text{if } \alpha \geq \beta, \\ n^{-\alpha/(1+2\alpha)} & \text{if } \alpha \leq \beta. \end{cases} \tag{2}$$

Here the index p corresponds to the tail behaviour of individual coefficients of the prior, with $p = 2$ recovering the Gaussian case (1) and $p = 1$ corresponding to Laplace (double-exponential) tails; and β again refers to Sobolev smoothness of the true regression function.

A few remarks in the light of (1)–(2) are:

- the optimal (minimax) rate is achieved in both cases for $\alpha = \beta$ (only);
- when $\alpha \geq \beta$ (the 'oversmoothing case'), when p decreases from 2 to 1, the rate slightly improves in terms of powers of n^{-1} from $\beta/(1+2\alpha)$ to $\beta/(1+\beta+\alpha)$;
- when $\beta \geq \alpha$ (the 'undersmoothing case'), the rate is always n^{-1} to the power $\alpha/(1+2\alpha)$, driven by the prior's own regularity.

As β is rarely known in practice, the previous prior distributions have to be made more complex if one wishes to derive *adaptation*. And indeed, two remarkable papers [35, 58] proved that adaptation in L^2-sense can be achieved from α-smooth Gaussian processes, up to logarithmic terms, by either using an additional rescaling variable, or by 'estimating' the prior's regularity α, either in a hierarchical Bayes or an empirical Bayes way (see also below for related references). Analogous adaptation results were derived very recently for p-exponential series priors in [5]; see also [31] for results on related priors in density estimation.

A natural question is thus what happens when tails heavier than Laplace are considered. This would formally correspond to taking the index p of p-exponential priors to be smaller than 1 and even going to 0. Let us 'do' the formal manipulation $p \to 0$ in (2), and see what happens. The 'limiting' rate would then become $n^{-\beta/(1+2\beta)}$ for $\alpha \geq \beta$. This would mean that at least for the 'oversmoothing' case, the 'adaptive' rate would be automatically obtained. Of course we just did a formal substitution that could perhaps not be valid; in particular, the techniques employed in [4] rely on the logarithmic concavity of p-exponential priors, a property which no longer holds for $p < 1$. We will see that, somewhat surprisingly at first, heavy tails enable, in a variety of settings, to derive adaptation in a fully automatic way.

The meaning of 'heavy tail' in the title and through the paper can be understood in the relatively loose sense of having tails that have a polynomial-type decrease, although the actual conditions under which we work are somewhat milder.

Heavy tailed series priors. Depending on the setting, we construct priors in $L^2 := L^2[0,1]$ via series expansions in either an orthonormal basis $\{\varphi_k : k \geq 1\}$ or an orthonormal boundary-corrected wavelet basis $\{\psi_{lk} : l \geq 0, k \in \mathcal{K}_l\}$ where $\mathcal{K}_l = \{0, \dots, 2^l - 1\}$ and where we denote the scaling function as the first wavelet ψ_{00}. Without loss of generality, we have taken the coarsest scale to be 1, while it is straightforward to accommodate for coarsest scales finer than 1. For more details, see [30], Section 4.3.

Prior Π *on functions.* For $\langle\cdot,\cdot\rangle$ the usual inner product on L^2, let $f_k := \langle f, \varphi_k\rangle$, respectively, $f_{lk} := \langle f, \psi_{lk}\rangle$, denote the coefficients of $f \in L^2$ onto the considered bases, so that

$$f = \sum_{k=1}^{\infty} f_k \varphi_k, \quad \text{or} \quad f = \sum_{l=0}^{\infty} \sum_{k \in \mathcal{K}_l} f_{lk} \psi_{lk}.$$

Let us define a prior Π on f by letting, for (σ_k), (s_l) sequences to be chosen below, and (ζ_k), (ζ_{lk}) independent identically distributed random variables of common law H with heavy tails, also to be specified,

$$f_k \overset{\text{ind.}}{\sim} \sigma_k \zeta_k, \tag{3}$$

in the case of a single-index basis (φ_k), or for a double-index basis

$$f_{lk} \overset{\text{ind.}}{\sim} s_l \zeta_{lk}. \tag{4}$$

A key choice of scale parameters σ_k and s_l throughout the paper is, for any $k \geq 1$ and $l \geq 0$,

$$\sigma_k = e^{-(\log k)^2}, \qquad s_l = 2^{-l^2}. \tag{5}$$

Another possible choice we consider, again for any such k, l and some $\alpha > 0$ is

$$\sigma_k = k^{-1/2-\alpha}, \qquad s_l = 2^{-l(1/2+\alpha)}. \tag{6}$$

The choice (6) corresponds to the same scaling as in (2) and, since here we consider heavy tails, to (formally at least) setting $p = 0$ for a p-exponential prior. Contrary to (6), where the value of α should be chosen, note that (5) is in principle free of any parameter. The choice of the square in (5) is mostly to fix ideas and results below also hold for $\sigma_k = e^{-a(\log k)^{1+\delta}}$ with any given constants $a, \delta > 0$. The fact that (σ_k) in (5) decreases faster than any polynomial in k^{-1}, but not exponentially fast (as, for instance, e^{-k}), is key for the results ahead. Similarly, for double-index bases we can use $s_l = 2^{-l^{1+\delta}}$ for any fixed $\delta > 0$ in (5).

To complete the prior's description, let us now specify the distribution H of the ζ variables as above. Suppose that H admits a density h on $\mathbb{R}$ and that for $c_1 > 0$ and $\kappa \geq 0$,

$$h \text{ is symmetric, positive, bounded and decreasing on } [0, \infty), \tag{7}$$

$$\log\bigl(1/h(x)\bigr) \leq c_1\bigl(1 + \log^{1+\kappa}(1+x)\bigr), \quad x \geq 0. \tag{8}$$

A leading example throughout the paper is the case $\kappa = 0$ corresponding to polynomial tails (sometimes called fat tails): Student distributions, including Cauchy, satisfy these conditions for c_1 large enough constant. Yet, some flexibility is allowed with $\kappa > 0$ permitting slightly lighter tails. Depending on the setting, we sometimes assume a mild integrability or moment condition, that will still accommodate most Student-type tails.

We call priors Π verifying (5)–(7)–(8) *oversmoothed heavy-tailed* priors or simply OT-priors while HT(α) for α *heavy-tailed* priors stand for those satisfying (6)–(7)–(8).

Frequentist analysis of posterior distributions. Consider a statistical model $\{P_f^{(n)}, f \in \mathcal{F}\}$ indexed by a function f with observations $X = X^{(n)}$. Examples considered below include nonparametric regression, density estimation and classification. Given a prior distribution Π on f, the Bayesian model sets $X|f \sim P_f^{(n)}$ and $f \sim \Pi$. The posterior distribution $\Pi[\cdot|X]$ is the conditional distribution $f|X$. Assuming the model is dominated, the posterior is given as usual by Bayes' formula. Taking a frequentist approach, we analyse the posterior $\Pi[\cdot|X]$ under the assumption that X has actually been generated from $P_{f_0}^{(n)}$ for some fixed true function f_0. We refer to the book [28] for more context and references.

Classical regularity balls. Before describing our main results, let us recall three types of standard smoothness assumptions on the underlying truth f_0: Sobolev, Hölder and Besov.

[Sobolev-type] When working with an orthonormal basis $\{\varphi_k\}$, we consider Sobolev-type assumptions. Recalling that $f_k = \langle f, \varphi_k \rangle$, for $\beta, L > 0$, denote

$$\mathcal{S}^\beta(L) = \Big\{ f = (f_k), \sum_{k \geq 1} k^{2\beta} f_k^2 \leq L^2 \Big\}. \tag{9}$$

For certain choices of φ_k, the sets $\mathcal{S}^\beta(L)$ correspond to balls of classical Hilbert–Sobolev spaces of functions in $L^2[0, 1]$ possessing β square integrable derivatives.

When working with an orthonormal wavelet basis $\{\psi_{lk}\}$ we consider either hyper-rectangle (Hölder-type) or Besov-type assumptions.

[Hölder-type] For $f_{lk} = \langle f, \psi_{lk} \rangle$ and $\beta, L > 0$, let

$$\mathcal{H}^\beta(L) = \Big\{ f = (f_{lk}), \max_{k \in \mathcal{K}_l} |f_{lk}| \leq 2^{-l(1/2+\beta)} L \text{ for all } l \geq 0 \Big\}. \tag{10}$$

For wavelet bases with classical Hölder regularity higher than β, the sets $\mathcal{H}^\beta(L)$ correspond to L-balls of the Hölder-Zygmund spaces $\mathcal{C}^\beta[0, 1]$, see [30], Section 4.3. For noninteger β the latter spaces coincide with the classical Hölder spaces $C^\beta[0, 1]$, while for β an integer it holds $\mathcal{C}^{\beta'} \subset C^\beta \subset \mathcal{C}^\beta$ for all $\beta' > \beta$ where inclusions are all strict.

[Besov-type] For $\beta, L > 0$ and $1 \leq r \leq 2$, let

$$\mathcal{B}^\beta_{rr}(L) = \Big\{ f = (f_{lk}), \sum_{l \geq 0} 2^{rl(\beta+1/2-1/r)} \sum_{k \in \mathcal{K}_l} |f_{lk}|^r < L^r \Big\}. \tag{11}$$

Again for appropriate wavelet bases, the sets $\mathcal{B}^\beta_{rr}(L)$ correspond to L-balls of Besov spaces $B^\beta_{rr}[0, 1]$ defined via moduli of continuity, see [30], Section 4.3. For $r = 2$, Besov spaces coincide with the Hilbert–Sobolev spaces, while for $r < 2$ Besov spaces are useful for modelling spatially inhomogeneous functions, that is functions which are smooth in some areas of the domain and irregular or even discontinuous in other areas, see [24] or [34], Section 9.6. Here, we restrict to Besov spaces $\mathcal{B}^\beta_{rq}$ with $r = q$ for simplicity, see Section 5 for a discussion.

Outline and informal description of the results. In what follows, we will substantiate the intuition that heavy-tailed series priors achieve adaptation to smoothness without the need to sample any hyperparameters. Our results show that OT-priors are fully adaptive, and that HT(α)-priors are partially adaptive (essentially) for smoothness of the truth $\beta \leq \alpha$. More precisely, in Section 2 we consider white noise regression and show (near-) adaptive posterior contraction rates in the minimax sense in the following settings:

- in L^2-loss for Sobolev regularity in both the direct and a linear inverse problem setting;
- in L^∞-loss under Hölder smoothness in the direct setting;
- in L^2-loss under (spatially inhomogeneous) Besov smoothness in the direct setting.

A result on the limiting shape of the posterior distribution is also given, in the form of an adaptive nonparametric Bernstein–von Mises theorem. In Section 3, we establish generic bounds for the mass that heavy-tailed priors put on L^2 and L^∞-balls around Sobolev and Hölder truths, respectively. By themselves, such bounds allow the derivation of contraction rates for *tempered* posterior distributions in general models, in terms of Rényi divergence. Indeed, we exemplify this approach in three nonparametric settings, in which we achieve (near-) adaptive rates of contraction of tempered posteriors in the minimax sense:

- in density estimation, in L^1-loss and under Hölder smoothness of the true log-density;
- in binary classification, in an L^1-type loss and under Sobolev smoothness of the logit of the true binary regression function;
- in (direct) white noise regression, in L^2-loss and under Besov smoothness of the truth.

A simulation study is presented in Section 4, while a brief discussion and review of open questions can be found in Section 5. Proofs are presented in Section 6 as well as in the Supplementary Material [3]. The Supplement [3] also includes additional simulations and a discussion on extending the results of Section 3 to contraction of standard posteriors.

Comparison with other priors. While our results below shall answer positively the question of obtaining adaptation with heavy-tailed series priors, it is of interest to compare our priors with other priors leading to adaptation. The list below is by far not exhaustive; we mention a few classes of priors that bear some similarity with the priors here considered.

