The Annals of Statistics
2015, Vol. 43, No. 5, 2259–2295
DOI: 10.1214/15-AOS1341

ON ADAPTIVE POSTERIOR CONCENTRATION RATES

BY MARC HOFFMANN[*,1], JUDITH ROUSSEAU[*,1]
AND JOHANNES SCHMIDT-HIEBER[†,2]

Université Paris-Dauphine[*] and University of Leiden [†]

We investigate the problem of deriving posterior concentration rates under different loss functions in nonparametric Bayes. We first provide a lower bound on posterior coverages of shrinking neighbourhoods that relates the metric or loss under which the shrinking neighbourhood is considered, and an intrinsic pre-metric linked to frequentist separation rates. In the Gaussian white noise model, we construct feasible priors based on a spike and slab procedure reminiscent of wavelet thresholding that achieve adaptive rates of contraction under L^2 or L^∞ metrics when the underlying parameter belongs to a collection of Hölder balls and that moreover achieve our lower bound. We analyse the consequences in terms of asymptotic behaviour of posterior credible balls as well as frequentist minimax adaptive estimation. Our results are appended with an upper bound for the contraction rate under an arbitrary loss in a generic regular experiment. The upper bound is attained for certain sieve priors and enables to extend our results to density estimation.

1. Introduction.

1.1. *Setting*. There has been a growing interest for posterior concentration rates in nonparametric Bayes over the last decade, initiated by the seminal papers of Schwartz [26], Barron [2] and Ghosal, Ghosh and van der Vaart [15]. Consider a statistical model or experiment $\mathcal{E}^n = \{P_\theta^n : \theta \in \Theta\}$ generated by data Y^n, with parameter space Θ equipped with a prior distribution π. The posterior distribution $P^\pi(\cdot|Y^n)$ concentrates at rate $\epsilon_n > 0$ under $P_{\theta_0}^n$ for the loss $\ell : \Theta \times \Theta \to [0, \infty)$ if

$$E_{\theta_0}^n\big[P^\pi\big(\theta : \ell(\theta, \theta_0) > \epsilon_n|Y^n\big)\big] = o(1). \tag{1.1}$$

Posterior concentration allows to uncover frequentist properties of Bayesian methods. It implies that the posterior probability of an ϵ_n-neighbourhood around the true parameter θ_0 converges to one. Thus, most of the posterior mass will be close to the truth in the frequentist sense.

Received May 2013; revised April 2015.
[1]Supported in part by the Agence Nationale de la Recherche (project BANDHIT and Blanc SIMI 1 2011 project CALIBRATION).
[2]Supported by DFG postdoctoral fellowship SCHM 2807/1-1.
MSC2010 subject classifications. Primary 62G20, 62G08; secondary 62G07.
Key words and phrases. Bayesian nonparametrics, minimax adaptive estimation, posterior concentration rates, sup-norm, rates of convergence.

Whenever (1.1) holds uniformly in $\theta_0 \in \Theta$ and if ϵ_n can be taken as constant multiple of the minimax rate of estimation over Θ for the loss ℓ, we say that the concentration rate is asymptotically minimax. We further say that the posterior distribution $P^\pi(\cdot|Y^n)$ concentrates adaptively over the collection $\{\Theta_\beta, \beta \in \mathcal{I}\}$ of subsets of Θ if

$$(1.2)\quad \sup_{\theta_0 \in \Theta_\beta} E^n_{\theta_0}\big[P^\pi\big(\theta : \ell(\theta, \theta_0) > \epsilon_n(\beta)|Y^n\big)\big] = o(1) \qquad \text{for every } \beta \in \mathcal{I},$$

where $\epsilon_n(\beta)$ is a constant multiple of the minimax rate of adaptive estimation over Θ_β. Recently, some families of prior distributions under various types of statistical models have been studied in this light and have been proved to lead to adaptive posterior concentration rates; see Section 5.4 for a more extensive discussion of these results. Similarly as for (1.1), existence of a result of type (1.2) implies that the Bayes estimator is minimax adaptive under fairly general conditions; see Section 5.1. Consequently, existence or nonexistence of adaptive estimators in some nonparametric situations (see, e.g., [4, 20]) yield limitations about the best possible achievable concentration rates $\epsilon_n(\beta)$ in (1.2).

In this paper, we are interested in understanding further the interplay between nonparametric minimax rates of convergence and the existence of adaptive concentration rates for appropriate priors in nonparametric estimation. We cover in particular the two paradigmatic examples of density estimation, when the data Y^n is drawn from a n-sample of an unknown distribution, and the case of a signal observed in Gaussian white noise. More specifically, we attempt to answer the following related questions:

(I) Can we formalise the connexion between posterior concentration rates and the minimax theory: given an experiment $\mathcal{E}^n$ and a loss ℓ, can we define some notion of lower bound associated to the posterior concentration rate? Can we derive a generic construction for a prior with posterior achieving the minimax rate of convergence in the sense of (1.1)? Can we further make this construction adaptive, in the sense of (1.2)?

(II) In the specific framework of the L^2 and L^∞ metric for the loss ℓ, can we construct a feasible prior in standard models such as Gaussian white noise or density estimation for which the posterior distribution contracts adaptively over Hölder balls?

1.2. *Main results.* A first answer to these problems is given in Section 2 in the form of a lower bound on the speed at which the posterior mass outside an ϵ_n-ball vanishes in the sense of (1.1). Assume that Θ is equipped with a pre-metric[3] d that controls the separation rate between two elements in $\mathcal{E}^n$. We prove in Theorem 2.1 that if $\mathcal{E}^n$ is dominated and admits a certain regularity condition then, for every

[3] That is, we only require that d is nonnegative and $d(\theta, \theta') = 0$ iff $\theta = \theta'$.

prior π such that the posterior $P^\pi(\cdot|Y^n)$ concentrates with rate ϵ_n over Θ, there exists a constant $c > 0$ such that

$$\sup_{\theta_0 \in \Theta} E^n_{\theta_0}\big[P^\pi\big(\theta : \ell(\theta, \theta_0) \geq \epsilon_n | Y^n\big)\big] \geq e^{-cn\Omega(\epsilon_n, \Theta, \ell)^2}, \tag{1.3}$$

where

$$\Omega(\epsilon_n, \Theta, \ell) = \inf\{d(\theta, \theta') : \ell(\theta, \theta') \geq 2\epsilon_n, \theta, \theta' \in \Theta\}.$$

The pre-metric d geometrises the statistical model and does not depend on the loss function nor the rate. At this point, one might think of d as the Hellinger distance. If $\epsilon_n \to 0$, only the local behaviour of d plays a role in the definition of $\Omega(\epsilon_n, \Theta, \ell)$, which gives slightly more flexibility and allows to take, for instance, d as the L^2-metric in the Gaussian white noise model. The precise conditions that determine d are stated in Theorems 2.1 and 4.1. Explicit computations are developed in Section 2.

The exponent $\Omega(\epsilon_n, \Theta, \ell)$ appearing in (1.3) is a dual formulation of the modulus of continuity introduced in [14] and further considered by Cai and Low [8], Cai, Low and Zhao [9]; see Section 5.4. Theorem 2.1 also admits a stronger local version: ϵ_n can be a function of θ also in a manner similar to the between classes modulus of continuity of Cai and Low [6]. Another important consequence is that there are limitations of the commonly employed proof strategy for derivation of posterior concentration rates; see Section 5.3.

In Section 3, we address question (II) and explicitly construct a prior—in the family of spike and slab priors—that achieves the lower bound of Theorem 2.1 in the white noise model simultaneously over a collection of Hölder balls $\mathcal{H}(\beta, L)$ for $\beta \in \mathcal{I}$, where $\mathcal{I}$ is a compact subset of $(0, \infty)$. Recasting Y^n into a regular wavelet basis (see, e.g., [12] and [13]), we obtain the sequence model

$$Y_{j,k} = \theta_{j,k} + n^{-1/2}\epsilon_{j,k}, \qquad k \in I_j, j = 0, 1, \dots,$$

where $k \in I_j$ is a location parameter at scale 2^{-j} with I_j having approximately 2^j terms, and the $\epsilon_{j,k}$ are i.i.d. standard normal. The spike and slab prior is constructed as follows: for j less than a maximal resolution level[4] J_n with $2^{J_n} \asymp n$, the $\theta_{j,k}$'s are drawn independently according to

$$\pi_j(dx) = (1 - w_{j,n})\delta_0(dx) + w_{j,n} g(x)\, dx, \tag{1.4}$$

for appropriate level-dependent weights $w_{j,n} > 0$. Here, $\delta_y(dx)$ is the Dirac mass at point y and g is a bounded density on $\mathbb{R}$. For $j > J_n$, we put $\theta_{j,k} = 0$. The construction of the spike and slab prior does not involve knowledge of the smoothness index β. Due to the point mass at zero, the posterior resembles many properties of a wavelet thresholding procedure. In Theorem 3.1, we prove adaptive concentration

[4]In the sequel, we adopt the notation for positive sequences: $a_n \lesssim b_n$ if $\limsup_n a_n/b_n < \infty$ and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$ simultaneously.

rates

$$\sup_{\theta_0 \in \Theta_\beta} E_{\theta_0}^n \big[P^\pi \big(\theta : \|\theta - \theta_0\|_{L^\infty} \geq \epsilon_n(\beta) | Y^n \big) \big] \leq n^{-B}$$

uniformly in $\beta \in \mathcal{I}$, where $\epsilon_n(\beta) = M(n/\log n)^{-\beta/(2\beta+1)}$ and for some constants $B, M > 0$ depending on π and $\mathcal{I}$ only. Moreover, the polynomial speed n^{-B} at which the contraction holds is sharp according to Theorem 2.1 (up to the exponent B). The spike and slab prior (1.4) therefore leads to an adaptive minimax posterior concentration rate over Hölder balls $\mathcal{H}(\beta, L)$ for the sup-norm loss, without additional $\log n$ term. To the best of our knowledge, this is the first construction of a prior leading to an optimal adaptive posterior concentration rate in sup-norm. However, we miss the optimal rate by a logarithmic term if, for the same prior, we consider contraction under the L^2-metric instead of L^∞. We show in Theorem 3.2 how to modify the spike and slab prior in order to remove the logarithmic terms in the L^2-metric and achieve exact adaptation in that setting too.

An answer to question (I) is presented in Section 4. We derive a generic upper bound, neither restricted to the white noise model nor to L^2 or L^∞ losses by considering priors which are uniform over well chosen discrete sieves of Θ. In this abstract framework, Theorem 4.1 provides conditions which imply that there exists a constant $C > 0$ such that

$$\sup_{\theta_0 \in \Theta} E_{\theta_0}^n \big[P^\pi \big(\theta : \ell(\theta_0, \theta) > \epsilon_n | Y^n \big) \big] \leq e^{-Cn\Omega(\epsilon_n, \Theta, \ell)^2}. \tag{1.5}$$

The interesting case is $n\Omega(\epsilon_n, \Theta, \ell)^2 \to \infty$, implying posterior concentration at rate ϵ_n. The rate can also be made adaptive by letting $\epsilon_n = \epsilon_n(\theta_0)$ vary with θ_0. Comparing (1.5) with the lower bound (1.3), we see in particular that the upper and lower bounds agree, up to the constants c and C, and are therefore sharp in that sense. The rather abstract conditions which are required for (1.5) are satisfied in the Gaussian white noise model and for density estimation (Propositions 1–3).

In Section 5, we discuss various implications of the lower and upper bounds (1.3) and (1.5). First, we outline how these bounds on posterior concentration rates lead to the construction of Bayesian estimators having asymptotic minimax (adaptive) frequentist risk, generalising the result of Ghosal, Ghosh and van der Vaart [15], Theorem 2.5. In Section 5.2, we point out the links between posterior coverage and confidence balls. Interestingly, the lower bound (1.3) implies that the classical strategy for derivation of concentration rates fails if an arbitrary loss is considered. This is developed in Section 5.3. Finally, we discuss the relation of the derived results to other works in Section 5.4, both from a frequentist and Bayesian point of view.

2. A generic lower bound.

In this section, we exhibit tractable conditions on the structure of a statistical experiment $\mathcal{E}^n = \{P_\theta^n : \theta \in \Theta\}$ generated by data Y^n in order to obtain an explicit lower bound on

$$E_{\theta_0}^n \big[P^\pi \big(\theta : \ell(\theta, \theta_0) > \epsilon_n | Y^n \big) \big], \qquad \theta_0 \in \Theta,$$

where ϵ_n can be either fixed or a function of θ_0, π is a prior on Θ, $P^\pi(\cdot|Y^n)$ denotes the posterior distribution associated to π and $\ell : \Theta \times \Theta \to [0, \infty)$ is a given loss function.

Assume that the parameter space Θ is equipped with a pre-metric d. Let $\Theta_0 \subset \Theta$ and let $(\epsilon(\theta), \theta \in \Theta_0)$ denote a collection of positive θ-dependent radii over Θ_0. We define a local and a global modulus of continuity related to $\epsilon(\cdot)$ between d and ℓ over a class Θ_0 by setting

$$\Omega(\epsilon(\cdot), \theta, \ell) = \inf\{d(\theta, \theta') : \ell(\theta, \theta') \geq \epsilon(\theta) + \epsilon(\theta'), \theta' \in \Theta_0\} \tag{2.1}$$

and

$$\Omega(\epsilon(\cdot), \Theta_0, \ell) = \inf_{\theta \in \Theta_0} \Omega(\epsilon(\cdot), \theta, \ell). \tag{2.2}$$

To illustrate the meaning of Ω, consider, for instance, the context of the Gaussian white noise model (3.1) developed in Section 3 below, where $\Theta_0 \subset \Theta = L^2([0, 1])$. Take $d = L^2$ and $\ell = L^\infty$ the sup-norm, and for $\beta, L > 0$, let $\Theta_0 = \mathcal{H}(\beta, L)$ be a Hölder ball. Set $\epsilon_n(\theta) = M(n/\log n)^{-\beta/(2\beta+1)}$ for $\theta \in \Theta_0$ and $M > 0$. Then

$$\Omega(\epsilon_n(\cdot), \theta, L^\infty) \lesssim \sqrt{\log n/n} \qquad \text{for every } \theta \in \Theta_0 \tag{2.3}$$

hence

$$\Omega(\epsilon_n(\cdot), \Theta_0, L^\infty) \lesssim \sqrt{\log n/n}$$

as well (for a proof see Section A.1). Similarly, if $\Theta_0 = \mathcal{H}(\beta_1, L) \supset \mathcal{H}(\beta_2, L)$ with $\beta_1 < \beta_2$ and if

$$\epsilon_n(\theta) = \begin{cases} M(n/\log n)^{-\beta_2/(2\beta_2+1)}, & \text{if } \theta \in \mathcal{H}(\beta_2, L), \\ M(n/\log n)^{-\beta_1/(2\beta_1+1)}, & \text{otherwise,} \end{cases} \tag{2.4}$$

then

$$\Omega(\epsilon_n(\cdot), \Theta_0, L^\infty) \lesssim \sqrt{\log n/n}.$$

Obviously, when $d = \ell$, then for all $\theta \in \Theta_0$ we have $\Omega(\epsilon_n(\cdot), \theta, \ell) \geq \epsilon_n(\theta)$ and $\Omega(\epsilon_n(\cdot), \Theta_0, \ell) \geq \inf\{\epsilon_n(\theta), \theta \in \Theta_0\}$.

THEOREM 2.1. *Let $\Theta_0 \subset \Theta$. Let d be a pre-metric on Θ. Assume that ℓ is a pseudo-metric*[5] *on Θ_0, and that the prior π and the family of positive sequences $(\epsilon_n(\theta), \theta \in \Theta_0)$ satisfy the posterior concentration condition*:

$$\sup_{\theta_0 \in \Theta_0} E^n_{\theta_0}\left[P^\pi\left(\theta : \ell(\theta_0, \theta) \geq \epsilon_n(\theta_0)|Y^n\right)\right] = o(1). \tag{2.5}$$

[5] That is, the axioms of a metric are required with $\ell(\theta, \theta) = 0$ but possibly $\ell(\theta, \theta') = 0$ for some distinct $\theta \neq \theta'$.

Assume that the family $\{P_\theta^n : \theta \in \Theta_0\}$ *is dominated by some* σ*-finite measure* μ *and that there exists a constant* $K > 0$ *such that*

$$P_{\theta'}^n\big(\mathcal{L}_n(\theta') - \mathcal{L}_n(\theta) \geq Knd(\theta, \theta')^2\big) = o(1), \tag{2.6}$$

uniformly over all $\theta, \theta' \in \Theta_0$ *satisfying*

$$\Omega(\epsilon_n(\cdot), \theta, \ell) \leq d(\theta, \theta') \leq 2\Omega(\epsilon_n(\cdot), \theta, \ell),$$

where $\mathcal{L}_n(\theta) = \log \frac{dP_\theta^n}{d\mu}(Y^n)$ *denotes the log-likelihood function w.r.t.* μ. *If* $n\Omega(\epsilon_n(\cdot), \Theta_0, \ell)^2 \to \infty$, *then, for all* $\theta_0 \in \Theta_0$ *and large enough* n

$$E_{\theta_0}^n\big[P^\pi\big(\theta : \ell(\theta_0, \theta) > \epsilon_n(\theta_0) | Y^n\big)\big] \geq e^{-3Kn\Omega(\epsilon_n(\cdot), \theta_0, \ell)^2}. \tag{2.7}$$

The proof is delayed until Section 6.

REMARK 1. By taking $\epsilon_n(\theta) = \epsilon_n$ constant on Θ_0, we retrieve the more stringent result (1.3) announced in Section 1.2.

REMARK 2 (About the assumptions). Assumption (2.6) is merely on the pre-metric d that must be related to the intrinsic geometry of the experiment $\mathcal{E}^n$: it shows in particular that d must be able to control locally the likelihood ratio. This can be the Hellinger distance used in the Birgé–Le Cam testing theory in density estimation or simply the L^2-distance in Gaussian white noise model linked to the Hilbert space structure on which relies the existence of an iso-normal process. Note also that since d is not required to be symmetric, the order $d(\theta, \theta')$ is important in assumption (2.6).

In Sections 3 and 4, we show that under some additional assumptions the lower bound (2.7) is sharp.

3. Upper bounds in the white noise model via spike and slab priors.

In this section, we prove that the lower bound obtained in (2.7) is sharp in the white noise model when ℓ is either the sup-norm L^∞ or the L^2-norm. This is done using spike and slab type priors. We observe

$$Y^n = \theta + n^{-1/2}\dot{W}, \tag{3.1}$$

where the signal of interest θ belongs to the Hilbert space

$$\Theta = L^2([0, 1]) = \left\{\theta : [0, 1] \to \mathbb{R} \text{ with } \int_{[0,1]} \theta(x)^2\, dx < \infty\right\}$$

and $\dot{W}$ is a Gaussian white noise on Θ. The noise $\dot{W}$ is not realisable as a random element of L^2; it is therefore viewed as the standard *iso-Gaussian process* for the Hilbert space Θ. Picking an orthonormal wavelet basis, we equivalently observe

$$Y_{j,k} = \theta_{j,k} + n^{-1/2}\epsilon_{j,k}, \qquad \epsilon_{j,k} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1), \qquad j \in \mathbb{N}, k \in I_j, \tag{3.2}$$

where $\theta_{j,k} = \int_0^1 \theta(x)\Psi_{j,k}(x)\,dx$ is the wavelet coefficient associated to a given compactly supported wavelet basis $(\Psi_{j,k})_{(j,k)\in\Lambda}$ of Θ with $\Lambda = \{(j,k), k \in I_j, j \in \mathbb{N}\}$. We append the basis with boundary conditions and assume that it is associated with a R-regular multi-resolution of $L^2([0,1])$; see [12] and [13]. The terms corresponding to $j=0$ incorporate the scaling function and we have that $|I_j|$ is of order 2^j. We identify $\Theta = L^2([0,1])$ with

$$\ell^2(\Lambda) = \left\{\theta = (\theta_{j,k})_{(j,k)\in\Lambda} : \sum_{(j,k)\in\Lambda} \theta_{j,k}^2 < \infty\right\}$$

and we transfer two loss functions on the sequence space model: the L^2-loss

$$\ell_2(\theta,\theta') = \left(\sum_{(j,k)\in\Lambda} (\theta_{j,k} - \theta'_{j,k})^2\right)^{1/2},$$

and the L^∞-loss

$$\ell_\infty(\theta,\theta') = \sum_{j\in\mathbb{N}} 2^{j/2} \max_{k\in I_j} |\theta_{j,k} - \theta'_{j,k}|.$$

Since $(\Psi_{j,k})_{(j,k)\in\Lambda}$ is orthonormal, ℓ_2 coincides with the $L^2([0,1])$ norm. However, the losses ℓ_∞ and L^∞ are not comparable on $\Theta = L^2([0,1])$ identified with $\ell^2(\Lambda)$, but rather on smooth subspaces of Θ. To that end, introduce the Hölder balls[6]

$$\mathcal{H}(\beta, L) = \{\theta = (\theta_{j,k})_{(j,k)\in\Lambda} : |\theta_{jk}| \le L2^{-j(\beta+1/2)}, (j,k) \in \Lambda\} \tag{3.3}$$

for $\beta > 0$, $L > 0$. Then we also have that $\ell_\infty(\theta,\theta')$ and $\|\theta - \theta'\|_{L^\infty([0,1])}$ are comparable on $\mathcal{H}(\beta, L) \subset \ell^2(\Lambda)$.