- *Sieve priors* (e.g., [8, 45, 47, 49]). Here adaptation is obtained by truncating the series prior and taking the truncation parameter K random; note that the distribution on the modelled K coefficients in general plays little role on the obtained rate. In regression, the above references show that (near)-optimal adaptive rates are achieved in the L^2-norm; but generally this is not the case in the L^∞ norm ([21], Theorem 5). In contrast, we will see that heavy-tailed series priors in white noise regression are adaptive in both norms.
- *Spike-and-slab priors and sparsity inducing priors.* Due to their links to thresholding rules, spike-and-slab (SAS) priors are also particularly natural: [32] show in white noise regression that SAS posteriors achieve adaptive rates both in L^2 (nearly) and L^∞ (there are few results in other models, except [18, 42] in density estimation). While heavy-tailed priors share the same desirable properties, they do not model sparsity so have no 'mass at 0' part; this can be an advantage computationally, as in more complex models, sampling from SAS posteriors typically requires exploration of a combinatorial number of models. While sampling from OT posteriors is relatively easy using MCMC, we are not aware of posterior samplers for SAS in density estimation, for instance (the posterior in [18] is computable but uses partial conjugacy and is limited to regularities up to 1). The horseshoe prior [14] is in a sense closer to our proposal as it has density with Cauchy tails. Note though that similarly to SAS priors and unlike our heavy-tailed prior construction it directly models sparsity through a diverging density at 0. We expect that horseshoe priors have good adaptation properties, although we do not know any proof in a nonparametric context ([14] present simulations in one such setting)—the techniques we develop here could be used precisely to derive such results.
- *Mixtures.* It may be argued that heavy-tailed distributions can be represented as mixtures of lighter tailed distributions: for instance, Laplace and Student distributions are scale mixtures of normals. So, one could view the heavy-tailed prior in a hierarchical manner with independent Gaussians and one hyper-parameter per coefficient. Note, however, that this in general does not suffice for adaptation: for instance, Laplace series priors do not adapt optimally; and even if the resulting distribution on coordinates is, for example, Student, the choice of scale parameters σ_k or s_l in (5)–(6) is essential as, for instance, (6) does not adapt if $\alpha < \beta$. This shows that only some well-designed mixture priors work. Furthermore, even if a heavy-tailed law has a mixture representation, this does not mean that it is advantageous computationally to use it (e.g., using a Gibbs sampling to approximate the posterior; this may face computational difficulties due to the high number of hyper-parameters), and in fact we do not do so in Section 4, where we use direct sampling from the posterior via MCMC in all considered examples. Also related to mixtures, [26] construct a hierarchical block-prior that enables to derive contraction rates in L^2-sense (or with testing distances) without additional logarithmic terms. The resulting construction requires a specific hyperprior, and may be difficult to sample from in complex settings (e.g., beyond white noise regression); also, although optimal in L^2 it is presumably suboptimal in the L^∞-sense.

2. Nonparametric regression.

To avoid technicalities independent of the ideas at stake, we focus in this section on the Gaussian white noise model, that can be seen as the prototypical nonparametric model [30, 56]. Up to dealing with discretisation effects, similar results as the ones below are expected to hold also, for example, for fixed-design nonparametric regression. For $f \in L^2$ and $n \geq 1$, the Gaussian white noise model writes

$$dY^{(n)}(t) = f(t)\,dt + dW(t)/\sqrt{n}, \quad t \in [0,1], \tag{12}$$

where W is standard Brownian motion.

2.1. *L^2-loss and Sobolev smoothness.*

By projecting (12) onto a single-index orthonormal basis $\{\varphi_k\}$ of L^2, one obtains the normal sequence model, with $f_k = \langle f, \varphi_k\rangle$,

$$X_k \,|\, f_k \sim \mathcal{N}(f_k, 1/n), \tag{13}$$

independently for $k \geq 1$, with $X_k = \int_0^1 \varphi_k(t)\,dY^{(n)}(t)$. We denote $X = X^{(n)} = (X_1, X_2, \dots)$ the corresponding observation sequence. Here for simplicity of notation, we consider only single-index bases, but the results in the present section and the next hold as well for double-indexed wavelet bases, such as ones considered in Section 2.3, with the corresponding appropriately chosen scalings as in (5)–(6).

Early results for Bayesian series priors (we discuss a few other priors in Section 5) in this setting include [63], who established nonadaptive convergence rates for the posterior mean under Gaussian priors, while [11] derived adaptive rates using a hyperprior over a discrete set of regularities. Still, for Gaussian series priors, in [54] partial adaptation was achieved with fixed regularity using either a hierarchical or an empirical Bayes choice of a universal scaling parameter provided the truth is not too smooth compared to the prior, while in the work [35] full adaptation (up to logarithmic factors) was established using either a hierarchical or an empirical Bayes choice of the prior regularity. Gaussian series priors on manifolds with an extra random time parameter were shown to be adaptive to smoothness in broad geometric contexts [17]. More recently, both fixed-regularity and adaptive results were derived for p-exponential priors in [4] and [5].

In this section, we consider series priors as in (3)–(5), defined via a heavy-tailed density h satisfying the moment assumption, for some $q \geq 1$,

$$\int_{-\infty}^{\infty} |x|^q h(x)\,dx < \infty. \tag{14}$$

THEOREM 1. *In the regression model* (13), *consider the heavy-tailed series prior* (3) *with parameters specified by* (5) *and* (7)–(8) *as well as* (14) *with $q = 2$. Suppose $f_0 \in \mathcal{S}^\beta(L)$ for some $\beta, L > 0$. Then as $n \to \infty$,*

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

where $\mathcal{L}_n = (\log n)^d$ for some $d > 0$. Further, the same result holds, with a possibly different d, for the choice of (σ_k) as in (6), *provided $\alpha \geq \beta$. Both results also hold for truncated priors at $k = n$, that are the same as the ones considered except they set $f_k = 0$ for $k > n$.*

Theorem 1 shows that the oversmoothed heavy-tailed (OT) prior as in (5)–(8) leads to full adaptation to smoothness $\beta > 0$, without any restriction to the range over Sobolev balls S^β (see also Remark 2) and without the need of tuning of any smoothness hyper-parameter. The fact that the second part of Theorem 1 holds shows that the heuristic presented below (2) letting $p \to 0$ is correct: if (σ_k) is polynomially decreasing as $\sigma_k = k^{-1/2-\alpha}$ as in (6), then adaptation holds in the range $\beta \in (0, \alpha]$, that is exactly in the case the prior 'oversmooths' the truth, as expected from formula (2). For a comparison with other priors and more discussion, we refer to Section 5.

A difficulty with the proof of Theorem 1 is that it does not seem possible to use the general approach to posterior convergence rates as in [27, 28], as the latter requires exponential decrease of probabilities of sieve sets (at least with infinite series priors or priors modelling high-frequencies, so excluding sieve priors, for which specific arguments can be used, see, for example, [8, 45, 49]), which is essential in being able to discard regions of the parameter spaces, as crucially used in results for Gaussian or p-exponential priors (the latter just allow for exponential decrease when $p = 1$). Our proof is based on a detailed analysis of the posterior induced on coefficients, the most delicate part being high-frequencies, for which careful compensations from numerator and denominator in the ratios arising from Bayes' formula are needed. We note that the results in the present section impose a moment condition on the heavy-tailed density h; this is mostly for technical convenience: it is expected that existence of a second moment is required in the theorem above, as its proof goes through controlling the posterior second moment. It is likely that one can remove the moment condition by requiring a control in probability only; such approach would allow to include the Cauchy density, but the proof would likely be more technical, so we refrain from pursuing this goal here; we only note that in this vein results for the Cauchy prior are derived in Section 3.

Numerical intuition behind the result. Underlying our proofs, is the behaviour of heavy-tailed priors on $\mu \in \mathbb{R}$ in the model $X|\mu \sim \mathcal{N}(\mu, 1/n)$, which we compare here to the behaviour of Gaussian priors. Consider $\mu \sim \sigma\Pi$ where σ is a positive scaling and Π is either standard normal or say a standard Student distribution with 3 degrees of freedom. Recall that in the Gaussian prior case, the posterior mean is given as $E[\mu|X] = n\sigma^2 X/(1 + n\sigma^2)$. Figure 1 depicts the posterior mean in the Student prior case as a function of X, for decreasing values of σ and noise level $1/\sqrt{n} = 10^{-3.5}$. We observe that in the Student prior case, for large prior scalings σ the posterior mean is given by the observation (this is similar to the Gaussian prior case), while for small σ the posterior mean resembles a thresholding estimator, nearly setting to zero small observations and preserving larger observations (this is unlike the Gaussian prior case where observations are shrunk by a constant factor determined by the size of σ relative to the noise precision n). In particular, Figure 1 suggests that with the Student prior, good recovery is achieved by the posterior mean independently of the size of the scaling σ, for $|X| \geq 0.002 \gg 1/\sqrt{n} \approx 0.0003$. Contrast this to the Gaussian prior case, where small σ leads to poor recovery of large observations. It thus appears, that an oversmoothing heavy-tailed prior may still have good nonparametric behaviour, despite the scaling being mismatched.

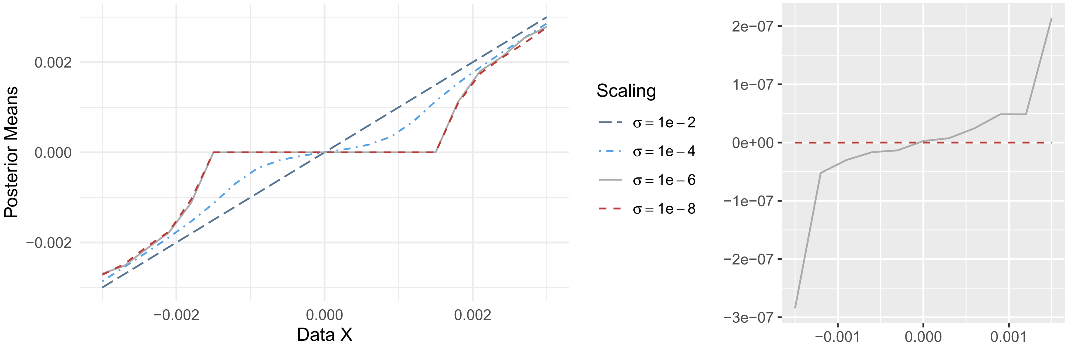

FIG. 1. *Left*: *posterior means for the univariate model* $X|\mu \sim N(\mu, 10^{-7})$, $\mu \sim \sigma\Pi$, *plotted against the observed data* X, *for* Π *a standard Student distribution with* 3 *degrees of freedom and for* 4 *values of the scaling* σ. *Right*: *detailed view of the center region of the plot on the left.*

REMARK 1 (Optimal rate and log factors). The rate in Theorem 1 is optimal up to a logarithmic factor (it can be checked, for instance, that one can take a power $d = 1$ in $\mathcal{L}_n$ for the OT prior), which we did not try to optimise. A main reason is that work by Tony Cai [13] shows that any method that is smoothness-adaptive in L^2 and separable, in the sense that it makes coordinates independent, must pay a logarithmic price in its convergence rate. Moreover, the squared-rate cannot be better than $(\log n/n)^{2\beta/(2\beta+1)}$ for some β, a rate (nearly) achieved for tempered posteriors in Theorem 10 below, see Remark 3 and (22).

REMARK 2 (Smoothness and order of basis). Theorem 1 and results below hold for any smoothness parameter $\beta > 0$, over regularity balls defined by coefficients as before. As usual with estimators defined over bases, if one wishes results over classical Hölder spaces or Besov spaces defined via moduli of continuity, one needs to assume a basis of large enough order/regularity (which means adaptation holds in that case over $\beta \le \beta_{\max}$, where $\beta_{\max}$ can be made as large as desired by choosing the order of the basis large enough).

2.2. *Linear inverse problems, Sobolev smoothness.* A synthetic prototypical model in linear inverse problems arises when projecting onto the SVD of the forward operator: the observation model is, for some $\nu \ge 0$, independently for $k \ge 1$,

$$X_k | f_k \sim \mathcal{N}(\kappa_k f_k, 1/n), \quad \kappa_k \asymp k^{-\nu}. \tag{15}$$

This generalises the former signal-in-white-noise setting with the 'inverse' nature of the problem represented by the sequence (κ_k). This model has been much studied in terms of minimaxity and adaptation over Sobolev smoothness in the frequentist literature [22]. Following a Bayesian approach with Gaussian priors, [36] derived posterior contraction rates in the nonadaptive case of fixed regularity; again in [35], the authors proved that empirical and hierarchical Bayes approaches could be used to derive adaptation for the previous class of Gaussian priors. Results for sieve priors were obtained in [45] (see also Section 5 for more on this).

THEOREM 2. *In the inverse regression model* (15) *with degree of ill-posedness of the forward operator* $\nu \ge 0$, *consider the heavy-tailed series prior* (3) *with parameters specified by* (5) *and* (7)–(8) *as well as* (14) *with* $q = 2$. *Suppose* $f_0 \in \mathcal{S}^\beta(L)$ *for some* $\beta, L > 0$. *Then, as* $n \to \infty$,

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

where $\mathcal{L}_n = (\log n)^d$ *for some* $d > 0$. *Further, the same result holds, with a possibly different* d, *for the choice of* (σ_k) *as in* (6), *provided* $\alpha \ge \beta$.

Theorem 2 generalises Theorem 1 (the case $\nu = 0$). With the technique and estimates of the proof of Theorem 1 in hand, its proof via coordinate-estimates is relatively simple; this is in contrast to existing empirical or hierarchical Bayes proofs in this setting for infinite series Gaussian priors [35], for which the study of the marginal maximum likelihood estimates (or of the posterior on the smoothness parameter) requires nontrivial work.