3.1. *Adaptive posterior concentration rates under sup-norm loss*: *Spike and slab prior.* Throughout the following, let g be a bounded density on $\mathbb{R}$, which satisfies

$$\inf_{x\in[-L_0,L_0]} g(x) > 0$$

for some $L_0 > 0$. We consider the following family of priors on $\Theta = \ell^2(\Lambda)$. Set $J_n = \lfloor \log n/\log 2 \rfloor$ and notice that $n/2 < 2^{J_n} \le n$. For $j \le J_n$ and $k \in I_j$, the $\theta_{j,k}$'s are drawn independently from

$$\pi_j(dx) = (1 - w_{j,n})\delta_0(dx) + w_{j,n} g(x)\,dx. \tag{3.4}$$

[6] Having $\beta = m + \{\beta\}$ with m and integer and $\beta \in (0,1]$, the class $\mathcal{H}(\beta, L)$ coincides with functions $f = \sum_{(j,k)\in\Lambda} \theta_{j,k}\psi_{j,k}$ that are m-times differentiable with $f^{(m)}$ being Hölder continuous of order $\{\beta\}$ provided the regularity of the multi-resolution exceeds β.

For $j > J_n$, $\pi_j(dx) = \delta_0(dx)$, or equivalently, $\theta_{j,k} = 0$. We assume that there are constants $K > 0$, $\tau > 1/2$, such that $n^{-K} \le w_{j,n} \le 2^{-j(1+\tau)}$, for all $j \le J_n$. This constraint on the mixture weights implies in particular that the prior favours sparse models since the individual probability to be nonnull becomes small as the resolution level j increases. We then have the following.

THEOREM 3.1. *Consider a prior in the family of spike and slab priors defined above. If Y^n is drawn from the white noise model* (3.2), *for every $0 < \beta_1 \le \beta_2$ and $L_0 - 1 \ge L > 0$, there exist $M, B > 0$ such that*

$$\sup_{\theta_0 \in \mathcal{H}(\beta, L)} E^n_{\theta_0}\big[P^\pi\big(\theta : \ell_\infty(\theta, \theta_0) \ge M(n/\log n)^{-\beta/(2\beta+1)} | Y^n\big)\big] \le n^{-B}$$

uniformly in $\beta \in [\beta_1, \beta_2]$.

The proof of Theorem 3.1 is given in Section 6.3. It is based on a fine description of the asymptotic behaviour of the posterior distribution on the selected sets of coefficients $\theta_{j,k}$, of the form $S = \{(j,k), \theta_{j,k} \neq 0\}$, that is, we consider coefficients that are not equal to 0 under the posterior distribution. Lemma 1 in Section 6.3 states that the posterior distribution is asymptotically neither forgetting nonnegligible coefficients $\theta^0_{j,k}$ nor selecting too small coefficients $\theta^0_{j,k}$ under $P^n_{\theta_0}$ with $\theta_0 = (\theta^0_{j,k})_{(j,k)\in\Lambda}$. As follows from the proof, if the prior density g is positive and continuous on $\mathbb{R}$, then the conclusion of Theorem 3.1 remains valid for every $L > 0$ and the procedure is independent of both the smoothness β and the radius L.

REMARK 3. Setting $\epsilon_n(\beta) = M(n/\log n)^{-\beta/(2\beta+1)}$, we have

$$\Omega\big(\epsilon_n(\beta), \mathcal{H}(\beta, L), \ell_\infty\big) = O\big(\sqrt{\log n/n}\big)$$

and according to Theorem 2.1, the best possible expectation of the posterior probability of complements on $\epsilon_n(\beta)$ neighbourhoods in ℓ_∞ is at most of polynomial order $n^{-B'}$ for some $B' > 0$. Thus, Theorem 3.1 is sharp up to the constants B' and B.

3.2. *Adaptive posterior concentration rates under L^2 loss: Block spike and slab prior.* Theorem 3.1 implies the existence of $\widetilde{M} > 0$ such that

$$E^n_{\theta_0}\big[P^\pi\big(\theta : \ell_2(\theta, \theta_0) \ge \widetilde{M}(n/\log n)^{-\beta/(2\beta+1)} | Y^n\big)\big] \le n^{-B} \tag{3.5}$$

uniformly in $\beta \in [\beta_1, \beta_2]$ since ℓ_2 is dominated by ℓ_∞. Therefore, an adaptive minimax posterior concentration rate in the ℓ_2-norm is also obtained up to a $\log n$ term. It can indeed be proved that for this prior the $\log n$ term cannot be avoided. Since the spike and slab prior (3.4) is a product measure on the wavelet coefficients, this might be viewed as a Bayesian analogue of the fact that separable rules do not give adaptation with the clean rates in ℓ_2 (cf. Cai [5]). To circumvent this drawback, we propose a block spike and slab prior which achieves the minimax adaptive rate for the ℓ_2-loss without additional $\log n$ term. The posterior associated to this prior is easier to simulate from numerical data since the space of possible selected sets is much smaller than the local spike and slab prior (3.4). It leads, however, to suboptimal posterior concentration rates under sup-norm loss.

For $j \le J_n$, pick a family of independent random vectors $\underline{\theta}_j = (\theta_{j,k})_{k \in I_j}$ for $j \in \mathbb{N}$ according to the distribution

$$\tilde{\pi}_j(dx) = (1 + \nu_{j,n})^{-1}\big(\delta_0(dx) + \nu_{j,n} g_j(x)\, dx\big) \qquad \forall x \in \mathbb{R}^{|I_j|}, \tag{3.6}$$

where g_j is a density with respect to Lebesgue measure on $\mathbb{R}^{|I_j|}$ which satisfies

$$\sup_{x \in \mathbb{R}^{|I_j|}} g_j(x) \le e^{G|I_j|}, \qquad \inf_{x \in [-L_0, L_0]^{|I_j|}} g_j(x) \ge e^{-G|I_j|}, \tag{3.7}$$

and $\nu_{j,n} = n^{|I_j|/2} e^{-c|I_j|}$ for some constants $G > 0$ and $c \ge 4 + G$. For $j > J_n$ put $\theta_{j,k} = 0$. Condition (3.7) is satisfied in particular if, given that group j is not 0, the $\theta_{j,k}$'s are i.i.d. with density g satisfying the same conditions as in the local spike and slab prior (3.4).

THEOREM 3.2. *Consider a prior in the family of spike and slab priors defined above. If Y^n is drawn from the white noise model (3.2), for every $0 < \beta_1 \le \beta_2$ and $L_0 - 1 \ge L > 0$, there exist $M, B > 0$ such that*

$$\sup_{\theta_0 \in \mathcal{H}(\beta, L)} E^n_{\theta_0}\big[P^\pi\big(\theta : \ell_2(\theta, \theta_0) \ge M n^{-\beta/(2\beta+1)} | Y^n\big)\big] \le e^{-B n^{1/(2\beta+1)}}$$

uniformly in $\beta \in [\beta_1, \beta_2]$.

The proof is given in Section 6.4.

REMARK 4. Note that not only do we recover the optimal posterior concentration rate (without any $\log n$ term) but we also bound from above the expectation of the posterior concentration rate by a term of the order

$$\exp\big(-cn\Omega\big(\zeta_n(\beta), \mathcal{H}(\beta, L), \ell_2\big)^2\big)$$

with $\zeta_n(\beta) = n^{-\beta/(2\beta+1)}$ when Ω is computed under the intrinsic metric $d = \ell_2$. The same rate is provided by the lower bound in Theorem 2.1 and is therefore sharp up to the constant $c > 0$.

REMARK 5. Since the prior (3.6) depends neither on β nor on L (in particular if g_j corresponds to $|I_j|$ identically distributed random variables with positive and continuous density g on $\mathbb{R}$) the posterior concentration rate obtained in Theorem 3.2 is moreover adaptive in the minimax sense of (1.2).

4. A generic upper bound. In this section, we explore a more general situation and show that the generic lower bound obtained in Theorem 2.1 is indeed sharp in a wider sense than the one considered in Section 3. In the context of an arbitrary experiment $\mathcal{E}^n$, we construct priors with finite and increasing support, usually referred to as sieve priors. Sieve priors have already been considered in the Bayesian nonparametric literature in some specific context; see [15] and [16]. In both cases, the interest of these priors is that they lead to optimal posterior concentration rates, without additional $\log n$ terms. From a practical point of view, however, the construction of their support and their implementation is close to being impossible. Moreover, they have poor behaviour in terms of credible and confidence sets. In this section, we shall use such priors in the same way, as a device for the existence of an optimal estimation procedure, not as priors to be used in practice.

We adopt the same framework as in Section 2: $\mathcal{E}^n = \{P_\theta^n, \theta \in \Theta\}$ is generated by the observation Y^n and is dominated by some σ-finite measure μ, and $\mathcal{L}_n(\theta) = \frac{dP_\theta^n}{d\mu}(Y^n)$ is a likelihood function. The loss function $\ell : \Theta \times \Theta \to [0, \infty)$ is a pseudo-metric. Let us be given a family

$$\epsilon_n = (\epsilon_n(\theta), \theta \in \Theta)$$

that we understand as the target posterior concentration rate at point θ relative to the loss ℓ. Typically, $\epsilon_n(\theta)$ is the minimax rate of estimation over a subclass $\Theta_0 \subset \Theta$ which contains θ. Let $(\Theta_n)_{n\geq 1}$ be an increasing sequence of compact subsets of Θ for the topology induced by the loss ℓ. More precisely, we only require that Θ_n can be covered by a finite collection of balls centered at θ with radius $\epsilon_n(\theta)$ in terms of the loss ℓ. We denote by N_n the number of such balls and by $\theta_{(l)}$ the centers of these balls for $l = 1, \dots, N_n$. Note that we do not necessarily require that N_n is the minimal number of such balls satisfying the coverage property. We define a *sieve prior* as follows:

$$\pi_n = \frac{1}{N_n} \sum_{l=1}^{N_n} \delta_{\theta_{(l)}}. \tag{4.1}$$

To control the posterior concentration rate, we need to partition the sieve $(\theta_{(l)}, 1 \leq l \leq N_n)$ into slices.

DEFINITION 1. For every $\theta_0 \in \Theta_n$, a partition $(\mathcal{J}_r, 1 \leq r \leq R_n)$ of $\{1, \dots, N_n\}$ (we omit the dependence upon θ_0 in the notation) is called θ_0-admissible if:

(i) There exists $A > 0$ such that $\mathcal{J}_0 = \{l : \ell(\theta_0, \theta_{(l)}) \leq A\epsilon_n(\theta_0)\}$.
(ii) For all $1 \leq r \leq R_n$, $|\mathcal{J}_r| \leq |\mathcal{J}_0|$.

THEOREM 4.1. *Assume that there exist constants $C_0, K_0, K_1 > 0$ and for every $\theta_0 \in \Theta_n$ a θ_0-admissible partition $(\mathcal{J}_r, 1 \leq r \leq R_n)$ together with injective* *maps* $j_r : \mathcal{J}_r \to \mathcal{J}_0$ *such that*

$$P^n_{\theta_0}\big(\exists r, \exists l : \mathcal{L}_n(\theta_{(l)}) - \mathcal{L}_n(\theta_{(j_r(l))}) > -K_0 n d(\theta_{(l)}, \theta_{(j_r(l))})^2\big) \leq C_0 \exp\big(-K_1 n \Omega(\epsilon_n(\cdot), \theta_0, \ell)^2\big) \tag{4.2}$$

and

$$\sum_{r=1}^{R_n} e^{-K_0 n u_r^2} \leq C_0 e^{-K_1 n \Omega(\epsilon_n(\cdot), \theta_0, \ell)^2}, \tag{4.3}$$

where $u_r = \min\{d(\theta_{(l)}, \theta_{(j_r(l))}), l \in \mathcal{J}_r\}$. *Then for all* $\theta_0 \in \Theta_n$,

$$E^n_{\theta_0}\big[P^{\pi_n}\big(\theta : \ell(\theta, \theta_0) > A\epsilon_n(\theta_0) | Y^n\big)\big] \leq 2C_0 e^{-K_1 n \Omega(\epsilon_n(\cdot), \theta_0, \ell)^2}. \tag{4.4}$$

The proof is delayed until Section 6. Conditions (4.2) and (4.3) on the admissible partition are rather abstract. Interestingly, (4.2) is the only condition which links the geometry of Θ to the model $\{P^n_\theta, \theta \in \Theta\}$. To illustrate conditions (4.2)–(4.3) and the admissible partition, consider the following setup:

$$\Theta = \bigcup_{\beta \in [\beta_1, \beta_2]} \mathcal{H}(\beta, L) = \mathcal{H}(\beta_1, L) \subset \ell_2(\Lambda),$$

where the Hölder ball $\mathcal{H}(\beta, L)$ is defined in Section 3. Put $\ell(\theta, \theta') = \ell_\infty(\theta, \theta')$ and $d(\theta, \theta') = \ell_2(\theta, \theta')$. Let $\Theta_n = \{\theta \in \Theta : \theta_{j,k} = 0, \forall j > J_n\}$ with $n < 2^{J_n} \leq 2n$ and set $\phi_n = \phi_0 (\log n/n)^{1/2}$, where $\phi_0 > 0$ is fixed. Define

$$\mathcal{D}_n = \big\{\theta = (a_{j,k}\phi_n, j \leq J_n, k \in I_j), a_{j,k} \in \mathbb{Z} \cap [-L-1, L+1]\big\} \tag{4.5}$$

and identify $\mathcal{D}_n$ as a subset of Θ_n by appending zeros, that is, $\theta_{j,k} = 0$ whenever $j > J_n$. The set $\mathcal{D}_n$ defines the sieve, which we can enumerate as $\{\theta_{(l)}, 1 \leq l \leq N_n\}$ with $N_n = |\mathcal{D}_n|$. For any $\theta_0 = (\theta^0_{j,k})_{(j,k) \in \Lambda} \in \mathcal{H}(\beta, L)$ with $\beta \in [\beta_1, \beta_2]$, there exists an integer $J_n(\beta)$ and a constant b_0 such that

$$\sup_{j > J_n(\beta)} \max_{k \in I_j} |\theta^0_{j,k}| \leq \phi_n/4, \qquad \sum_{j > J_n(\beta)} 2^{j/2} \max_{k \in I_j} |\theta^0_{j,k}| \leq \epsilon_n(\beta), \qquad 2^{J_n(\beta)} \leq b_0 (n/\log n)^{1/(2\beta+1)}, \tag{4.6}$$

and we can pick $\theta^* \in \mathcal{D}_n$ satisfying

$$\forall (j,k) \in \Lambda, \qquad \big|\theta^0_{j,k} - \theta^*_{j,k}\big| \leq \phi_n/2.$$

This implies in particular that $\forall j > J_n(\beta), \forall k \in I_j, \theta^*_{j,k} = 0$, and $\ell_\infty(\theta_0, \theta^*) \leq (\phi_0 + 2)\epsilon_n(\beta)$. We are ready to construct an admissible partition. First, consider the semi-metric $d_1 : \mathcal{D}_n \times \mathcal{D}_n \to [0, \infty)$ defined by

$$d_1(\theta, \theta')^2 = \sum_{j \leq J_n, k \in I_j} \big((\theta_{j,k} - \theta'_{j,k})^2 - \phi_n^2 \mathbf{1}_{\{(j,k) \in \mathcal{U}^c, \theta_{j,k} \wedge \theta'_{j,k} < \theta^0_{j,k} < \theta_{j,k} \vee \theta'_{j,k}\}}\big),$$

where

$$\mathcal{U} = \Big\{(j,k) \in \Lambda,\ j \le J_n,\ \min_{t\in\mathbb{Z}} |\theta^0_{j,k} - t\phi_n| \le \phi_n/4\Big\}. \tag{4.7}$$

Using the semi-metric d_1, we say that $\theta, \theta' \in \mathcal{D}_n$ are equivalent if $d_1(\theta, \theta') = 0$, which defines an equivalence relation. Denote by $\mathcal{I}_r$ the elements of the corresponding quotient space, and let $\mathcal{I}_0$ be the equivalence class of θ^*. Then, for any $\theta \in \mathcal{I}_0$

$$\ell_\infty(\theta_0, \theta) \le \frac{3\phi_n}{4} \sum_{j=0}^{J_n(\beta)} 2^{j/2} + \epsilon_n(\beta) \le \big(3\phi_0 b_0^{1/2} + 1\big)\epsilon_n(\beta).$$

Eventually, we can define for $A \ge 4(3\phi_0 b_0^{1/2} + 1)$ the sets

$$\mathcal{J}_0 = \big\{\theta \in \mathcal{D}_n : \ell_\infty(\theta, \theta_0) \le A\epsilon_n(\beta)\big\}, \qquad \mathcal{J}_r = \mathcal{I}_r \cap \mathcal{J}_0^c, \tag{4.8}$$

where we have identified the partition of the indices with the partition of the elements of $\mathcal{D}_n$. We then have the following.

PROPOSITION 1. *Assume that* $\theta_0 \in \bigcup_{\beta\in[\beta_1,\beta_2]} \mathcal{H}(\beta, L) = \mathcal{H}(\beta_1, L)$, *and consider the partition* $(\mathcal{J}_r, r \ge 0)$ (*depending on* θ_0) *defined as in* (4.8) *above. Then, if* $\ell = \ell_\infty$, *the partition* $(\mathcal{J}_r, r \ge 0)$ *is* θ_0-*admissible and satisfies* (4.3).

Moreover, if Y^n *is drawn from the white noise model* (3.2), *for every* $0 < \beta_1 \le \beta_2$ *and* $L > 0$, *there exist* $M, B > 0$ *such that*

$$\sup_{\theta_0\in\mathcal{H}(\beta,L)} E^n_{\theta_0}\big[P^{\pi_n}\big(\theta : \ell_\infty(\theta, \theta_0) \ge M(n/\log n)^{-\beta/(2\beta+1)} | Y^n\big)\big] \le n^{-B}$$

uniformly in $\beta \in [\beta_1, \beta_2]$.

The proof of Proposition 1 is given in Appendix A.4. The generic upper bound allows us to prove posterior concentration in L^2 loss with the "clean" adaptive rate $\epsilon_n(\beta) = n^{-\beta/(2\beta+1)}$ as well. In fact, we obtain an analogous result to Theorem 3.2 by constructing an appropriate sieve prior and using Theorem 4.1. For sake of brevity, we give the statement without a proof.

PROPOSITION 2. *There exists a sieve prior* π_n, *such that if* Y^n *is drawn from the white noise model* (3.2), *for every* $0 < \beta_1 \le \beta_2$ *and* $L > 0$, *there exist* $M, B > 0$ *with*

$$\sup_{\theta_0\in\mathcal{H}(\beta,L)} E^n_{\theta_0}\big[P^{\pi_n}\big(\theta : \ell_2(\theta, \theta_0) \ge Mn^{-\beta/(2\beta+1)} | Y^n\big)\big] \le \exp\big(-Bn^{1/(2\beta+1)}\big)$$

uniformly in $\beta \in [\beta_1, \beta_2]$.

Even more interesting is that the generic upper bound can be also applied to prove adaptive rates for density estimation, with respect to ℓ_∞ loss. In this model, we observe $Y^n = (Y_1, \dots, Y_n)$, where $Y_i, i = 1, \dots, n$ are independent and identically distributed on [0, 1] with density f_θ and write

$$\sqrt{f_\theta(x)} = \sum_{(j,k)\in\Lambda} \theta_{j,k}\Psi_{j,k}(x). \tag{4.9}$$

Here, the parameter space consists of vectors $\theta = (\theta_{j,k})_{(j,k)\in\Lambda} \in \mathcal{H}(\beta, L)$ such that the right-hand side of (4.9) is larger than some constant $c > 0$ and $\|\theta\|_{\ell_2} = 1$. We refer to this restricted Hölder space as $\mathcal{H}'(\beta, L)$ in the sequel. In this case, we can take $d = \ell_2$ again.

PROPOSITION 3. *There exists a sieve prior π_n, such that if Y^n is drawn from the density model* (4.9), *for every $1/2 < \beta_1 \le \beta_2$ and $L > 0$, there exist $M, B > 0$ with*

$$\sup_{\theta_0\in\mathcal{H}'(\beta,L)} E_{\theta_0}^n\big[P^{\pi_n}\big(\theta : \ell_\infty(\theta, \theta_0) \ge M(n/\log n)^{-\beta/(2\beta+1)}|Y^n\big)\big] \le n^{-B}$$

uniformly in $\beta \in [\beta_1, \beta_2]$.

The proof of Proposition 3 is given in Section A.5.

5. Further results and discussion.

5.1. *Construction of minimax adaptive estimators given adaptive concentration.* The main focus of this work is to study the full posterior distribution under the frequentist assumption of a true parameter θ_0. As a statistical implication of the results let us shortly comment on convergence rates of Bayesian point estimators. In the nonadaptive case, Theorem 2.5 in [15] asserts the existence of an estimator that converges with the posterior concentration rate to the true parameter. However, the construction of the estimator crucially depends on knowledge of the rate ϵ_n and is therefore not applicable in the adaptive setup. Not surprisingly, the Bayes estimator

$$\widehat{\theta} \in \operatorname*{argmin}_{\delta} E^\pi\big[\ell(\delta, \theta)|Y^n\big], \tag{5.1}$$

(assumed to be well defined) (see, e.g., [23]), Chapter 2, will achieve the adaptive rate under quite general conditions. To see this, assume that $\ell(\theta, \theta') = \ell(\theta', \theta)$ for all $\theta, \theta' \in \Theta$ and observe that for any $\theta_0 \in \Theta$,

$$\ell(\widehat{\theta}, \theta_0) \le E^\pi\big[\ell(\widehat{\theta}, \theta) + \ell(\theta, \theta_0)|Y^n\big] \le 2E^\pi\big[\ell(\theta, \theta_0)|Y^n\big]. \tag{5.2}$$

If the loss is bounded, say $\sup_{\theta\in\Theta} \ell(\theta, \theta_0) \le M$, this can be further controlled by

$$2\big(\epsilon_n(\beta) + MP^\pi\big(\theta : \ell(\theta, \theta_0) > \epsilon_n(\beta)|Y^n\big)\big).$$

Consider now a subset $\Theta_\beta \subset \Theta$ such that for any $\theta_0 \in \Theta_\beta$ the posterior concentration rate at θ_0 is bounded by $\epsilon_n(\beta)$ in the slightly stricter sense

$$\sup_{\theta_0 \in \Theta_\beta} E^n_{\theta_0}\big[P^\pi\big(\theta : \ell(\theta, \theta_0) > \epsilon_n(\beta)|Y^n\big)\big] = o\big(\epsilon_n(\beta)\big).$$

Then

$$\sup_{\theta_0 \in \Theta_\beta} P^n_{\theta_0}\big(\ell(\widehat{\theta}, \theta_0) > 2(M+1)\epsilon_n(\beta)\big) = o(1)$$

and

$$\sup_{\theta_0 \in \Theta_\beta} E^n_{\theta_0}\big[\ell(\widehat{\theta}, \theta_0)\big] = O\big(\epsilon_n(\beta)\big).$$

Consequently, $\widehat{\theta}$ achieves the rate $\epsilon_n(\beta)$ over Θ_β. In the case of an unbounded loss functions ℓ, slightly refined arguments can be applied. Consider, for instance, the framework of Theorem 3.2. Here, adaptation is meant over Hölder balls $\mathcal{H}(\beta, L) \subset \mathcal{H}(\beta_1, L)$ with $\beta_1 > 0$. Since $\sup_{\theta, \theta' \in \mathcal{H}(\beta_1, L)} \ell_2(\theta, \theta') \leq M_2 < \infty$ for some constant M_2, we can improve any estimator by projection on $\mathcal{H}(\beta_1, L)$. The projected estimator lies then within ℓ_2-distance M_2 from θ_0. Thus, considering risk of estimators, we may replace the ℓ_2-loss by the modified bounded loss function $\widetilde{\ell}_2 = \min(\ell_2, M_2)$. Together with Theorem 3.2 and the steps described above, the Bayes estimator with respect to $\widetilde{\ell}_2$ yields then an adaptive estimator.