2.3. *L^∞-loss in white noise regression.* Let us now consider estimation under the L^∞-loss, which is a particularly desirable cost function in curve estimation, as a bound in supremum norm guarantees visual closeness between curves. In this case, we expand in a wavelet orthonormal basis, and the projection of model (12) becomes a normal sequence model

$$X_{lk} | f_{lk} \sim \mathcal{N}(f_{lk}, 1/n), \tag{16}$$

independently over relevant indices l, k. Again, we denote by $X^{(n)}$ the observation sequence. Minimaxity and adaptation in supremum loss for Hölder smoothness have been studied extensively in the frequentist literature [30, 33, 41]. Deriving results for Bayesian procedures (or more generally for likelihood-based procedures, see, for example, the notes of Chapter 7 of [30]) for this loss is in general quite delicate. Generic results in this direction include [16, 29], but most existing results are concerned with specific models and priors. In white noise regression, supremum norm adaptation has been derived so far for spike-and-slab priors [32], and tree priors *à la* Bayesian CART [21] (up to an unavoidable log factor).

THEOREM 3. *In the regression model, let us consider the prior on f induced by the heavy-tailed wavelet series prior* (4) *on coefficients f_{lk} in* (16), *with parameters specified by* (5) *and* (7)–(8) *as well as* (14) *with $q \geq 1$. Suppose $f_0 \in \mathcal{H}^\beta(L)$ for some $\beta, L > 0$. Then, as $n \to \infty$,*

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

where $\mathcal{L}_n = (\log n)^d$ for some $d > 0$. Further, the same result holds, with a possibly different d, for the choice of (s_l) as in (6) *and under* (14) *with any $q \geq 1$, provided $\alpha \geq \beta + 1/q$.*

Theorem 3 shows that the OT prior attains the adaptive minimax supremum-norm rate up to a logarithmic term. This is the first result of this kind for a prior distribution that does not have a 'spike' part that sets coefficients to 0 (as opposed to spike-and-slab and tree priors, that select a subset of coefficients and set all others to 0). For the HT(α) prior, the condition $\alpha \geq \beta + 1/q$ is for technical reasons and may be suboptimal (note though that as $q \to \infty$ the condition becomes milder and approaches the one in Theorem 1 for the L^2-loss).

2.4. *Besov classes.* We now consider the case of unknown functions f_0 that can be spatially inhomogeneous in the projected white noise model (16). In particular, we study adaptation of heavy-tailed priors for underlying true functions with Besov smoothness $\mathcal{B}^\beta_{rr}(L)$ for $1 \leq r < 2$. Minimaxity and adaptation over spatially inhomogeneous Besov spaces in this model have been studied in the frequentist setting in [24]. A distinctive feature is that linear estimators are provably suboptimal by a polynomial factor. More recently, rates of contraction in the nonadaptive case of fixed regularity were established using undersmoothing and appropriately rescaled p-exponential priors with $p \leq r$ in [4]. Adaptation with p-exponential priors, $p \leq r$, was achieved in [5], Theorem 2.5, using either a hierarchical or an empirical (marginal maximum likelihood) Bayes choice of both the regularity and scaling parameters (simultaneously). Importantly, it was established in [6] that Gaussian priors suffer from the same suboptimality as linear estimators in the frequentist setting. The next result establishes adaptation with heavy-tailed priors in this setting as well.

THEOREM 4. *In the white noise regression model, let Π be a heavy-tailed prior generated by* (4) *with parameters specified by* (5) *and* (7)–(8) *as well as* (14) *with $q = 2$. Suppose $f_0 \in \mathcal{B}^\beta_{rr}(L)$ for some $\beta, L > 0$, $r \in [1, 2]$ and $\beta > 1/r - 1/2$. Then, as $n \to \infty$,*

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

where $\mathcal{L}_n = (\log n)^d$ for some $d > 0$. Further, the same result holds, with a possibly different d, for the choice of (σ_k) as in (6), *provided $\alpha \geq \beta$.*

The assumption $\beta > 1/r - 1/2$ is sharp, in the sense that it is the weakest assumption on β ensuring $B^\beta_{rr} \subset L^2$, [55], Theorem 3.3.1, a necessary condition since we consider contraction in L^2-loss. This is in contrast to [5], Theorem 2.5, which has a more stringent assumption on β (arising due to the necessity of controlling the prior mass uniformly with respect to the scaling prior-parameter, see Remark 2.6(a) in that source). Another advantage of the above result is that it provides a very simple and practical framework for achieving adaptation for spatially inhomogeneous functions which, as discussed in Section 3.4 below, goes beyond the regression setting. In particular, from the practical point of view and unlike [5], Theorem 2.5, approximating posteriors in this framework does not require the use of a Gibbs sampler, which may mix poorly in high dimensions when the data are informative [1], neither does it require implementing marginal maximum likelihood estimators of prior parameters which can be technically difficult, especially in more involved models.

2.5. *Adaptive nonparametric Bernstein–von Mises theorem.* Beyond convergence rates, one may be interested in the limiting shape of the posterior distribution, as formalised in nonparametric settings in [19, 20], from which one can deduce, for instance, uncertainty quantification results on certain functionals of f, as in the application below Theorem 5.

For monotone increasing weighting sequences $w = (w_l)_{l\geq 0}$, $w_l \geq 1$, we define multi-scale sequence spaces

$$\mathcal{M} \equiv \mathcal{M}(w) \equiv \left\{ x = \{x_{lk}\} : \|x\|_{\mathcal{M}(w)} \equiv \sup_l \frac{\max_k |x_{lk}|}{w_l} < \infty \right\}. \tag{17}$$

The space $\mathcal{M}(w)$ is a nonseparable Banach space (it is isomorphic to ℓ_∞). However, the following space $\mathcal{M}_0$ forms a separable closed subspace for the same norm

$$\mathcal{M}_0 = \mathcal{M}_0(w) = \left\{ x \in \mathcal{M}(w) : \lim_{l\to\infty} \max_k \frac{|x_{lk}|}{w_l} = 0 \right\}. \tag{18}$$

The white noise model (12) can be rewritten as the sequence model $X^{(n)} = f + \mathbb{W}/\sqrt{n}$, where in slight abuse of notation f is identified with the sequence of its coefficients $f = (f_{lk})$ and $\mathbb{W} = (\int \psi_{lk}\, dW(t))_{l,k}$ has the distribution of an i.i.d. sequence of $\mathcal{N}(0, 1)$ variables. It is not hard to see that $\mathbb{W}$ almost surely belongs to $\mathcal{M}_0$ (and $X^{(n)}$ as well for $f \in L^2$) under the condition that w_l diverges faster than $\sqrt{l}$, see [20]. Denote $\tau : f \to \sqrt{n}(f - X^{(n)})$. Then $\Pi[\cdot|X^{(n)}] \circ \tau^{-1}$ denotes the induced posterior distribution on $\mathcal{M}_0$, shifted and rescaled by τ.

For S a given metric space, let $\beta_S(P, Q)$ denote the bounded-Lipschitz metric over probability distributions P, Q on S. It is well known that β_S metrises weak convergence on S.

THEOREM 5. *In the regression model, let us consider the heavy-tailed wavelet series prior* (4) *on coefficients* f_{lk} *in* (16), *with parameters specified by* (5) *and* (7)–(8). *Consider the multiscale space* $\mathcal{M}_0$ *as in* (18) *with* $w_l = l^{1+\kappa+\varepsilon}$ *for some* $\varepsilon > 0$. *Suppose* $f_0 \in \mathcal{H}^\beta(L)$ *for some* $\beta, L > 0$. *Then*

$$E_{f_0} \beta_{\mathcal{M}_0}(\Pi[\cdot|X^{(n)}] \circ \tau^{-1}, \mathcal{L}(\mathbb{W})) \to 0$$

as $n \to \infty$, *where* $\mathcal{L}(\mathbb{W})$ *denotes the law of a Gaussian white noise in* $\mathcal{M}_0$. *Further, the same result holds for* (s_l) *as in* (6) *with* $w_l = l^{(1+\kappa+\varepsilon)/2}$, *for some* $\varepsilon > 0$ (*and any* $\alpha > 0$).

The heavy-tailed priors we consider thus automatically satisfy an adaptive nonparametric Bernstein–von Mises theorem [46]. As opposed to results of this type obtained in the literature so far, note that we do not need to modify the prior to impose a *flat initialisation*; [46] proves that this is necessary for spike-and-slab priors: for these one needs to remove the spikes from a slowly increasing number of coordinates to allow for Gaussian finite-dimensional distributions in the limit (otherwise small signals can erroneously be classified into the spike part by the posterior; the same holds for Bayesian CART, see [21]). Here the heavy-tailed prior is continuous, so asymptotic normality holds even for arbitrarily low frequencies: this is because the prior induced on the first coordinates has a continuous and positive density on the whole real line, so the conditions of (a version of) the parametric Bernstein–von Mises theorem are satisfied, see the proof of Theorem 5 in the Supplementary Material [3].

Application. An implication of Theorem 5 is the following (using [20], Theorem 4): a Donsker-type theorem holds for the posterior distribution with heavy-tailed priors when estimating the primitive $F(\cdot) = \int_0^{\cdot} f(u)\,du$ of f: for B standard Brownian motion and $\mathcal{L}(\|B\|_\infty)$ the distribution of its supremum on $[0, 1]$, in P_0-probability,

$$\beta_{\mathbb{R}}(\mathcal{L}(\sqrt{n}\|F(\cdot) - X^{(n)}(\cdot)\|_\infty | X^{(n)}), \mathcal{L}(\|B\|_\infty)) \to 0,$$

where $\mathcal{L}(F(\cdot)|X^{(n)})$ denotes the posterior distribution on F induced from the posterior on f through the primitive map $f \to F$. From this, one immediately deduces that supremum-norm quantile credible bands for F are asymptotically optimal (efficient) confidence bands for F_0, see [20] for details and discussion.

3. Prior mass bounds and ρ-posterior convergence.

In this section, we first derive lower bounds for the prior mass that heavy-tailed priors put on L^2- and L^∞-balls, around Sobolev and Hölder functions. These bounds next enable us to obtain contraction rates for tempered posteriors (ρ-posteriors) in a variety of nonparametric settings: as examples, we consider density estimation, binary classification and regression (the latter under Besov regularity of the truth). As a slight variant to the moment assumption (14), here we require the tail condition: for some $c_2 > 0$,

$$\overline{H}(x) := \int_x^\infty h(u)\,du \le c_2/x^2, \quad x \ge 1. \tag{19}$$

Condition (19) allows for most Student distributions; Cauchy tails can also be accommodated, see Remark 5 below and Remark B.1 in the Supplementary Material [3].

3.1. *Generic prior mass results.*

THEOREM 6 (Generic prior mass condition in L^2). *For Π a prior generated by (3):*

- *Suppose (σ_k) is as in (6) for $\alpha > 1/2$ and assume (7)–(8)–(19). For $\beta > 0$, let*

$$\varepsilon_n = (\log n)^{\frac{1+(1+\kappa)\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}}. \tag{20}$$

Then for any $\beta \le \alpha$, $L > 0$ and $f_0 \in \mathcal{S}^\beta(L)$, it holds that for any $d_2 > 0$ there exists $d_1 > 0$ sufficiently large such that

$$\Pi[\|f - f_0\|_2 < d_1\varepsilon_n] \ge e^{-d_2 n\varepsilon_n^2}.$$

- *Suppose (σ_k) is defined, for some $a, \delta > 0$, by*

$$\sigma_k = e^{-a(\log k)^{1+\delta}}, \tag{21}$$

and assume (7)–(8)–(19). For $\beta > 0$, let

$$\varepsilon_n = (\log n)^{\frac{(1+\kappa)(1+\delta)\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}}. \tag{22}$$

Then for any $\beta > 0$, $L > 0$ and $f_0 \in \mathcal{S}^\beta(L)$, it holds that for any $d_2 > 0$ there exists $d_1 > 0$ sufficiently large such that

$$\Pi[\|f - f_0\|_2 < d_1\varepsilon_n] \ge e^{-d_2 n\varepsilon_n^2}.$$

REMARK 3 (Logarithmic factor). For the OT prior with $\kappa = 0$, the logarithmic term is nearly the best possible one (see Remark 1) up to a power $C \cdot \delta$ (for some constant $C = C(\beta)$ depending on β only) that can be made arbitrarily small by taking a small δ.