An alternative modification to incorporate unbounded loss functions goes via a slicing of $\ell(\theta, \theta_0)$ in $E^n_{\theta_0}[E^\pi[\ell(\theta, \theta_0)|Y^n]]$. With (5.2),

$$E^n_{\theta_0}\big[\ell(\widehat{\theta}, \theta_0)\big] \leq 2\epsilon_n(\beta) + 2\sum_{j \geq 1}(j+1)\epsilon_n(\beta) E^n_{\theta_0}\big[P^\pi\big(\theta : \ell(\theta, \theta_0) > j\epsilon_n(\beta)|Y^n\big)\big].$$

The second term of the upper bound will typically be negligible (uniformly over Θ_β) since it involves the posterior concentration. In fact, under the conditions of Theorem 3.1, the Bayes estimator in (5.1) adapts to Hölder balls with respect to the sup-norm loss.

PROPOSITION 4. *Consider the spike and slab prior* (3.4) *with* $w_{j,n} \leq n^{-6} 2^{-j(1+\tau)}$ *and* $\tau > 1/2$. *If* Y^n *is drawn from the white noise model* (3.2), *for any* $0 < \beta_1 \leq \beta_2$ *and* $L_0 - 1 \geq L > 0$, *then there exists* $M > 0$ *such that with* $\epsilon_n(\beta) = M(n/\log n)^{-\beta/(2\beta+1)}$,

$$\sup_{\beta \in [\beta_1, \beta_2]} \sup_{\theta_0 \in \mathcal{H}(\beta, L)} P^n_{\theta_0}\big(\ell_\infty(\widehat{\theta}, \theta_0) \geq \epsilon_n(\beta)\big) = o(1)$$

and

$$\sup_{\beta \in [\beta_1, \beta_2]} \epsilon_n(\beta)^{-1} \sup_{\theta_0 \in \mathcal{H}(\beta, L)} E^n_{\theta_0}\big[\ell_\infty(\widehat{\theta}, \theta_0)\big] < \infty.$$

The proof of Proposition 4 is given together with the proof of Theorem 3.1 in Section 6.3.

5.2. *Posterior concentration and confidence balls.* The posterior distribution does not only provide point estimators but also Bayesian measures of uncertainty. Apart from regular parametric models, it is not clear whether such credible sets have a frequentist interpretation as measures of confidence. In this section we discuss some consequences on the asymptotic behaviour of posterior credible balls.

Assume that the prior π leads to a concentration rate ϵ_n over some subset Θ_0 of the parameter space, that is,

$$\sup_{\theta_0\in\Theta_0} E^n_{\theta_0}\big[P^\pi\big(\theta:\ell(\theta,\theta_0)>\epsilon_n|Y^n\big)\big]\le e^{-cn\Omega(\epsilon_n,\Theta_0,\ell)^2}\to 0.$$

As discussed in Section 5.1, this implies under mild conditions existence of a point estimator $\widehat{\theta}_n$ satisfying

$$\sup_{\theta_0\in\Theta_0} P^n_{\theta_0}\big(\ell(\widehat{\theta}_n,\theta_0)>\epsilon_n\big)=o(1).$$

Let $\alpha_n\in(0,1)$ be a sequence, possibly tending to zero, that satisfies $e^{-cn\Omega(\epsilon_n,\Theta_0,\ell)^2}=o(\alpha_n)$. Construct the credible ball

$$C_n=\big\{\theta:\ell(\theta,\widehat{\theta}_n)\le q^\pi_{\alpha_n}\big\},$$

where $q^\pi_{\alpha_n}$ is the $1-\alpha_n$ posterior quantile of $\ell(\theta,\widehat{\theta}_n)$ so that

$$P^\pi\big(\theta\in C_n|Y^n\big)\ge 1-\alpha_n. \tag{5.3}$$

We then have the following two properties for C_n:

$$\begin{aligned}&\int_\Theta P^n_\theta(\theta\in C_n)\,d\pi(\theta)\ge 1-\alpha_n,\\ &\sup_{\theta\in\Theta_0} P^n_\theta\big(\ell(C_n)>4\epsilon_n\big)=o(1),\end{aligned} \tag{5.4}$$

where $\ell(C_n)=\sup\{\ell(\theta,\theta'):\theta,\theta'\in C_n\}=2q^\pi_{\alpha_n}$ is the diameter of C_n.

PROOF OF (5.4). The first inequality is a consequence of the Fubini theorem since (5.3) is true for all Y^n so that

$$\int_\Theta P^n_\theta(\theta\in C_n)\,d\pi(\theta)=\int_{\mathcal{Y}^n} P^\pi\big(\theta\in C_n|Y^n\big)\,dm_\pi\big(Y^n\big)\ge 1-\alpha_n,$$

where m_π is the marginal distribution of Y^n. The second statement of (5.4) follows from the fact that for all $\theta\in C_n$,

$$\ell(\widehat{\theta}_n,\theta)\ge -\ell(\theta_0,\widehat{\theta}_n)+\ell(\theta_0,\theta).$$

Thus, on the event $\{\ell(\theta_0,\widehat{\theta}_n)\le\epsilon_n\}$, for all $t<q^\pi_{\alpha_n}$ and every $\theta_0\in\Theta_0$,

$$\alpha_n\le P^\pi\big(\ell(\theta,\widehat{\theta}_n)>t|Y^n\big)\le P^\pi\big(\ell(\theta_0,\theta)>t-\epsilon_n|Y^n\big)$$

implying

$$P^n_{\theta_0}\big(q^\pi_{\alpha_n} > 2\epsilon_n\big) \le P^n_{\theta_0}\big(\ell(\theta_0, \widehat{\theta}_n) > \epsilon_n\big) + P^n_{\theta_0}\big(P^\pi\big(\ell(\theta_0, \theta) > \epsilon_n | Y^n\big) \ge \alpha_n\big)$$
$$= o(1) + \frac{e^{-cn\Omega(\epsilon_n, \Theta_0, \ell)^2}}{\alpha_n} = o(1),$$

uniformly over $\theta_0 \in \Theta_0$. This completes the proof of (5.4). □

A natural question is then whether the first inequality of (5.4) can be turned into

$$\inf_{\theta \in \Theta} P^n_\theta(\theta \in C_n) \ge 1 - \alpha$$

at least for some reasonably small α. Of particular interest is the case of adaptive posterior concentration rate, which we illustrate considering the sup-norm loss ℓ_∞ over a collection of Hölder balls $\bigcup_{\beta \in [\beta_1, \beta_2]} \mathcal{H}(\beta, L) = \mathcal{H}(\beta_1, L)$ with $0 < \beta_1 \le \beta_2$ and $L > 0$ fixed. Assume that

$$\sup_{\beta \in [\beta_1, \beta_2]} \sup_{\theta_0 \in \mathcal{H}(\beta, L)} E^n_{\theta_0}\big[P^\pi\big(\theta : \ell_\infty(\theta, \theta_0) > \epsilon_n(\beta) | Y^n\big)\big] \le n^{-B}$$

with $\epsilon_n(\beta) = M(n/\log n)^{-\beta/(2\beta+1)}$ for some $M, B > 0$ and

$$\sup_{\beta \in [\beta_1, \beta_2]} \sup_{\theta_0 \in \mathcal{H}(\beta, L)} P^n_{\theta_0}\big(\ell_\infty(\widehat{\theta}, \theta_0) \ge \epsilon_n(\beta)\big) = o(1).$$

By Theorem 3.1 and Proposition 4, this is, for instance, achieved by the prior in (3.4) and the Bayes estimator $\widehat{\theta}$. Let $\alpha_n \ge n^{-B+t}$ for some $t > 0$, then following from (5.4) we obtain

$$\int_\Theta P^n_\theta(\theta \in C_n)\, d\pi(\theta) \ge 1 - \alpha_n,$$
(5.5)
$$\sup_{\beta \in [\beta_1, \beta_2]} \sup_{\theta_0 \in \mathcal{H}(\beta, L)} P^n_{\theta_0}\big(\ell(C_n) > 2\epsilon_n(\beta)\big) = o(1).$$

In this case, there exists no adaptive confidence band (see, e.g., [18]), so that (5.5) implies that for all $\alpha > 0$,

$$\lim_n \inf_{\beta \in [\beta_1, \beta_2]} \inf_{\theta_0 \in \mathcal{H}(\beta, L)} P^n_{\theta_0}(\theta_0 \in C_n) = 0.$$

The nonexistence of adaptive confidence bands means that requiring both honest frequentist coverage and adaptive length of the band is too strong. Integrating out the confidence band is a weaker notion and a possible alternative to the approach of [18] which modifies confidence bands by taking off some points θ and by demanding coverage and adaptive length over this restricted set. Further notice that the first inequality of (5.5) implies that

$$\pi\big(\theta : P^n_\theta(\theta \in C_n) \le 1 - \alpha\big) \le \frac{\alpha_n}{\alpha}.$$

This is, however, not enough to characterise the parameter values θ_0 for which $P^n_{\theta_0}(\theta_0 \in C_n)$ is small. This question is of interest but beyond the scope of the present paper.