THEOREM 7 (Generic prior mass condition in L^∞). *For Π a prior generated by* (4):

- *Suppose (s_l) is as in* (6) *for $\alpha > 1/2$ and assume* (7)–(8)–(19). *For $\beta > 0$, let*

$$\varepsilon_n = (\log\log n)^{\frac{2}{1+2\beta}} (\log n)^{\frac{1+(1+\kappa)\beta}{1+2\beta}} n^{-\frac{\beta}{2\beta+1}}. \tag{23}$$

Then for any $\beta \leq \alpha$, $L > 0$ and $f_0 \in \mathcal{H}^\beta(L)$, it holds that for any $d_2 > 0$ there exists $d_1 > 0$ sufficiently large such that, for large enough n,

$$\Pi\big[\|f - f_0\|_\infty < d_1\varepsilon_n\big] \geq e^{-d_2 n\varepsilon_n^2}.$$

- *Suppose (s_l) is as in* (5) *and assume* (7)–(8)–(19). *For $\beta > 0$, let*

$$\varepsilon_n = (\log n)^{\frac{(2+2\kappa)\beta}{1+2\beta}} n^{-\frac{\beta}{2\beta+1}}. \tag{24}$$

Then for any $\beta > 0$, $L > 0$ and $f_0 \in \mathcal{H}^\beta(L)$, it holds that for any $d_2 > 0$ there exists $d_1 > 0$ sufficiently large such that, for large enough n,

$$\Pi\big[\|f - f_0\|_\infty < d_1\varepsilon_n\big] \geq e^{-d_2 n\varepsilon_n^2}.$$

Prior mass results as obtained in Theorem 6–7 are a key preliminary step for obtaining posterior contraction rates. Yet, as noted above, the general theory in [27, 28] also typically requires exponentially decreasing prior masses for certain portions of the parameter space (which can then be 'tested out'). A major difficulty with (nontruncated) heavy-tailed series priors is that prior masses that are exponentially decreasing correspond to extremely small (or 'far-away') sets, so usual approaches via entropy control of sieve sets seem out of reach. While we were able in Section 2 to derive all the results for classical posteriors, here we use instead tempered posteriors: these are defined as, for $0 < \rho \leq 1$, by

$$\Pi_\rho[B|X] = \frac{\int_B (p_f^{(n)}(X))^\rho \, d\Pi(f)}{\int (p_f^{(n)}(X))^\rho \, d\Pi(f)},$$

for measurable B. The usual posterior corresponds to $\rho = 1$ while $\rho < 1$ 'tempers' the influence of the likelihood. Tempered posteriors require only a prior mass condition to converge [59, 62]. Inference in terms of uncertainty quantification can also be conducted with these, see [38]. We refer to the Supplementary Material [3] for more context (Section A therein) and a precise statement (Section D). We also note that the results to follow are obtained for any $\rho < 1$ but not for $\rho = 1$ (except in white noise regression where results hold for both). Deriving results for $\rho = 1$ in general models is beyond the scope of the present work but is an interesting avenue of future research. A detailed discussion on possible approaches to this can be found in the Supplementary Material [3], Section A.

REMARK 4. The condition $\alpha > 1/2$ for the HT(α) prior is a technical condition; it can be checked using similar bounds as in the proof of Theorem 7 that, under that condition, a draw f from the prior (4) is bounded Π-almost surely, which in particular ensures that e^f is integrable, a fact used for inducing a density in (25) below.

3.2. *Density estimation.* From a prior defined by (4) and (5), a prior on densities on [0, 1] is easily defined by exponentiation and renormalisation: for f bounded and measurable, let

$$g(x) = g_f(x) = \frac{e^{f(x)}}{\int_0^1 e^{f(u)}\,du}. \tag{25}$$

THEOREM 8 (Density estimation). *Consider data $X = (X_1, \dots, X_n)$ sampled independently from a density g_0 on $[0, 1]$ that is bounded away from 0 and suppose $f_0 := \log g_0 \in \mathcal{H}^\beta(L)$ for some $\beta > 0$. Let Π be a prior on densities g generated by* (25), *with parameters on the prior on f as in the statement of Theorem* 7 (*e.g.*, $\alpha > 1/2$ *for the HT*(α) *prior*) *with corresponding rates ε_n as in* (23) *or* (24). *For any $\rho < 1$, there exists $M = M(\rho) > 0$ with*

$$E_{g_0}\Pi_\rho\big[\|g - g_0\|_1 > M\varepsilon_n | X\big] \to 0$$

as $n \to \infty$, and where $P_0 = P_{g_0}$.

Theorem 8 derives adaptive posterior contraction for the ρ-posterior at minimax rate (up to a logarithmic term) in density estimation. The term $(1 - \rho)^{-1}$ is expected since it is known that usual posteriors may not converge without entropy and sieve conditions. This result can be seen as a counterpart for heavy-tailed priors to results for usual posteriors in density estimation for Gaussian priors [58], recently obtained also for Laplace priors in [31]. Sampling from tempered posteriors is generally of comparable difficulty compared to classical posteriors, and since there is no hyper-posterior on the smoothness parameter to sample from with the considered heavy-tailed priors, sampling in density estimation is relatively easy and carried out in Section 4- and this even though the model does not tensorise over coordinates.

3.3. *Classification.* Consider independent observations $(X_1, Y_1), \dots, (X_n, Y_n)$ from a given distribution of a random variable (X, Y), where $Y \in \{0, 1\}$ is binary and X takes values in $\mathcal{X} = [0, 1]^d$ for $d \geq 1$. The interest is in estimating the binary regression function $h_0(x) = P(Y = 1|X = x)$. Consider the logistic link function $\Lambda(u) = 1/(1 + e^{-u})$ and denote its inverse by Λ^{-1}. From a function f sampled from (4)–(6) (or (4)–(5)), setting

$$h_f(x) = \Lambda\big(f(x)\big) \tag{26}$$

induces a prior distribution Π on binary regression functions. The density of the data (X, Y) given f equals $p_f(x, y) = h_f(x)^y(1 - h_f(x))^{1-y}g(x)$, where $g(x)$ denotes the marginal density of X. Denote by $\|\cdot\|_{G,1}$ the $L^1(G)$ norm on $\mathcal{X}$ and by P_0 the true data generating distribution with regression f_0 and marginal g (note that Bayesian modelling of g is not needed for inference on f as it factorises from the likelihood).

THEOREM 9 (Classification). *Consider data (X, Y) from the binary classification model. Suppose $f_0 = \Lambda^{-1}h_0$ belongs to the Sobolev ball $\mathcal{S}^\beta(L)$ for $\beta, L > 0$. Let Π be a prior generated by* (26), *with parameters on the prior on f as in the statement of Theorem* 6 *with corresponding rates ε_n as in* (20) *or* (22). *Then for any given $\rho < 1$, there exists $M = M(\rho)$ such that, as $n \to \infty$,*

$$E_{P_0}\Pi_\rho\big[\|p_f - p_{f_0}\|_{G,1} > M\varepsilon_n | X_1, Y_1, \dots, X_n, Y_n\big] \to 0.$$

Theorem 9 derives adaptation for binary classification. Once again, simulation from (an approximation of) the ρ-posterior can be carried out using a direct MCMC method without the need of hyperparameter sampling, see Subsection E.4 in the Supplementary Material [3] for details.

3.4. *Besov classes.* We now provide results for possibly spatially inhomogeneous functions and ρ-posteriors. We restrict for simplicity to white noise regression and to variances as in (5). The following theorem shows that Theorem 4 also holds for ρ-posteriors, $\rho < 1$.

THEOREM 10. *In the white noise regression model, for* Π *a prior generated by* (4) *and* (5), *assume* (7)–(8)–(19) *hold. Suppose* $f_0 \in \mathcal{B}^{\beta}_{rr}(L)$ *for some* $\beta, L > 0$, $r \in [1, 2]$ *and* $\beta > 1/r - 1/2$. *Then, for any given* $\rho < 1$, *as* $n \to \infty$,

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

where $\mathcal{L}_n = (\log n)^d$ *for some* $d > 0$.

We observe excellent empirical behaviour in simulations of the corresponding posterior distributions in terms of adaptation and signal fit on a variety of inhomogeneous test signals, see Section E.2 in the Supplementary Material [3]. A proof of Theorem 10 could be given following similar arguments as for Theorem 4 (i.e., relying on the approach of Theorems 1–3). The proof we provide in the Supplementary Material relies on prior mass arguments. Indeed, the latter are easier to generalise to more other settings (such as density estimation as above) modulo slight adaptation of the conditions. A more systematic study of convergence in Besov spaces in different models and for different losses is beyond the scope of the present work and is the object of forthcoming work.

4. A simulation study.

We consider the following four simulation settings:

(a) inverse regression with Sobolev/spatially homogeneous truth,
(b) spatially inhomogeneous truth in white noise regression,
(c) density estimation with Hölder-regular truth,
(d) binary classification with Sobolev-regular truth.

Here we present the setting (a) and a simulation for (c) for illustration, and refer to the Supplementary Material, Section E, for more details.

Inverse regression. We consider the model studied in [35], Section 3, and [53], Section 4, where one observes the process

$$X_t = \int_0^t \int_0^s f(u)\,du\,ds + \frac{1}{\sqrt{n}}B_t, \quad t \in [0, 1],$$

for B_t a standard Brownian motion and $f \in L^2[0, 1]$ the unknown function. This is a linear inverse problem with the Volterra integral operator $Kf(t) = \int_0^t f(u)\,du$ as forward operator, which has eigenfunctions $e_k(t) = \sqrt{2}\cos(\pi(k - 1/2)t)$ and corresponding eigenvalues $\kappa_k = \pi/(k - 1/2)$, for $k \geq 1$. Equivalently, we study the normal sequence model

$$X_k|f_k \overset{\text{ind}}{\sim} \mathcal{N}(\kappa_k f_k, 1/n), \quad k \geq 1,$$

where f_k are the coefficients of the unknown with respect to the orthonormal system formed by the eigenfunctions (e_k), so that we are in the setting of Section 2.2. As underlying truth we use a function with coefficients with respect to (e_k) given by $f_{0,k} = k^{-3/2}\sin(k)$. In particular, the truth can be thought of as having Sobolev regularity (almost) $\beta = 1$.

We consider priors on the coefficients of the unknown of the form $f_k = \sigma_k\zeta_k$ for i.i.d. ζ_k, for three different choices of the standard deviations σ_k and/or the distribution of ζ_1:

- Gaussian hierarchical prior: $\sigma_k = k^{-1/2-\alpha}$ with $\alpha \sim \text{Exp}(1)$, ζ_1 standard normal;
- HT(α) prior: $\sigma_k = k^{-1/2-\alpha}$ with $\alpha = 5$, ζ_1 Student distribution with 3 degrees of freedom;
- OT prior: $\sigma_k = e^{-a(\log k)^{1+\delta}}$, with $a = 1$, $\delta = 0.5$ and ζ_1 again a Student t_3 distribution.

Note that, on purpose, in order to test the robustness of the method, we take a very 'unfavourable' α for the HT(α) prior, with $\alpha = 5$ much larger than the true smoothness $\beta = 1$ here. Furthermore, for the OT prior we use $\delta = 0.5$ instead of $\delta = 1$ used in our analysis. As noted in Section 1, the contraction rates are identical for any $\delta > 0$, however we found 'the finite' n behaviour to be slightly better for $\delta = 0.5$ compared to $\delta = 1$ (although the difference seemed relatively small in all conducted experiments), so we kept this choice through the simulations.

To sample the posterior arising from the Gaussian hierarchical prior, we employ a Metropolis-within-Gibbs sampler which alternates between updating the $\alpha|f$, X and $f|\alpha$, X (with an appropriate parametrization, centered or noncentered depending on the size of the noise, to optimize the mixing of the α-chain, see [1]). For the two Student priors, due to independence, the posterior decomposes into an infinite product of univariate posteriors. We use Stan, with random initialization uniformly on the interval $(-2, 2)$, to sample each of the univariate posteriors [50] (it is also possible to code this manually, e.g., via a Random Walk Metropolis). In all three 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.

In Figure 2, we present posterior sample 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 OT prior appears to perform at least as well as the Gaussian hierarchical prior at all noise levels both in terms of the posterior sample mean as well as uncertainty quantification (this is expected from the theory, since the OT posterior is guaranteed to converge (near)-optimally for both quadratic and supremum norm). For the HT(α) prior, although α is very far off the true smoothness, we see that as n increases the posterior is still able to approximately match the unknown truth. Compared to the OT prior, in this setting the HT(α) prior appears to be overconfident in all but the lowest noise levels. This is a 'finite n' phenomenon, which can be explained by the behaviour of the univariate Student prior in the model $X \sim \mathcal{N}(f, 1/n)$ as detailed in Section 2 (but adapted to accommodate κ_k): since $\kappa_k \sigma_k = k^{-6.5}$ becomes very small already for small k, among coefficients with small signal-to-noise ratio, very few get a significant value under the posterior and hence there is very little variance in the posterior. Although asymptotically for $k \to \infty$ the scalings of the OT prior decay even faster, $\kappa_k \sigma_k = k^{-1} e^{-(\log k)^{3/2}}$ remains large (relatively to $1/\sqrt{n}$) for more frequencies, hence more frequencies get a significant value under the posterior and the posterior on the function f exhibits more variability.

In Section E of the Supplementary Material [3], we additionally study ρ-posteriors in this setting for the two Student priors, with similar conclusions as for the classical posteriors. We also compare the behaviour of the HT(α) prior and a corresponding Gaussian process prior both with $\alpha = 5$, as an illustration of the 'tail-adaptation' property which takes place for the HT prior but, as expected, not for the Gaussian prior.

Density estimation. We consider the density estimation setting of Section 3.2, for a true density g_0 defined via (25) with f_0 a 2-Hölder smooth function defined via its coefficients in a certain wavelet basis. We consider an α-smooth Gaussian prior, and Cauchy HT(α) and OT priors, where $\alpha = 5$. In Figure 3, we present the corresponding posteriors. Even though we are not in a product space, hence we have to use a function space MCMC algorithm, Cauchy priors show excellent performance without the requirement of a Gibbs Sampler for sampling a smoothness hyper-parameter. For more details, additional experiments and a comparison to a standard frequentist estimator, see Section E in the Supplementary Material [3]. The code for all our experiments with heavy-tailed priors is available at https://www.mas.ucy.ac.cy/sagapi01/assets/code/code-HT-BNP-adapt.zip.