5.3. *Consequences on proving strategies for posterior concentration rates.* The lower bound in Theorem 2.1 has an interesting consequence for nonparametric Bayes in general. So far, the state of the art techniques for deriving posterior consistency and concentration rates date back to the work of [26]. Her approach relies on two key ideas. First, treat the numerator and denominator in the Bayes formula separately. Second, introduce an abstract test and express the upper bound in terms of errors of the first and second type. These methods were later refined by [2, 15] and [16]. In particular, from the proof of Theorem 1 of [16], if for ϵ_n their conditions (2.4), (2.5) (associated to the loss ℓ) are satisfied and

$$P_{\theta_0}^n\big[\mathcal{L}_n(\theta) - \mathcal{L}_n(\theta_0) < -n\epsilon_n^2\big] \le e^{-c_1 n\epsilon_n^2}$$

for some positive c_1, then

$$E_{\theta_0}^n\big[P^\pi\big(\theta : \ell(\theta,\theta_0) > M\epsilon_n | Y^n\big)\big] \lesssim e^{-c_2 n\epsilon_n^2}$$

for some $c_2 > 0$. The lower bound of Theorem 2.1, however, implies that

$$\Omega(\epsilon_n, \theta_0, \ell) \gtrsim \epsilon_n.$$

Therefore, if the targeted concentration rate (say the minimax estimation rate over some given class) ϵ_n^* satisfies

$$\Omega(\epsilon_n^*, \theta_0, \ell) = o(\epsilon_n^*)$$

then the approach of [16], Theorem 1, leads to a suboptimal posterior concentration rate. The core of the problem comes from the decomposition of the posterior probability which treats separately the denominator D_n and the numerator N_n (see Section A.2) where the main steps of the arguments of [16] are recalled. Denote by Φ_n the test for $H_0 : \theta = \theta_0$ versus $H_1 : \ell(\theta, \theta_0) > \epsilon_n$. Then the derived upper bound can be written as follows: There exists a positive sequence u_n such that

$$\begin{aligned} &E_{\theta_0}^n\big[P^\pi\big(\theta : \ell(\theta_0,\theta) > \epsilon_n | Y^n\big)\big] \\ &\quad \le E_{\theta_0}^n[\Phi_n] + e^{cnu_n^2} \sup_{\theta:\ell(\theta_0,\theta)>\epsilon_n} E_\theta^n[1-\Phi_n] + e^{-c'nu_n^2}, \end{aligned} \tag{5.6}$$

with finite constants $c, c' > 0$ on which we do not have good control. For the right-hand term of (5.6) to be small, we need

$$\sup_{\theta\in\Theta_n} \mathbf{1}_{\{\ell(\theta_0,\theta)>\epsilon_n\}} E_\theta^n[1-\Phi_n] = o\big(e^{-cnu_n^2}\big).$$

Hence, ϵ_n shall verify the constraint $\Omega(\epsilon_n, \theta_0, \ell) \gtrsim u_n$; if the minimax estimation rate ϵ_n^* over a given class satisfies $\Omega(\epsilon_n^*, \theta_0, \ell) = o(u_n)$, the approach through tests typically leads to suboptimal posterior concentration rates. To illustrate this, consider the white noise model where d is the L^2 loss, $\ell = \ell_\infty$, and θ_0 belonging to a Hölder ball with smoothness β. Assume that $\theta_1 \in \Theta$ satisfies $\ell(\theta_0, \theta_1) > \epsilon_n$ and $\|\theta_1 - \theta_0\|_{L^2} \le C\Omega(\epsilon_n(\cdot), \theta_0, \ell)$ for some fixed arbitrary C. Any test Φ_n with error of first kind smaller than some small ϵ must have a second kind error greater than that of the likelihood ratio test ϕ_{n,θ_1} for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. In other words,

$$E^n_{\theta_1}[1 - \Phi_n] \geq E^n_{\theta_1}[1 - \phi_{n,\theta_1}] \gtrsim e^{-c_1 n \|\theta_1 - \theta_0\|^2_{L^2}} \geq e^{-nc_1 C \Omega(\epsilon_n, \theta_0, \ell)^2}$$

for some $c_1 > 0$. This implies $\Omega(\epsilon_n, \theta_0, \ell) \gtrsim u_n$. The above argument can be generalised to other models, in particular, to density estimation with $\theta_0 \in \mathcal{H}(\beta, L)$ for $L, \beta > 0$. If we rely on the bound (5.6), the best achievable concentration rate is given by the (ℓ, d)-modulus of continuity $\omega(u_n)$ as defined in (5.7) below. As an example, consider density estimation. Any prior which leads to the minimax estimation error $n^{-\beta/(2\beta+1)}$ in the Hellinger metric gives $u_n = n^{-\beta/(2\beta+1)}$ in (5.6) (possibly up to $\log n$ terms). Since for ℓ the sup-norm and $\theta_0 \in \mathcal{H}(\beta, L)$, $\omega(n^{-\beta/(2\beta+1)}) \lesssim n^{-(\beta-1/2)/(2\beta+1)}$, this explains the (suboptimal) rate observed in [17] which was derived using the standard approach, and thus a bound of the type (5.6).

5.4. *Relation to other works*. In the last decade, a variety of posterior concentration rates have been derived. These studies include density estimation in the case of independent and identically distributed observations as in [15], nonparametric regression (Ghosal and van der Vaart [16]) and the white noise model (Zhao [31], Belitser and Ghosal [3]), Markov models (Tang and Ghosal [29]), Gaussian time series (Choudhuri, Ghosal and Roy [11] and Rousseau, Chopin and Liseo [25]) to name but a few, or the recent canonical statistical setting of [10]. For each of these models, a variety of families of priors have been investigated. An interesting feature of the Bayesian nonparametric approaches considered in these papers is that minimax adaptive concentration rates are achieved using hierarchical types of priors where, at the highest level of hierarchy some hyperparameter, somehow related to the class index β, is itself given a prior distribution. For instance, in the case of density estimation, the renown class of Dirichlet process mixtures or related types of mixtures lead to adaptive posterior concentration rates over collections of Hölder balls of regularity β, up to a $\log n$ term, see, for instance, [19, 24, 28] and [27] under the Hellinger or the L^1 losses on the densities. Gaussian random fields, with inverse Gamma bandwidth as prior models also lead to adaptive posterior concentration rates up to a $\log n$ term for a large class of models, including the nonlinear regression model under the empirical quadratic loss on the design and the classification problem under the L^2 loss; see [30]. Similarly, orthonormal basis expansions with random truncation generically yield adaptive posterior concentration rates too, provided the loss function is well chosen; see [1]. All these results, however, are proved using the approach proposed by [16], which relies on the existence of tests with exponentially small error of the second kind outside ℓ-neighbourhoods of the true parameter. Therefore, these results are applicable to loss functions which behave similarly to d.

Previous to this work, suboptimal asymptotic behaviour of posterior distributions has been observed for specific loss functions. Arbel, Gayraud and and Rousseau [1] shows, for instance, that a random truncation prior with minimax adaptive (up to a $\log n$ term) posterior concentration rate under L^2 loss leads to significantly suboptimal posterior concentration rate and (and risk) under pointwise loss.

To our knowledge, the question of the existence of adaptive minimax posterior concentration rates when ℓ is the pointwise loss or even the sup-norm loss L^∞ has been an open question until now. A consequence of our results is the explicit construction of priors that lead to adaptive concentration rates for various loss functions (including the sup-norm L^∞). Given a prior π and a loss function ℓ, the best achievable rate of concentration of the posterior distribution is intimately linked to the geometry of the experiment $\mathcal{E}^n = \{P_\theta^n, \theta \in \Theta\}$ in the most classical sense of Le Cam (see, e.g., [21]), expressed through the pre-metric d. The behaviour of such a pair (ℓ, d) is reminiscent of several phenomena in minimax theory: these include estimation of linear functionals [14], constrained risk inequalities [4, 7, 9] or the existence of adaptive confidence sets [6, 22]. In all these studies, a key ingredient is the behaviour of a (ℓ, d) modulus of continuity

$$\omega(\epsilon) = \sup\{\ell(\theta, \theta') : d(\theta, \theta') \le \epsilon, \theta, \theta' \in \Theta\}, \qquad \epsilon > 0 \tag{5.7}$$

that quantifies the maximal error in the desired ℓ-loss for a prescribed statistical distance ϵ induced by the experiment $\mathcal{E}^n$ via the intrinsic pre-metric d. More precisely, if there are two sequences $\epsilon_n > 0$ and $\theta_n \in \Theta$, such that $d(\theta_0, \theta_n) \le \epsilon_n$ implies that there exists no convergent test of

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_n : \theta = \theta_n,$$

then $\omega(\epsilon_n)$ yields a lower bound for the minimax estimation rate of θ in ℓ-loss. The nonexistence of adaptive confidence intervals over Hölder balls in the Gaussian white noise model lies at the heart of this simple phenomenon: In that case, ℓ is the pointwise or L^∞-norm and d is the L^2-metric. The fact that an irregular function can be close to a smooth functions in $L^2([0, 1])$ while away from the smooth target in L^∞ explains the negative result of Low [22] [see also [17, 18]] and is quantified by $\omega(\epsilon_n)$. Interestingly, in the Bayesian framework, the lower bound derived in Theorem 2.1 is of similar nature and the modulus of continuity defined in (2.2) is the dual of the modulus of continuity considered in the frequentist minimax literature and defined in (5.7).

6. Proofs.

6.1. *Proof of Theorem* 2.1. We prove Theorem 2.1 by contradiction. Assume that there exist $\theta_0 \in \Theta_0$ such that

$$E_{\theta_0}^n\big[P^\pi\big(\theta : \ell(\theta_0, \theta) > \epsilon_n(\theta_0) | Y^n\big)\big] < e^{-3Kn\Omega(\epsilon_n(\cdot), \theta_0, \ell)^2}, \tag{6.1}$$

infinitely often, which without loss of generality we can assume to be satisfied for all n. By definition of $\Omega(\epsilon_n(\cdot), \theta_0, \ell)$, we can choose a sequence $(\theta_n^*)_n \subset \Theta_0$ satisfying

$$\Omega(\epsilon_n(\cdot), \theta_0, \ell) \le d(\theta_0, \theta_n^*) \le 2\Omega(\epsilon_n(\cdot), \theta_0, \ell)$$

and

$$\ell(\theta_0, \theta_n^*) \ge \epsilon_n(\theta_0) + \epsilon_n(\theta_n^*)$$

simultaneously. Then for every $\theta \in \Theta$,

$$\ell(\theta, \theta_n^*) < \epsilon_n(\theta_n^*) \quad \Rightarrow \quad \ell(\theta, \theta_0) > \ell(\theta_0, \theta_n^*) - \epsilon_n(\theta_n^*) \ge \epsilon_n(\theta_0)$$

so that

$$\begin{aligned}
&E_{\theta_n^*}^n\left[P^\pi\left(\theta : \ell(\theta_n^*, \theta) < \epsilon_n(\theta_n^*)|Y^n\right)\right] \\
&\quad \le E_{\theta_n^*}^n\left[P^\pi\left(\theta : \ell(\theta_0, \theta) > \epsilon_n(\theta_0)|Y^n\right)\right] \\
&\quad \le e^{Knd(\theta_0,\theta_n^*)^2} E_{\theta_0}^n\left[P^\pi\left(\theta : \ell(\theta_0, \theta) > \epsilon_n(\theta_0)|Y^n\right)\right] \\
&\qquad + P_{\theta_n^*}^n\left(\mathcal{L}_n(\theta_n^*) - \mathcal{L}_n(\theta_0) > Knd(\theta_0, \theta_n^*)^2\right) \\
&\quad \le e^{-(3K-2K)n\Omega(\epsilon_n(\cdot),\theta_0,\ell)^2} + P_{\theta_n^*}^n\left(\mathcal{L}_n(\theta_n^*) - \mathcal{L}_n(\theta_0) > Knd(\theta_0, \theta_n^*)^2\right) \\
&\quad = o(1)
\end{aligned}$$

in contradiction with the posterior concentration (2.5).