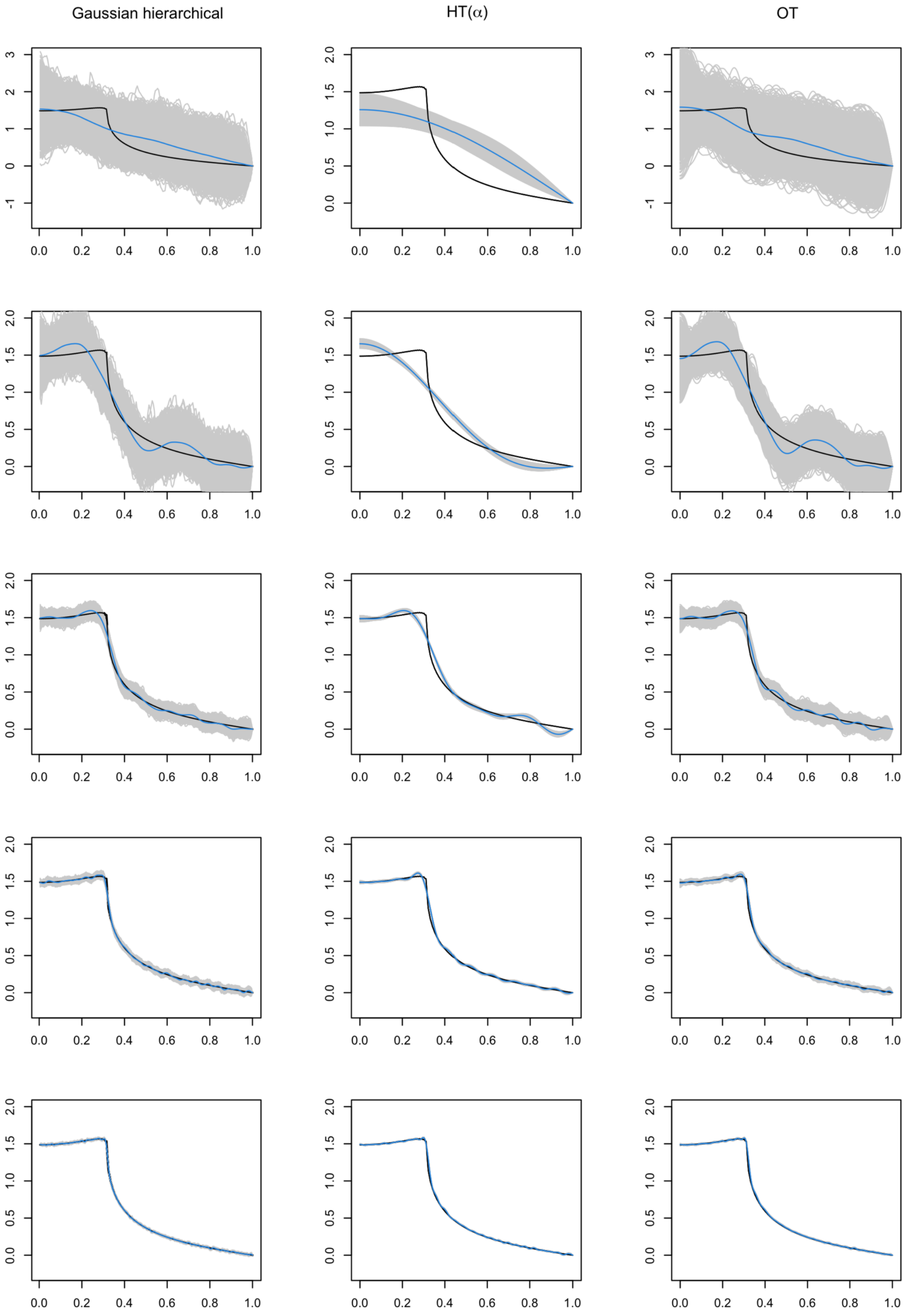

FIG. 2. *White noise model: true function (black), posterior mean (blue), 95% credible regions (grey), for* $n = 10^3, 10^5, 10^7, 10^9, 10^{11}$ *top to bottom and for the three considered priors left to right.*

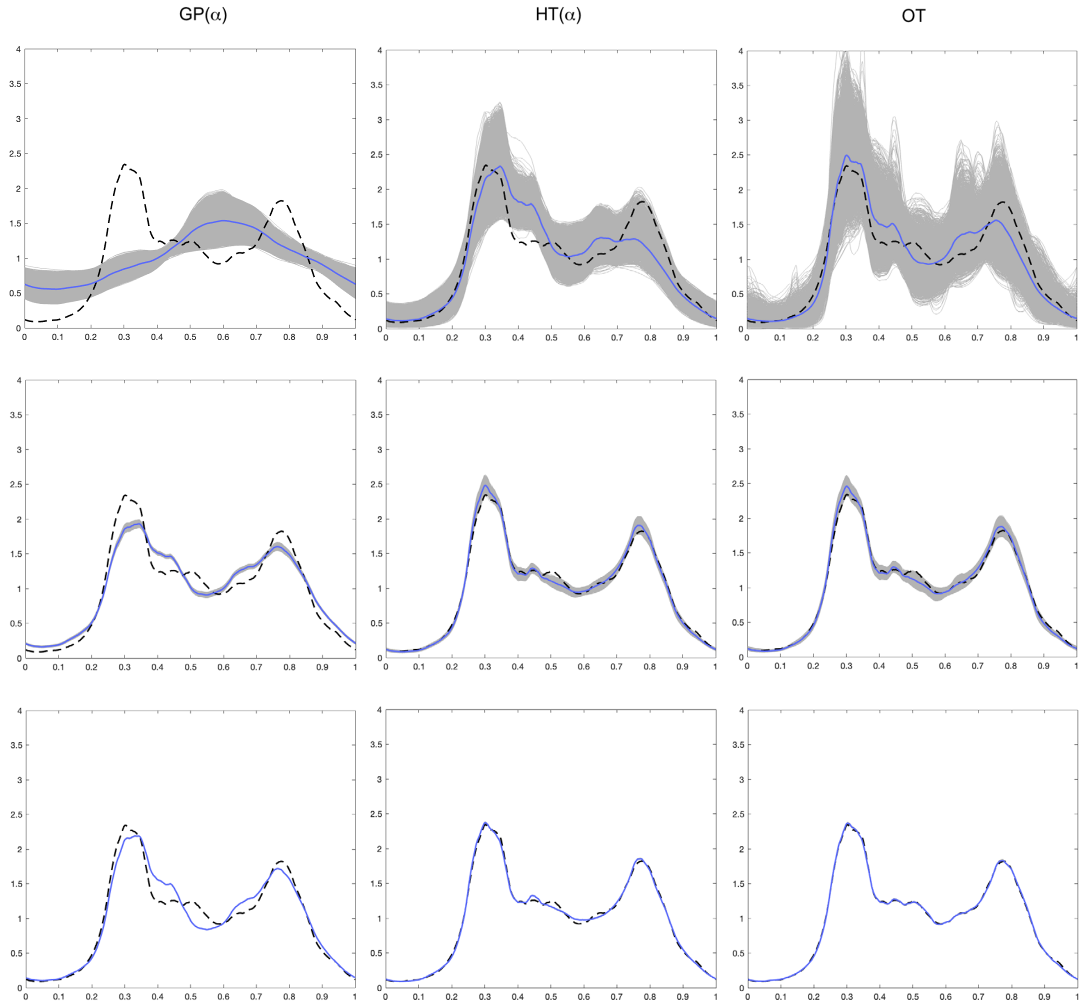

FIG. 3. *Density estimation*: *true density* (*black dashed*), *posterior mean* (*blue*), 95% *credible regions* (*grey*), *for* $n = 10^2, 10^4, 10^6$ *top to bottom and for the three considered priors left to right*.

5. Discussion. We have introduced a new prior, the *oversmoothed heavy-tailed* (OT) prior, which we show leads to Bayesian nonparametric adaptation to smoothness in a wide array of settings. One main appeal is that it is only defined from the basis coefficients one wishes to model, without extra need of hyper-parameters to derive adaptation.

While some prior classes can be seen as Bayesian analogues of nonparametric adaptation methods (e.g., sieve priors↔model selection, spike-and-slab↔thresholding), the OT prior achieves adaptation through the prior distribution's (heavy) *tails*, so in a sense is distinctively Bayesian in spirit.

Open directions. While this work proposes a novel class of adaptive priors and closes a gap in the literature by showing that the phase transition from light (e.g., Gaussian) and moderate (e.g., exponential) tails in infinite series priors to heavy-tails comes with obtaining automatic adaptation, it opens a number of questions for future work. First, the fact that the prior performs a 'soft' model selection—by this we mean that many coefficients are close to zero but not exactly zero under the posterior—is of interest for complex models, where performing a 'hard' model selection (as is the case for spike and slab priors that set coefficients exactly to 0) can be computationally intensive. We are currently working on adaptation on deep ReLU neural networks using such heavy-tailed priors; another interesting direction is that of sparse settings. Among computationally less demanding alternatives to hard model selection, let us also mention the possibility to use a variational Bayes approach, see, for example, [7, 60, 61], although then approximating a different object than the posterior or tempered posterior itself. Second, we conjecture that the results of Section 3 also hold for classical posteriors. This is in particular supported by numerical evidence from Section 4. Although beyond the scope of the present contribution, this suggests that one could possibly extend the general posterior rates theory [27, 28] beyond cases where exponential decrease of sieve set probabilities is available. Third, it would also be particularly interesting to study the computational complexity of heavy-tailed priors in the spirit of the recent works [9, 43].

6. Proof of the main results.

Here we prove Theorems 1 and 6, and related technical lemmas. Proofs of the remaining theorems can be found in the Supplementary Material [3].

PROOF OF THEOREM 1. Let $\varepsilon_n = n^{-\beta/(2\beta+1)}$ and K_n be the (closest integer) solution to

$$K_n = n^{1/(2\beta+1)}. \tag{27}$$

For $\mathcal{A}_n$ a suitable event to be defined below, using Markov's inequality,

$$E_{f_0}\Pi[\|f-f_0\|_2 > v_n | X^{(n)}] \le P_{f_0}(\mathcal{A}_n^c) + v_n^{-2} E_{f_0}\bigg[\int \|f-f_0\|_2^2 \, d\Pi(f|X^{(n)}) 1\!\!1_{\mathcal{A}_n}\bigg],$$

where we choose $v_n = \mathcal{L}_n\varepsilon_n$. 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 $f^{[K_n^c]} := f - f^{[K_n]}$. Then

$$\begin{aligned}\|f-f_0\|_2^2 &= \|f^{[K_n]} - f_0^{[K_n]}\|_2^2 + \|f^{[K_n^c]} - f_0^{[K_n^c]}\|_2^2 \\ &\le \|f^{[K_n]} - f_0^{[K_n]}\|_2^2 + 2\|f^{[K_n^c]}\|_2^2 + 2\|f_0^{[K_n^c]}\|_2^2.\end{aligned}$$

By definition of $\mathcal{S}^\beta(L)$, we have $\|f_0^{[K_n^c]}\|_2^2 \lesssim K_n^{-2\beta} \lesssim \varepsilon_n^2$. Next

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

Under E_{f_0}, the second sum on the right-hand side is bounded by K_n/n. For the first term, Lemma 1 used on coordinate k combined with Lemma 2 gives, for any $t > 0$,

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

This is optimised in t by taking t^4 of the order of the log term in the last display, leading to

$$nE_{f_0}\int (f_k - f_{0,k})^2 \, d\Pi(f|X^{(n)}) \lesssim \log^{1+\kappa}\bigg(1 + \frac{L + 1/\sqrt{n}}{\sigma_k}\bigg). \tag{28}$$

Using $\sigma_k^{-1} \le \sigma_{K_n}^{-1}$, the last term is at most of logarithmic order in n for both choices of σ, so

$$E_{f_0}\int \|f^{[K_n]} - f_0^{[K_n]}\|_2^2 \, d\Pi(f|X^{(n)}) \lesssim (\log n)^d \sum_{k=1}^{K_n}\frac{1}{n} \lesssim (\log n)^d K_n/n \lesssim (\log n)^d n^{-\frac{2\beta}{2\beta+1}}.$$

Let us now turn to bounding the term $\int \|f^{[K_n^c]}\|_2^2 \, d\Pi(f|X)$. We first claim that it is enough to focus on the set of indices k for which $|f_{0,k}| \le 1/\sqrt{n}$. Indeed, if N_n is the cardinality of

$$\mathcal{N}_n := \{k : |f_{0,k}| > 1/\sqrt{n}\}, \tag{29}$$

we have $L^2 \ge \sum_{k\in\mathcal{N}_n} k^{2\beta} f_{0,k}^2 \ge n^{-1}\sum_{k=1}^{N_n} k^{2\beta} \asymp N_n^{2\beta+1}/n$, so that $N_n \lesssim K_n$. Further, if $k \in \mathcal{N}_n$ then $k \le (L^2 n)^{1/(2\beta)}$. This means that for any index $k \in \mathcal{N}_n$, one can use bound (28) (which holds for any k) similarly to the case $k \le K_n$, just using $\sigma_k^{-1} \le \sigma_{(L^2 n)^{1/(2\beta)}}^{-1}$ this time, giving a bound in (28) which is logarithmic in n. In other words, for some $d' > 0$,

$$E_{f_0} \int \sum_{k \in \mathcal{N}_n} (f_k - f_{0,k})^2 \, d\Pi(f \mid X^{(n)}) \lesssim (\log n)^{d'} \sum_{k \in \mathcal{N}_n} \frac{1}{n} \lesssim (\log n)^{d'} \frac{K_n}{n} \lesssim (\log n)^{d'} n^{-\frac{2\beta}{2\beta+1}}.$$

Further note that for both choices of σ's (6) and (5), for any $k > K_n$, one has $\sigma_k \lesssim 1/\sqrt{n}$: this results from the definitions, using $\alpha \ge \beta$ for the choice (6).

We now focus on indices $k \in \{k > K_n : |f_{0,k}| \le 1/\sqrt{n}\}$ and bound $E_{f_0} \int f_k^2 \, d\Pi(f \mid X)$. Using Bayes' formula, it is enough to bound the following terms individually, where ϕ denotes the standard normal density,

$$\int f_k^2 \, d\Pi(f \mid X^{(n)}) = \frac{\int \theta^2 \phi(\sqrt{n}(X_k - \theta)) h(\theta/\sigma_k) \, d\theta}{\int \phi(\sqrt{n}(X_k - \theta)) h(\theta/\sigma_k) \, d\theta} =: \frac{N}{D}.$$

To bound the numerator N, we use $|\phi| \le \|\phi\|_\infty$ and $\int \theta^2 h(\theta/\sigma_k) \, d\theta = \sigma_k^3 \int u^2 h(u) \, du \lesssim \sigma_k^3$, using (14) with $q = 2$, so that $N \lesssim \sigma_k^3$ regardless of X_k.

The denominator is bounded as follows. By symmetry of both ϕ and h, it is enough to focus on the case $X_k \ge 0$ (denoting $D = D(X_k)$, we have $D = D(-X_k)$).

We first deal with the case of super-light variances (5). Let (x_k) be a deterministic nonnegative sequence to be chosen. By restricting the denominator to the set $[X_k - x_k, X_k + x_k]$,

$$D \ge \phi(\sqrt{n} x_k) \int_{X_k - x_k}^{X_k + x_k} h(\theta/\sigma_k) \, d\theta.$$

Assuming $X_k \le x_k$, the integral in the last display can be further bounded from below by $\int_0^{x_k} h(\theta/\sigma_k) \, d\theta = \sigma_k \int_0^{x_k/\sigma_k} h(u) \, du$, recalling $X_k \ge 0$. Further assuming that $\sigma_k \lesssim x_k$, the latter integral is further bounded below, for some $c > 0$, by $\int_0^c h(u) \, du \gtrsim 1$. Putting everything together and using symmetry, one gets, if $\sigma_k \lesssim x_k$,

$$\frac{N}{D} \mathbb{1}_{|X_k| \le x_k} \le \sigma_k^2 \frac{\mathbb{1}_{|X_k| \le x_k}}{\phi(\sqrt{n} x_k)}.$$

Define the events, for $j \ge 0$, $k \ge 1$ and with $a \vee b = \max(a, b)$,

$$\mathcal{A}_{k,j} := \left\{ |X_k| \le \sqrt{\frac{4 \log\{n(j^2 \vee 1)\}}{n}} \right\}, \tag{30}$$

$$\mathcal{A}_n(\mathcal{N}_n) := \bigcap_{K_n < k \le n, k \in \mathcal{N}_n^c} \mathcal{A}_{k,0} \cap \bigcap_{j \ge 1} \bigcap_{jn < k \le (j+1)n, k \in \mathcal{N}_n^c} \mathcal{A}_{k,j}. \tag{31}$$

Let us set $x_k = (4n^{-1} \log\{n(j^2 \vee 1)\})^{1/2}$ whenever $jn < k \le (j+1)n$. The constraint $\sigma_k \lesssim x_k$ is trivially satisfied when $k > K_n$ for this choice of x_k and large enough n since then $\sigma_k \lesssim 1/\sqrt{n}$ as noted above. Then, with $\mathcal{A}_n = \mathcal{A}_n(\mathcal{N}_n)$, using that (σ_k) is decreasing,

$$\begin{aligned}
&E_{f_0} \sum_{k > K_n, k \in \mathcal{N}_n^c} \int f_k^2 \, d\Pi(f \mid X^{(n)}) \mathbb{1}_{\mathcal{A}_n} \\
&\quad \le \sum_{K_n < k \le n, k \in \mathcal{N}_n^c} \frac{\sigma_k^2}{\phi(\sqrt{n} x_0)} + \sum_{j \ge 1} \sum_{jn < k \le (j+1)n, k \in \mathcal{N}_n^c} \frac{\sigma_k^2}{\phi(\sqrt{n} x_k)} \\
&\quad \le n \sigma_{K_n}^2 / \phi(\sqrt{n} x_0) + \sum_{j \ge 1} \sum_{jn < k \le (j+1)n, k \in \mathcal{N}_n^c} \sigma_{jn}^2 / \phi(\sqrt{n} x_{(j+1)n}) \\
&\quad \lesssim n^3 \sigma_{K_n}^2 + \sum_{j \ge 1} n \sigma_{jn}^2 (nj^2)^2 \lesssim n^3 \Big(e^{-2\{\log K_n\}^2} + \sum_{j \ge 1} e^{-2\{\log(nj)\}^2} j^4 \Big).
\end{aligned}$$

The latter bound is $o(n^{-M})$ for arbitrary $M > 0$, which, combined with Lemma 3, concludes the proof for σ_k as in (5).

Let us now turn to the case of variances as in (6). In this case, one slightly updates the definition of $\mathcal{N}_n$ by choosing

$$\mathcal{M}_n = \{k : |f_{0,k}| > \delta_n/\sqrt{n}\}, \tag{32}$$

with $\delta_n := 1/\sqrt{\log n}$. In slight abuse of notation we still denote $\mathcal{A}_n = \mathcal{A}_n(\mathcal{M}_n)$ below. Note that as above one can first deal with indices $k \in \mathcal{M}_n$ using the same bounds as before; reasoning as above, there are at most $m_n \le (n \log n)^{1/(2\beta+1)} \lesssim K_n(\log n)^{1/(2\beta+1)}$ such indices, so their overall contribution to the quadratic risk is within a logarithmic factor of K_n/n.

Thus it is enough to focus on the indices $k > K_n$ and $k \notin \mathcal{M}_n$. If $\sqrt{n}|X_k| \le 1$, then the above bounds for N, D can be used with $1/\sqrt{n}$ in place of x_k, leading to, for $k > K_n$,

$$\frac{N}{D} 1\!\!1_{\sqrt{n}|X_k|\le 1} \lesssim \sigma_k^2 \phi(1)^{-1}.$$

One splits, recalling the definition of $x_k = (4n^{-1}\log\{n(j^2 \vee 1)\})^{1/2}$,

$$\begin{aligned}\frac{N}{D} 1\!\!1_{\sqrt{n}|X_k|>1} 1\!\!1_{|X_k|\le x_k} &= \sum_{p\ge 1} \frac{N}{D} 1\!\!1_{\sqrt{p}<\sqrt{n}|X_k|\le\sqrt{p+1}} 1\!\!1_{|X_k|\le x_k} \\ &\le \sum_{p=1}^{4\log(n(j^2\vee 1))} \frac{N}{D} 1\!\!1_{\sqrt{p}<\sqrt{n}|X_k|\le\sqrt{p+1}}.\end{aligned}$$

We bound from below D, on the event $\{\sqrt{p} < \sqrt{n}|X_k| \le \sqrt{p+1}\}$, as follows, for $X_k \ge 0$,

$$D \ge \phi(\sqrt{n}X_k) \int_0^{X_k} h(\theta/\sigma_k)\, d\theta \ge \sigma_k \phi(\sqrt{p+1}) \int_0^{X_k/\sigma_k} h(u)\, du \gtrsim \sigma_k \phi(\sqrt{p+1}),$$

by restricting the denominator to $[0, X_k]$ and using $X_k/\sigma_k \ge 1/(\sqrt{n}\sigma_k) \gtrsim 1$ for $k > K_n$. By symmetry, the same bound also holds in case $X_k \le 0$. So, for given k,

$$\sum_{p=1}^{4\log(n(j^2\vee 1))} \frac{N}{D} 1\!\!1_{\sqrt{p}<\sqrt{n}|X_k|\le\sqrt{p+1}} \lesssim \sigma_k^2 \sum_{p=1}^{4\log(n(j^2\vee 1))} \frac{1\!\!1_{\sqrt{p}<\sqrt{n}|X_k|\le\sqrt{p+1}}}{\phi(\sqrt{p+1})},$$

where we use the universal bound $N \lesssim \sigma_k^3$ obtained above. Then

$$\begin{aligned}&E_{f_0} \sum_{k>K_n, k\in\mathcal{M}_n^c} \int f_k^2\, d\Pi(f|X) 1\!\!1_{\mathcal{A}_n} \\ &\quad\lesssim E_{f_0} \sum_{K_n<k\le n, k\in\mathcal{M}_n^c} \sigma_k^2 \left\{1 + \sum_{p=1}^{4\log n} \frac{1\!\!1_{\sqrt{p}<\sqrt{n}|X_k|\le\sqrt{p+1}}}{\phi(\sqrt{p+1})}\right\} \\ &\qquad + E_{f_0} \sum_{j\ge 1} \sum_{jn<k\le(j+1)n, k\in\mathcal{M}_n^c} \sigma_k^2 \left\{1 + \sum_{p=1}^{4\log(nj^2)} \frac{1\!\!1_{\sqrt{p}<\sqrt{n}|X_k|\le\sqrt{p+1}}}{\phi(\sqrt{p+1})}\right\}.\end{aligned}$$

We have, for $k \in \mathcal{M}_n^c$, denoting $\bar\Phi(u) = \int_u^{+\infty} \phi(t)\, dt$,

$$\begin{aligned}E_{f_0} 1\!\!1_{\sqrt{p}<\sqrt{n}|X_k|\le\sqrt{p+1}} &= P(|\mathcal{N}(0,1) + f_{0,k}\sqrt{n}| \in [\sqrt{p}, \sqrt{p+1}]) \\ &\le 2\{\bar\Phi(\sqrt{p} - \delta_n) - \bar\Phi(\sqrt{p+1} + \delta_n)\} \le 2\phi(\sqrt{p} - \delta_n)/(\sqrt{p} - \delta_n),\end{aligned}$$

by removing the negative term and using the standard bound $\bar{\Phi}(x) \le \phi(x)/x$ for $x > 0$. Now

$$\frac{1}{\sqrt{p}-\delta_n}\frac{\phi(\sqrt{p}-\delta_n)}{\phi(\sqrt{p+1})} = \frac{e^{1/2}}{\sqrt{p}-\delta_n}e^{-\delta_n^2/2+\sqrt{p}\delta_n}.$$

First dealing with the term $k \le n$, one deduces, recalling $\delta_n = (\log n)^{-1/2}$,

$$\begin{aligned}
&\sum_{K_n<k\le n, k\in\mathcal{M}_n^c} \sigma_k^2 \sum_{p=1}^{4\log n} \frac{P_{f_0}(\sqrt{p} < \sqrt{n}|X_k| \le \sqrt{p+1})}{\phi(\sqrt{p+1})} \\
&\quad\lesssim \sum_{K_n<k\le n, k\in\mathcal{M}_n^c} \sigma_k^2 \sum_{p=1}^{4\log n} \frac{1}{\sqrt{p}} e^{\sqrt{p}\delta_n} \\
&\quad\lesssim \sum_{K_n<k\le n, k\in\mathcal{M}_n^c} \sigma_k^2 \sum_{p=1}^{4\log n} \frac{1}{\sqrt{p}} \lesssim \sum_{K_n<k\le n, k\in\mathcal{M}_n^c} \sigma_k^2\sqrt{\log n},
\end{aligned}$$

where one uses the previous bounds. Similarly,

$$\begin{aligned}
&\sum_{j\ge1}\sum_{jn<k\le(j+1)n, k\in\mathcal{M}_n^c} \sigma_k^2 \sum_{p=1}^{4\log(nj^2)} \frac{P_{f_0}(\sqrt{p} < \sqrt{n}|X_k| \le \sqrt{p+1})}{\phi(\sqrt{p+1})} \\
&\quad\le \sum_{j\ge1}\sum_{jn<k\le(j+1)n} \sigma_k^2 \sum_{p=1}^{4\log(nj^2)} \frac{1}{\sqrt{p}} e^{\sqrt{p}\delta_n} \lesssim \sum_{j\ge1}\sum_{jn<k\le(j+1)n} \sigma_k^2 \sqrt{\log(nj^2)}e^{2\sqrt{\log(nj^2)}\delta_n}.
\end{aligned}$$