6.2. *Proof of Theorem* 4.1. For $\theta_0 \in \Theta_n$, let $(\mathcal{J}_r, 0 \le r \le R_n)$ be a θ_0-admissible partition satisfying (4.2) and (4.3). Let $A_n(\theta_0) = \{\theta \in \Theta_n : \ell(\theta_0, \theta) > A\epsilon_n(\theta_0)\}$, where A is defined via the admissible partition in (4.8). Set $p_{n,\theta}(Y^n) = \frac{dP_\theta^n}{d\mu}(Y^n)$ so that $\mathcal{L}_n(\theta) = \log p_{n,\theta}$. For the sieve prior π_n defined in (4.1),

$$\begin{aligned}
P^{\pi_n}\left(A_n(\theta_0)|Y^n\right) &= \frac{\sum_{l=1}^{N_n} \mathbf{1}_{A_n(\theta_0)}(\theta_{(l)}) p_{n,\theta_{(l)}}(Y^n)}{\sum_{l=1}^{N_n} p_{n,\theta_{(l)}}(Y^n)} \\
&\le \sum_{r=1}^{R_n} \frac{\sum_{l\in\mathcal{J}_r} p_{n,\theta_{(l)}}(Y^n)}{\sum_{l\in\mathcal{J}_0} p_{n,\theta_{(l)}}(Y^n)} \le \sum_{r=1}^{R_n} \max_{l\in\mathcal{J}_r} e^{\mathcal{L}_n(\theta_{(l)})-\mathcal{L}_n(\theta_{(j_r(l))})}.
\end{aligned}$$

Let

$$\Omega_n(\theta_0) = \left\{\forall r \ge 1, \forall l \in \mathcal{J}_r : \mathcal{L}_n(\theta_{(l)}) - \mathcal{L}_n(\theta_{(j_r(l))}) \le -K_0 nd(\theta_{(l)}, \theta_{(j_r(l))})^2\right\}.$$

On $\{Y^n \in \Omega_n(\theta_0)\}$,

$$\begin{aligned}
P^{\pi_n}\left(A_n(\theta_0)|Y^n\right) &\le \sum_{r=1}^{R_n} e^{-K_0 nd(\theta_{(l)},\theta_{(j_r(l))})^2} \\
&\le \sum_{r=1}^{R_n} e^{-K_0 n u_r^2} \le C_0 e^{-K_1 n\Omega(\epsilon_n(\cdot),\theta_0,\ell)^2},
\end{aligned}$$

thanks to assumption (4.3), which combined with assumption (4.2) completes the proof.

6.3. *Proof of Theorem* 3.1 *and of Proposition* 4. Recall that the prior can be written in the following hierarchical way: First, select a set of nonzero components S, with distribution P the product of independent Bernoulli random variables $\mathcal{B}(w_{j,n})$ for $j \le J_n$. Given S, draw independently $\theta_{j,k} \sim g$ for all $(j,k) \in S$, and put $\theta_{j,k} = 0$ otherwise.

Asymptotically, the posterior concentrates on supports S containing only indices (j,k) with $|\theta^0_{j,k}| > \underline{\gamma}\sqrt{\log n/n}$ and all indices (j,k) with $|\theta^0_{j,k}| > \overline{\gamma}\sqrt{\log n/n}$, where $0 < \underline{\gamma} < \overline{\gamma} < \infty$ are appropriate constants. In this respect, the posterior behaves similar as hard thresholding. Indeed, for

$$\mathcal{J}_n(\gamma) = \{(j,k) \in \Lambda : |\theta^0_{j,k}| > \gamma\sqrt{\log n/n}\} \qquad \text{with } \gamma > 0$$

we have the following.

LEMMA 1. *Under the conditions of Theorem* 3.1, *for every* $0 < \beta_1 \le \beta_2$, $L \le L_0 - 1$, *and any* $B > 0$, *there exists* $\overline{\gamma} > 0$ *such that*

$$\sup_{\beta_1 \le \beta \le \beta_2} \sup_{\theta_0 \in \mathcal{H}(\beta,L)} E^n_{\theta_0}\big[P^\pi\big(S^c \cap \mathcal{J}_n(\overline{\gamma}) \ne \varnothing | Y^n\big)\big] \lesssim \frac{\log n}{n^B}. \tag{6.2}$$

Suppose that the mixing weights in the spike and slab prior (3.4) *satisfy* $w_{j,n} \le \min(\frac{1}{2}, n^{(\tau\wedge 1)/2 - 1/4 - 2B} 2^{-j(1+\tau)})$ *with* $B > 0$, $\tau > 1/2$. *Then, for sufficiently small* $0 < \underline{\gamma}$,

$$\sup_{\beta_1 \le \beta \le \beta_2} \sup_{\theta_0 \in \mathcal{H}(\beta,L)} E^n_{\theta_0}\big[P^\pi\big(S \cap \mathcal{J}^c_n(\underline{\gamma}) \ne \varnothing | Y^n\big)\big] \lesssim \frac{\log n}{n^B}. \tag{6.3}$$

The proof of Lemma 1 is delayed until Appendix A.3. Suppose for the moment that for any $B > 0$ the bound $\overline{\gamma}$ can be chosen large enough such that

$$\sup_{\theta_0 \in \mathcal{H}(\beta,L)} E^n_{\theta_0}\Big[P^\pi\Big(\max_{(j,k)\in\mathcal{J}_n(\underline{\gamma})} |\theta_{j,k} - \theta^0_{j,k}| > \overline{\gamma}\sqrt{\log n/n} | Y^n\Big)\Big] \lesssim \frac{\log n}{n^B}, \tag{6.4}$$

uniformly in $\beta \in [\beta_1, \beta_2]$. The last estimate ensures that the posterior concentrates around $\theta^0_{j,k}$ with the good rate $\sqrt{\log n/n}$ on every component (j,k) on which signal might be detected. Now we are ready to complete the proof of Theorem 3.1. The definition of a Hölder ball in (3.3) implies that there exists a $J_n(\beta)$ with $2^{J_n(\beta)} \le k_2(n/\log n)^{1/(2\beta+1)}$ for some constant $k_2 > 0$ such that $\mathcal{J}_n(\underline{\gamma}) \subset \{(j,k) : j \le J_n(\beta), k \in I_j\}$ and

$$\sup_{\theta_0 \in \mathcal{H}(\beta,L)} \sum_{j > J_n(\beta)} 2^{j/2} \max_{k \in I_j} |\theta^0_{j,k}| \le \frac{1}{2} M (n/\log n)^{-\beta/(2\beta+1)} =: \frac{1}{2}\epsilon_n(\beta),$$

for M a sufficiently large constant. In order to prove the theorem, it is sufficient to show that $\ell_\infty(\theta,\theta_0) \le \epsilon_n(\beta)$ for all θ with $\max_{(j,k)\in\mathcal{J}_n(\underline{\gamma})}|\theta_{j,k}-\theta_{j,k}^0| \le \overline{\gamma}\sqrt{\log n/n}$ and support S satisfying the constraints $S^c \cap \mathcal{J}_n(\overline{\gamma}) = \varnothing$ and $S \cap \mathcal{J}_n^c(\underline{\gamma}) = \varnothing$. Using the properties of $J_n(\beta)$,

$$\ell_\infty(\theta,\theta_0) \le \sum_{j=0}^{J_n(\beta)} 2^{j/2}\max_{k\in I_j}|\theta_{j,k}-\theta_{j,k}^0| + \frac{1}{2}\epsilon_n(\beta) \le \overline{\gamma}2^{J_n(\beta)/2}\sqrt{\log n/n} + \frac{1}{2}\epsilon_n(\beta)$$

and the right-hand side can further be uniformly bounded by $\epsilon_n(\beta)$. This establishes Theorem 3.1 provided (6.4) is true.

For Theorem 3.1, it therefore remains to show (6.4). By Lemma 1, we can restrict ourselves to parameters with support S satisfying $S^c \cap \mathcal{J}_n(\overline{\gamma}) = \varnothing$. Using a union bound and considering the cases $\underline{\gamma}\sqrt{\log n/n} < |\theta_{j,k}^0| \le \overline{\gamma}\sqrt{\log n/n}$ and $\overline{\gamma}\sqrt{\log n/n} < |\theta_{j,k}^0|$ separately,

$$P^\pi\Big(\max_{(j,k)\in\mathcal{J}_n(\underline{\gamma})}|\theta_{j,k}-\theta_{j,k}^0| > \overline{\gamma}\sqrt{\log n/n} \text{ and } S^c\cap\mathcal{J}_n(\overline{\gamma}) = \varnothing | Y^n\Big)$$
$$\lesssim n \max_{(j,k)\in\mathcal{J}_n(\underline{\gamma})} P^\pi\big(|\theta_{j,k}-\theta_{j,k}^0| > \overline{\gamma}\sqrt{\log n/n} \text{ and } \theta_{j,k}\neq 0|Y^n\big).$$

Consider the event

$$\Omega_{n,B} = \big\{\sqrt{n}|Y_{j,k}-\theta_{j,k}^0| \le \big(2\log|I_j| + 2B\log n\big)^{1/2}, \forall j \le J_n, \forall k \in I_j\big\}.$$

Then

$$P_{\theta_0}^n(\Omega_{n,B}^c) \le 2n^{-B}J_n \le \frac{2\log n}{n^B}. \tag{6.5}$$

For all $j \le J_n$, $k \in I_j$, on $\Omega_{n,B}$, $|Y_{j,k}| \le |\theta_{j,k}^0| + \frac{1}{2} \le L_0 - \frac{1}{2}$ and so if $a = \inf\{g(x) : |x| \le L_0\} > 0$, then, setting $u_0 = \Phi^{-1}((1+1/\sqrt{2})/2)$

$$\int_{\mathbb{R}} e^{-(n/2)(Y_{j,k}-\theta)^2} g(\theta)\,d\theta \ge a(2\pi/n)^{1/2}\big(2\Phi(u_0)-1\big) \ge a(\pi/n)^{1/2}, \tag{6.6}$$

where $\Phi(x) = \Pr(\mathcal{N}(0,1) \le x)$. For any $(j,k) \in \mathcal{J}_n(\underline{\gamma})$,

$$P^\pi\big(|\theta_{j,k}-\theta_{j,k}^0| > \overline{\gamma}\sqrt{\log n/n} \text{ and } \theta_{j,k}\neq 0|Y^n\big)$$
$$\le a^{-1}\sup_x g(x)\Big(\frac{n}{\pi}\Big)^{1/2}\int_{\mathbb{R}} \mathbf{1}\big\{|\theta-\theta_{j,k}^0| > \overline{\gamma}\sqrt{\log n/n}\big\}e^{-(n/2)(Y_{j,k}-\theta)^2}\,d\theta.$$

On $\Omega_{n,B}$,

$$\big\{|\theta-\theta_{j,k}^0| > \overline{\gamma}\sqrt{\log n/n}\big\} \subset \big\{|\theta - Y_{j,k}| > \tfrac{1}{2}\overline{\gamma}\sqrt{\log n/n}\big\}$$

provided $\overline{\gamma}$ is large enough. Therefore, for any $(j,k) \in \mathcal{J}_n(\underline{\gamma})$,

$$P^{\pi}\big(|\theta_{j,k} - \theta^0_{j,k}| > \overline{\gamma}\sqrt{\log n/n} \text{ and } \theta_{j,k} \neq 0 | Y^n\big) \leq a^{-1} \sup_x g(x) 2^{3/2} e^{-\overline{\gamma}^2 \log n/8}$$

and together with the union bound and the estimate of $P^n_{\theta_0}(\Omega^c_{n,B})$ above, equation (6.4) follows for $\overline{\gamma}$ sufficiently large. The proof of Theorem 3.1 is complete.