Using that $nj^2 \le (nj)^2 \le k^2$ for $k > (nj)$ one gets that the last display is bounded up to a constant multiplicative factor by

$$\sum_{j\ge1}\sum_{jn<k\le(j+1)n} \sigma_k^2\sqrt{\log k}e^{4\sqrt{\log k}\delta_n} = \sum_{k>n}\sigma_k^2\sqrt{\log k}e^{4\sqrt{\log k}\delta_n}.$$

Since $\sqrt{\log k}e^{4\sqrt{\log k}\delta_n} \le e^{\eta\log k}$ for any $k \ge n$ for $\eta > 0$ fixed as small as desired for large enough n, the last display is bounded by $\sum_{k\ge n}\sigma_k^2 k^\eta \lesssim n^{-2\alpha+\eta} = o(n^{-2\alpha/(2\alpha+1)})$ for small enough η. Putting the previous bounds together one gets

$$E_{f_0}\sum_{k>K_n, k\in\mathcal{M}_n^c}\int f_k^2\, d\Pi(f|X^{(n)})\mathbb{1}_{\mathcal{A}_n} \lesssim \sum_{k=K_n}^{n}\sigma_k^2(1+\sqrt{\log n}) + o(n^{-\frac{2\alpha}{2\alpha+1}}) \lesssim \frac{\sqrt{\log n}}{n^{2\alpha/(2\alpha+1)}},$$

using (6). This bound is $O((\sqrt{\log n})n^{-2\beta/(2\beta+1)})$ if $\alpha \ge \beta$, which, combined with Lemma 3, concludes the proof for σ_k as in (6). □

PROOF OF THEOREM 6. Let $K \ge 2$ be an integer, and for a function f in L^2, let as before $f^{[K]}$ denote its projection onto the linear span of $\varphi_1, \dots, \varphi_K$ and $f^{[K^c]} = f - f^{[K]}$. Then

$$\begin{aligned}
&\Pi[\|f - f_0\|_2 < \varepsilon] \\
&\quad\ge \Pi[\|f^{[K]} - f_0^{[K]}\|_2 < \varepsilon/2, \|f^{[K^c]} - f_0^{[K^c]}\|_2 < \varepsilon/2] \\
&\quad\ge \Pi\left[\forall k \le K, |f_k - f_{0,k}| \le \frac{\varepsilon}{2\sqrt{K}}; \forall k > K, |f_k| \le \frac{\varepsilon}{D\sqrt{k}\log k}\right]\mathbb{1}_{\|f_0^{[K^c]}\|_2<\varepsilon/4} \\
&\quad= \prod_{k=1}^{K}\Pi\left[|f_k - f_{0,k}| \le \frac{\varepsilon}{2\sqrt{K}}\right]\Pi\left[\forall k > K, |f_k| \le \frac{\varepsilon}{D\sqrt{k}\log k}\right]\mathbb{1}_{\|f_0^{[K^c]}\|_2<\varepsilon/4},
\end{aligned}$$

where D is a large enough constant, and where we have used independence of the coefficients under the prior and the fact that $k^{-1/2}/\log(k)$ is a square-summable sequence.

Suppose the indicator in the last display equals one, which imposes $\|f_0^{[K^c]}\|_2 < \varepsilon/4$, for which a sufficient condition is, if f_0 is in $\mathcal{S}^\beta(L)$,

$$K^{-2\beta}L^2 < \varepsilon^2/16. \tag{33}$$

Let us now bound each individual term $p_k := \Pi[|f_k - f_{0,k}| \le \varepsilon/(2\sqrt{K})]$. By symmetry, for any $k \le K$, one can assume $f_{0,k} \ge 0$ and

$$p_k \ge \int_{f_{0,k}}^{f_{0,k}+\varepsilon/(2\sqrt{K})} \sigma_k^{-1} h(x/\sigma_k)\,dx \ge \frac{\varepsilon}{2\sqrt{K}} h(C/\sigma_K),$$

where we have used that (σ_k) is decreasing as well as $x \to h(x)$ on $[0,\infty)$ by assumption, and that $f_{0,k} + \varepsilon/(2\sqrt{K}) \le C$ since $|f_{0,k}|$ are bounded by L for $f_0 \in \mathcal{S}^\beta(L)$. For either choice of σ_k it holds $\log(2\sqrt{K}) \le \log(1 + C/\sigma_K)$ for large K, hence combining with (8) we get

$$p_k \ge \varepsilon e^{-C_1 \log^{1+\kappa}(1+C/\sigma_K)}.$$

One deduces, for a new value of the constant C_1,

$$\mathcal{P}_1 := \prod_{k=1}^K p_k \ge \varepsilon^K \exp\{-C_1 K \log^{1+\kappa}(C/\sigma_K)\}.$$

Let us deal first with the case of σ_k given by (6). We now also need to bound

$$\begin{aligned}\mathcal{P}_2 &:= \Pi\left[\forall k > K, |f_k| \le \frac{\varepsilon}{D\sqrt{k}\log k}\right] = \prod_{k>K} \left(1 - 2\overline{H}\left(\varepsilon/\{D\sigma_k\sqrt{k}\log k\}\right)\right) \\ &= \prod_{k>K} \left(1 - 2\overline{H}\left(\varepsilon k^\alpha/\{D\log k\}\right)\right).\end{aligned}$$

Note that if

$$\varepsilon \ge D' K^{-\beta} \log K, \tag{34}$$

for some universal constant $D' > 0$ to be chosen. Then, for $k > K$,

$$\varepsilon k^\alpha/\{D \log k\} \ge k^\alpha K^{-\beta} \frac{\log K}{\log k} \frac{D'}{D}.$$

Hence, as long as $\alpha \ge \beta$, D' can be chosen sufficiently large so that the last term is at least 1. Then, by using (19),

$$\overline{H}\left(\varepsilon k^\alpha/\{D \log k\}\right) \le C_3 K^{2\beta} k^{-2\alpha} (\log k/\log K)^2.$$

Then, if $\alpha > 1/2$ so that the series $\sum k^{-2\alpha}$ is converging, and possibly enlarging D' in (34) further in order to have that the right-hand side of the last display is less than 1/4, using the inequality $\log(1 - 2x) \ge -4x$ for all $x \in [0, 1/4)$, we have

$$\begin{aligned}\mathcal{P}_2 &\ge \exp\left\{\sum_{k>K} \log\left(1 - 2C_3 K^{2\beta} k^{-2\alpha} (\log k/\log K)^2\right)\right\} \\ &\ge \exp\{-4C_3 K^{2\beta} \sum_{k>K} k^{-2\alpha} (\log k/\log K)^2) \ge \exp\left(-C_4 K \cdot K^{2(\beta-\alpha)}\right),\end{aligned}$$

where we have used, whenever $\alpha > 1/2$, that $\sum_{k>K} k^{-2\alpha}(\log k)^2 = O(K^{-2\alpha+1}\log^2 K)$ as $K \to \infty$. The bound of the last display is at least $\exp(-C_4 K)$ assuming $\alpha \ge \beta$.

Putting everything together one gets

$$\begin{aligned}\Pi\big[\|f-f_0\|_2<\varepsilon\big] &\geq \varepsilon^K \exp\big\{-C_1 K \log^{1+\kappa}(C/\sigma_K) - C_4 K\big\}\\ &\geq \exp\big\{-K\log\big(K^\beta/(D'\log K)\big) - C_1 K\log^{1+\kappa}(C/\sigma_K) - C_4 K\big\}\\ &\geq \exp\big\{-C_5 K\log^{1+\kappa} K\big\}.\end{aligned}$$

Recall the definition of ε_n as in (20) and the constants d_1, d_2 from the statement of the Theorem. Set $K = K_n = (d_1/D')^{-1/\beta} n^{1/(1+2\beta)} \log^q n$, with $q = (1-\kappa)/(1+2\beta)$ and d_1 to be chosen below. Then (34) holds for $\varepsilon = d_1\varepsilon_n$ and large enough n. Also,

$$C_5 K\log^{1+\kappa} K \leq C d_1^{-1/\beta} n\varepsilon_n^2,$$

where C is a constant independent of d_1, d_2. For given $d_2 > 0$, the right-hand side of the last display is bounded from above by $d_2 n\varepsilon_n^2$, provided d_1 is chosen sufficiently large, which concludes the proof for σ_k as in (6).

Let us now turn to the case of σ_k given by (21). The term $\mathcal{P}_1$ above is bounded below by

$$\prod_{k=1}^K p_k \geq \varepsilon^K \exp\big\{-C_1 K\log^{1+\kappa}(C/\sigma_K)\big\} \geq \varepsilon^K \exp\big\{-C_1 K \log^{(1+\kappa)(1+\delta)} K\big\}.$$

On the other hand, we also have

$$\begin{aligned}\mathcal{P}_2 &= \prod_{k>K}\big(1-2\overline{H}\big(\varepsilon/\{D\sigma_k\sqrt{k}\log k\}\big)\big)\\ &= \prod_{k>K}\big(1-2\overline{H}\big(\varepsilon e^{a\log^{1+\delta}k}/\{D\sqrt{k}\log k\}\big)\big).\end{aligned}$$

Let $\varepsilon \geq DK^{-\beta}$, then for large enough K, $\varepsilon e^{a\log^{1+\delta}k}/\{D\sqrt{k}\log k\} \geq 1$ and by (19)

$$\overline{H}\big(\varepsilon e^{a\log^{1+\delta}k}/\{D\sqrt{k}\log k\}\big) \leq c_2\big(\varepsilon e^{a\log^{1+\delta}k}/\{D\sqrt{k}\log k\}\big)^{-2} \leq e^{-a\log^{1+\delta}k}$$

so that $\mathcal{P}_2 \geq \exp\{-C\sum_{k>K} e^{-a\log^{1+\delta}k}\} \geq \exp\{-C'e^{-a\log^{1+\delta}K}\}$. The latter is bounded from below by a constant, so the final bound obtained for the probability at stake is $\exp\{-C'K\log^{(1+\kappa)(1+\delta)}K\}$ (for a new value of the constant C'). Let us set $\varepsilon = d_1\varepsilon_n$ and $K := (d_1\varepsilon_n/D)^{-1/\beta}$, for ε_n as in (22). Then

$$C'K\log^{(1+\kappa)(1+\delta)}K \leq C d_1^{-1/\beta} n\varepsilon_n^2,$$

where C is independent of d_1, d_2. The last display is less than $d_2 n\varepsilon_n^2$ for large d_1, which concludes the proof. □

REMARK 5 (Cauchy tails). We note that the case of H equal to the Cauchy distribution, corresponding to $\overline{H}(x) \leq c_2/x$ for $x \geq 1$, can be accommodated up to a slight variation on the condition for the prior HT(α). Suppose in this case that $\alpha > 1$ (recall that we have the choice of α, and that in view of the theorem, the larger α is, the larger the range for which adaptation occurs, so we can always choose $\alpha > 1$ beforehand). Indeed, in this case for $\alpha > 1$ one gets, for σ_k as in (6),

$$\begin{aligned}\mathcal{P}_2 &\geq \exp\Big\{\sum_{k>K}\log\big(1-2C_3\big(K^\beta/k^\alpha\big)(\log k/\log K)\big)\Big\}\\ &\geq \exp\big\{-C_4K^{1+\beta-\alpha}\big\} = \exp\big\{-C_4 K K^{\beta-\alpha}\big\},\end{aligned}$$

and from there on the proof is identical to that of Theorem 6. A similar remark applies to the case of σ_k as in (5), this time with no extra condition (the latter choice is free of α).

Technical lemmas.

LEMMA 1. *Consider the model $\theta \sim \pi$ and $X|\theta \sim \mathcal{N}(\theta, 1/n)$. Suppose π is the law of $\sigma \cdot \zeta$, for $\sigma > 0$ and ζ a real random variable with density h satisfying* (7)–(8). *Then for some $C, C_1 > 0$, it holds, for all $t \in \mathbb{R}$, $\theta_0 \in \mathbb{R}$, $\sigma > 0$,*

$$\log E_{\theta_0} E^{\pi}\big[e^{t\sqrt{n}(\theta - X)}|X\big] \le t^2/2 + C \log^{1+\kappa}\left(1 + \frac{|\theta_0| + 1/\sqrt{n}}{\sigma}\right) + C_1.$$

PROOF. For any $t \in \mathbb{R}$, we have

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

Using that h is bounded, one gets

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

The latter integral can be lower bounded using (7) and (8),

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

As a result, we have

$$E_{\theta_0} E^{\pi}\big[e^{t\sqrt{n}(\theta-X)}|X\big] \lesssim e^{t^2/2 + c_1 \log^{1+\kappa}(1 + \frac{|\theta_0| + 1/\sqrt{n}}{\sigma})} E_{\xi \sim \mathcal{N}(0,1)}\left[\left(\int_{-1}^{1} e^{-\frac{(v-\xi)^2}{2}}\,dv\right)^{-1}\right],$$

and the claim follows since the expectation appearing on the right-hand side can be bounded by a universal constant as in [19], pages 2015–2016. □

LEMMA 2. *Let Y be a real random variable. Then for $t > 0$ and $\mathcal{L}(t) = E[\exp(t|Y|)]$,*

$$E\big[Y^2\big] \le t^{-2}\big\{8 + 2\log^2 \mathcal{L}(t)\big\}.$$

PROOF. Let us write $E[Y^2] = t^{-2}E[\log^2 \exp(t|Y|)]$. Using concavity of the map $x \to \log^2(x)$ for $x > e$, one may write

$$\begin{aligned} E\big[\log^2 \exp(t|Y|)\big] &\le E\big[\log^2\big(e + \exp(t|Y|)\big)\big] \le \log^2\big\{e + \mathcal{L}(t)\big\} \\ &\le \big(2 + \log \mathcal{L}(t)\big)^2 \le 8 + 2\log^2 \mathcal{L}(t), \end{aligned}$$

where the last inequality uses $\log(e + b) \le 2 + \log(b)$ valid for $b \ge 1$. □

LEMMA 3. *Let $\mathcal{A}_n(N)$ be the event as in* (31), *where either $N = \mathcal{N}_n$ or $N = \mathcal{M}_n$ as in* (29) *and* (32). *Then for large enough n it holds*

$$P_{f_0}\big[\mathcal{A}_n^c\big] \lesssim 1/\sqrt{n}.$$

PROOF. First, for any $k \in N^c$ and $j \geq 0$, we have

$$P_{f_0}[\mathcal{A}_{k,j}^c] \leq P\big[|\mathcal{N}(0,1)| > \sqrt{3\log n(j^2 \vee 1)}\big] \leq 2\big(n(j^2 \vee 1)\big)^{-3/2},$$

where one uses that $|f_{0,k}| \leq 1/\sqrt{n}$ for $k \in N^c$. From this, one deduces

$$P_{f_0}[\mathcal{A}_n^c] \lesssim \sum_{j \geq 0} n \frac{1}{\{n(j^2 \vee 1)\}^{3/2}} \lesssim 1/\sqrt{n}.$$

□

Acknowledgments. The authors would like to thank the Associate Editor and two referees for a number of insightful comments.

Funding. IC's work was partly funded by ANR BACKUP grant ANR-23-CE40-0018-01 and a fellowship from the Institut Universitaire de France.

SUPPLEMENTARY MATERIAL

Supplementary Material to 'Heavy-tailed Bayesian nonparametric adaptation' (DOI: 10.1214/24-AOS2397SUPP; .pdf). We provide the rest of the proofs of the results contained in the main article, some technical lemmas as well as additional simulations corroborating the theory. We also provide a discussion on the challenges involved in extending some of our results on contraction of ρ-posteriors to standard posteriors.

REFERENCES

[1] Agapiou, S., Bardsley, J. M., Papaspiliopoulos, O. and Stuart, A. M. (2014). Analysis of the Gibbs sampler for hierarchical inverse problems. *SIAM/ASA J. Uncertain. Quantificat.* **2** 511–544. MR3283919 https://doi.org/10.1137/130944229

[2] Agapiou, S., Burger, M., Dashti, M. and Helin, T. (2018). Sparsity-promoting and edge-preserving maximum *a posteriori* estimators in non-parametric Bayesian inverse problems. *Inverse Probl.* **34** 045002, 37. MR3774703 https://doi.org/10.1088/1361-6420/aaacac

[3] Agapiou, S. and Castillo, I. (2024). Supplement to "Heavy-tailed Bayesian nonparametric adaptation." https://doi.org/10.1214/24-AOS2397SUPP

[4] Agapiou, S., Dashti, M. and Helin, T. (2021). Rates of contraction of posterior distributions based on p-exponential priors. *Bernoulli* **27** 1616–1642. MR4278794 https://doi.org/10.3150/20-bej1285

[5] 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 https://doi.org/10.3150/23-bej1673

[6] Agapiou, S. and Wang, S. (2024). Laplace priors and spatial inhomogeneity in Bayesian inverse problems. *Bernoulli* **30** 878–910. MR4699538 https://doi.org/10.3150/22-bej1563

[7] Alquier, P. and Ridgway, J. (2020). Concentration of tempered posteriors and of their variational approximations. *Ann. Statist.* **48** 1475–1497. MR4124331 https://doi.org/10.1214/19-AOS1855

[8] Arbel, J., Gayraud, G. and Rousseau, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scand. J. Stat.* **40** 549–570. MR3091697 https://doi.org/10.1002/sjos.12002

[9] Bandeira, A. S., Maillard, A., Nickl, R. and Wang, S. (2023). On free energy barriers in Gaussian priors and failure of cold start MCMC for high-dimensional unimodal distributions. *Philos. Trans. R. Soc. A* **381** Paper No. 20220150, 29. MR4590496 https://doi.org/10.1098/rsta.2022.0150

[10] Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028 https://doi.org/10.1007/s004400050210

[11] Belitser, E. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.* **31** 536–559. Dedicated to the memory of Herbert E. Robbins. MR1983541 https://doi.org/10.1214/aos/1051027880

[12] Bui-Thanh, T. and Ghattas, O. (2015). A scalable algorithm for MAP estimators in Bayesian inverse problems with Besov priors. *Inverse Probl. Imaging* **9** 27–53. MR3305885 https://doi.org/10.3934/ipi.2015.9.27

[13] Cai, T. T. (2008). On information pooling, adaptability and superefficiency in nonparametric function estimation. *J. Multivariate Anal.* **99** 421–436. MR2396972 https://doi.org/10.1016/j.jmva.2006.11.010

[14] CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751 https://doi.org/10.1093/biomet/asq017

[15] CASTILLO, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* **2** 1281–1299. MR2471287 https://doi.org/10.1214/08-EJS273

[16] CASTILLO, I. (2014). On Bayesian supremum norm contraction rates. *Ann. Statist.* **42** 2058–2091. MR3262477 https://doi.org/10.1214/14-AOS1253

[17] CASTILLO, I., KERKYACHARIAN, G. and PICARD, D. (2014). Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields* **158** 665–710. MR3176362 https://doi.org/10.1007/s00440-013-0493-0

[18] 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 https://doi.org/10.1214/20-aihp1132

[19] CASTILLO, I. and NICKL, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** 1999–2028. MR3127856 https://doi.org/10.1214/13-AOS1133

[20] CASTILLO, I. and NICKL, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** 1941–1969. MR3262473 https://doi.org/10.1214/14-AOS1246

[21] CASTILLO, I. and ROČKOVÁ, V. (2021). Uncertainty quantification for Bayesian CART. *Ann. Statist.* **49** 3482–3509. MR4352538 https://doi.org/10.1214/21-aos2093

[22] CAVALIER, L. (2011). Inverse problems in statistics. In *Inverse Problems and High-Dimensional Estimation. Lect. Notes Stat. Proc.* **203** 3–96. Springer, Heidelberg. MR2868199 https://doi.org/10.1007/978-3-642-19989-9_1

[23] DASHTI, M., HARRIS, S. and STUART, A. (2012). Besov priors for Bayesian inverse problems. *Inverse Probl. Imaging* **6** 183–200. MR2942737 https://doi.org/10.3934/ipi.2012.6.183

[24] DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921. MR1635414 https://doi.org/10.1214/aos/1024691081

[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] GAO, C. and ZHOU, H. H. (2016). Rate exact Bayesian adaptation with modified block priors. *Ann. Statist.* **44** 318–345. MR3449770 https://doi.org/10.1214/15-AOS1368

[27] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

[28] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference. Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834

[29] 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 https://doi.org/10.1214/11-AOS924

[30] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge Univ. Press, New York. MR3588285 https://doi.org/10.1017/CBO9781107337862

[31] GIORDANO, M. (2023). Besov–Laplace priors in density estimation: Optimal posterior contraction rates and adaptation. *Electron. J. Stat.* **17** 2210–2249. MR4649387 https://doi.org/10.1214/23-ejs2161

[32] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **43** 2259–2295. MR3396985 https://doi.org/10.1214/15-AOS1341

[33] IBRAGIMOV, I. A. and HAS'MINSKIĬ, R. Z. (1980). An estimate of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.* (*LOMI*) **98** 61–85, 161–162, 166. Studies in mathematical statistics, IV. MR0591862

[34] JOHNSTONE, I. M. (2019). Gaussian estimation: Sequence and wavelet models. Unpublished manuscript.

[35] KNAPIK, B. T., SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2016). Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Related Fields* **164** 771–813. MR3477780 https://doi.org/10.1007/s00440-015-0619-7

[36] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39** 2626–2657. MR2906881 https://doi.org/10.1214/11-AOS920

[37] KOLEHMAINEN, V., LASSAS, M., NIINIMÄKI, K. and SILTANEN, S. (2012). Sparsity-promoting Bayesian inversion. *Inverse Probl.* **28** 025005, 28. MR2876856 https://doi.org/10.1088/0266-5611/28/2/025005

[38] L'HUILLIER, A., TRAVIS, L., CASTILLO, I. and RAY, K. (2023). Semiparametric inference using fractional posteriors. *J. Mach. Learn. Res.* **24** Paper No. [389], 61. MR4720845 https://doi.org/10.4995/agt.2023.18504

[39] LASSAS, M., SAKSMAN, E. and SILTANEN, S. (2009). Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Probl. Imaging* **3** 87–122. MR2558305 https://doi.org/10.3934/ipi.2009.3.87

[40] LEPSKIĬ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatn. Primen.* **35** 459–470. MR1091202 https://doi.org/10.1137/1135065

[41] LEPSKIĬ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatn. Primen.* **36** 645–659. MR1147167 https://doi.org/10.1137/1136085

[42] NAULET, Z. (2022). Adaptive Bayesian density estimation in sup-norm. *Bernoulli* **28** 1284–1308. MR4388939 https://doi.org/10.3150/21-bej1387

[43] NICKL, R.and WANG, S. (2024). On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms *J. Eur. Math. Soc.* (*JEMS*) **26** 1031–1112. MR4721029 https://doi.org/10.4171/jems/1304

[44] RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

[45] RAY, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat.* **7** 2516–2549. MR3117105 https://doi.org/10.1214/13-EJS851

[46] RAY, K. (2017). Adaptive Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **45** 2511–2536. MR3737900 https://doi.org/10.1214/16-AOS1533

[47] SCRICCIOLO, C. (2006). Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *Ann. Statist.* **34** 2897–2920. MR2329472 https://doi.org/10.1214/009053606000000911

[48] SHAH, A., WILSON, A. and GHAHRAMANI, Z. (2014). Student-t processes as alternatives to Gaussian processes. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* (S. Kaski and J. Corander, eds.). *Proceedings of Machine Learning Research* **33** 877–885. PMLR, Reykjavik , Iceland.

[49] SHEN, W. and GHOSAL, S. (2015). Adaptive Bayesian procedures using random series priors. *Scand. J. Stat.* **42** 1194–1213. MR3426318 https://doi.org/10.1111/sjos.12159

[50] STAN DEVELOPMENT TEAM (2024). Stan Modelling Language Users Guide and Reference Manual v. 2.34.

[51] SULLIVAN, T. J. (2017). Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors. *Inverse Probl. Imaging* **11** 857–874. MR3681971 https://doi.org/10.3934/ipi.2017040

[52] SUURONEN, J., CHADA, N. K. and ROININEN, L. (2022). Cauchy Markov random field priors for Bayesian inversion. *Stat. Comput.* **32** Paper No. 33, 26. MR4402180 https://doi.org/10.1007/s11222-022-10089-z

[53] SZABÓ, B., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** 1391–1428. MR3357861 https://doi.org/10.1214/14-AOS1270

[54] 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 https://doi.org/10.1214/13-EJS798

[55] TRIEBEL, H. (1983). *Theory of Function Spaces*. *Monographs in Mathematics* **78**. Birkhäuser, Basel. MR0781540 https://doi.org/10.1007/978-3-0346-0416-1

[56] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats. MR2724359 https://doi.org/10.1007/b13794

[57] 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 https://doi.org/10.1214/009053607000000613

[58] 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 https://doi.org/10.1214/08-AOS678

[59] WALKER, S. and HJORT, N. L. (2001). On Bayesian consistency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 811–821. MR1872068 https://doi.org/10.1111/1467-9868.00314

[60] YANG, Y., PATI, D. and BHATTACHARYA, A. (2020). α-variational inference with statistical guarantees. *Ann. Statist.* **48** 886–905. MR4102680 https://doi.org/10.1214/19-AOS1827

[61] ZHANG, F. and GAO, C. (2020). Convergence rates of variational posterior distributions. *Ann. Statist.* **48** 2180–2207. MR4134791 https://doi.org/10.1214/19-AOS1883

[62] ZHANG, T. (2006). From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *Ann. Statist.* **34** 2180–2210. MR2291497 https://doi.org/10.1214/009053606000000704

[63] ZHAO, L. H. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.* **28** 532–552. MR1790008 https://doi.org/10.1214/aos/1016218229