The proof of Proposition 4 relies on the computations above. Define $A_1 = \{\max_{(j,k)\in\mathcal{J}_n(\underline{\gamma})} |\theta_{j,k} - \theta^0_{j,k}| \leq \overline{\gamma}\sqrt{\log n/n}\}$, $A_2 = \{S : S^c \cap \mathcal{J}_n(\overline{\gamma}) = \varnothing\}$ and $A_3 = \{S : S \cap \mathcal{J}_n(\underline{\gamma}) = \varnothing\}$. On $A_1 \cap A_2 \cap A_3$, $\ell_\infty(\theta, \theta_0) \leq M(n/\log n)^{-\beta/(2\beta+1)}$ for some $M > 0$. Thus, with (5.2), Proposition 4 is proved if

$$E_{\theta_0}\big[E^{\pi}\big(\ell_\infty(\theta,\theta_0)(\mathbf{1}_{A_1^c\cap A_2\cap A_3} + \mathbf{1}_{A_2^c} + \mathbf{1}_{A_3^c})|Y^n\big)\big] \leq \frac{\log n}{\sqrt{n}}. \tag{6.7}$$

Let A be a measurable subset of the parameter set, then using the Cauchy–Schwarz inequality twice,

$$\begin{aligned}
&E^n_{\theta_0}\big[E^{\pi}\big(\ell_\infty(\theta,\theta_0)\mathbf{1}_A|Y^n\big)\big] \\
&\quad\lesssim \sum_{j,k} 2^{j/2} E^n_{\theta_0}\big[E^{\pi}\big(|\theta_{j,k} - \theta^0_{j,k}|^2|Y^n\big)\big]^{1/2} E^n_{\theta_0}\big[P^{\pi}\big(A|Y^n\big)\big]^{1/2} \\
&\quad\leq 2\sum_{j,k} 2^{j/2}\bigg(E^n_{\theta_0}\big[E^{\pi}\big(|\theta_{j,k} - Y_{j,k}|^2|Y^n\big)\big] + \frac{1}{n}\bigg)^{1/2} E^n_{\theta_0}\big[P^{\pi}\big(A|Y^n\big)\big]^{1/2}.
\end{aligned}$$

We apply this inequality to $A = A_1^c \cap A_2 \cap A_3$, A_2^c, and A_3^c. Using the bounds above, it is sufficient to control $E^n_{\theta_0}[E^{\pi}((\theta_{j,k} - Y_{j,k})^2|Y^n)]$. Recall the definition of the spike and slab prior (3.4) and observe

$$\begin{aligned}
&E^{\pi}\big((\theta_{j,k} - Y_{j,k})^2|Y^n\big) \\
&\quad\leq Y^2_{j,k} + \frac{2w_{j,n}\sup_x g(x)}{a}\frac{\int_{\mathbb{R}}(\theta - Y_{j,k})^2 e^{-n(\theta - Y_{j,k})^2/2}\,d\theta}{\int_{-L_0}^{L_0} e^{-n(\theta - Y_{j,k})^2/2}\,d\theta} \\
&\quad\leq Y^2_{j,k} + \frac{2w_{j,n}\sup_x g(x)}{an} \\
&\qquad\times\big[\Phi\big(\sqrt{n}\big(L_0 - \theta^0_{j,k}\big) - \epsilon_{j,k}\big) - \Phi\big(-\sqrt{n}\big(L_0 + \theta^0_{j,k}\big) - \epsilon_{j,k}\big)\big]^{-1}
\end{aligned}$$

with $\epsilon_{j,k} = \sqrt{n}(Y_{j,k} - \theta^0_{j,k})$ and Φ the distribution function of a standard normal random variable. Since $|\theta^0_{j,k}| \leq L_0 - 1$,

$$\begin{aligned}
&\int_0^\infty e^{-\epsilon^2/2}\big(\Phi\big(\sqrt{n}\big(L_0 - \theta^0_{j,k}\big) - \epsilon\big) - \Phi\big(-\sqrt{n}\big(L_0 + \theta^0_{j,k}\big) - \epsilon\big)\big)^{-1}\,d\epsilon \\
&\quad\leq \int_0^\infty e^{-\epsilon^2/2}\big(\Phi(\sqrt{n} - \epsilon) - \Phi(-\sqrt{n} - \epsilon)\big)^{-1}\,d\epsilon
\end{aligned}$$

$$\lesssim \int_0^{\sqrt{n}} e^{-\epsilon^2/2}\, d\epsilon + \int_{\sqrt{n}}^{\infty} \epsilon e^{n/2-\epsilon\sqrt{n}}\, d\epsilon$$
$$\lesssim 1 + e^{-n/2}.$$

The same inequality can be obtained for the integral over $(-\infty, 0)$. Consequently, there exists a universal constant $C > 0$ for which

$$E_{\theta_0}^n\big[E^\pi\big((\theta_{j,k} - Y_{j,k})^2 | Y^n\big)\big] \le \big(\theta_{j,k}^0\big)^2 + \frac{1}{n} + \frac{2C w_{j,n} \sup_x g(x)}{a}.$$

Since $|\theta_{j,k}^0| \lesssim 2^{-j/2}$ and $w_{j,n} \le 2^{-j}$, we obtain that for any measurable set A and uniformly over $\theta_0 \in \mathcal{H}(\beta, L)$,

$$E_{\theta_0}^n\big[E^\pi\big(\ell_\infty(\theta, \theta_0)\mathbf{1}_A | Y^n\big)\big] \lesssim n E_{\theta_0}^n\big[P^\pi\big(A|Y^n\big)\big]^{1/2}.$$

From the proof of Theorem 3.1 above, we find that the right-hand side is of order $\log n/\sqrt{n}$, provided that the exponent B in Lemma 1 and (6.4) is three. This completes the proof of Proposition 4.

6.4. *Proof of Theorem* 3.2. We set $\underline{Y}_j = (Y_{j,k}, k \in I_j)$ and similarly $\underline{\theta}_j = (\theta_{j,k}, k \in I_j)$. Whenever convenient, we identify $\underline{Y}_j$ and $\underline{\theta}_j$ as sequences indexed by the whole set of indices Λ, setting their value to be 0 on the complement of I_j. Thus, if $\|\cdot\|$ denotes the usual Euclidean norm on $\mathbb{R}^{|I_j|}$, we have $\ell_2(\underline{\theta}_j, \underline{\theta}'_j) = \|\underline{\theta}_j - \underline{\theta}'_j\|$ with a slight abuse of notation.

The proof of Theorem 3.2 follows the classical line for studying posterior concentration rates as proposed in [16], with some extra care that has to be taken in order to avoid the usual $\log n$ term that appears in this case. Set $u_n(\beta) = n^{-\beta/(2\beta+1)}$ and let $\widetilde{J}_n(\beta)$ satisfy $K_1 n^{1/(2\beta+1)} \le 2^{\widetilde{J}_n(\beta)} \le 2K_1 n^{1/(2\beta+1)}$, where K_1 is a constant to be large enough. Define

$$\Theta_n(\beta) = \{\theta : \theta_{j,k} = \theta_{j,k}\mathbf{1}_{\{j \le \widetilde{J}_n(\beta), k \in I_j\}}\}.$$

We first prove that for some $c_1, K_1 > 0$,

$$P^\pi\big(\Theta_n(\beta)^c | Y^n\big) \le e^{-c_1 n u_n^2(\beta)}. \tag{6.8}$$

Let $\theta_0 = (\theta_{j,k}^0)_{(j,k)\in\Lambda} \in \mathcal{H}(\beta, L)$ with $L \le L_0 - 1$ and L_0 is the constant appearing in condition (3.7). Denote by B_n the intersection of the events

$$\big\{Y^n : n\big\|\underline{Y}_j - \underline{\theta}_j^0\big\|^2 \le e|I_j|, \forall j \text{ with } \widetilde{J}_n(\beta) \le j \le J_n\big\}$$

and

$$\big\{Y^n : \big|Y_{j,k} - \theta_{j,k}^0\big| \le 1/2, \forall j \le \widetilde{J}_n(\beta), k \in I_j\big\}.$$

Set $c_e = e/2 - 1$ and $C_e = (1 - e^{-c_e})^{-1}$. For a χ_p^2 distributed random variable ξ, we have $\Pr(\xi > eq) \le e^{-c_e q}$ whenever $q \ge p$. Hence,

$$P_{\theta_0}^n(B_n^c) \le 2ne^{-n/8} + \sum_{j=\widetilde{J}_n(\beta)}^{J_n} e^{-c_e|I_j|} \le 2ne^{-n/8} + C_e e^{-c_e|I_{\widetilde{J}_n(\beta)}|} \le e^{-An^{1/(2\beta+1)}},$$

for n large enough, with A proportional to K_1. Since

$$\Theta_n(\beta)^c = \bigcup_{j \ge \widetilde{J}_n(\beta)} \{\theta : \theta_{I_j} \ne 0\}$$

(here $\theta_{I_j} \ne 0$ means $\theta_{j,k} \ne 0$ for at least one $k \in I_j$) we conclude

$$\begin{aligned} P^\pi\big(\Theta_n(\beta)^c|Y^n\big) &\le \sum_{j \ge \widetilde{J}_n(\beta)} \frac{(1+\nu_{j,n})^{-1}\nu_{j,n} \int_{\mathbb{R}^{|I_j|}} e^{-(n/2)\|\underline{\theta}_j - \underline{Y}_j\|^2} g_j(\underline{\theta}_j)\, d\underline{\theta}_j}{\int_{\mathbb{R}^{|I_j|}} e^{-(n/2)\|\underline{\theta}_j - \underline{Y}_j\|^2} d\widetilde{\pi}_j(\underline{\theta}_j)} \\ &\le \sum_{j \ge \widetilde{J}_n(\beta)} e^{G|I_j|}\nu_{j,n}\left(\frac{2\pi}{n}\right)^{|I_j|/2} \exp\left(\frac{n\|\underline{Y}_j\|^2}{2}\right). \end{aligned}$$

For all $j \ge \widetilde{J}_n(\beta)$, we have $\|\underline{\theta}_j^0\|^2 \le L^2|I_j|2^{-j(2\beta+1)} \le C|I_j|/n$, for the radius of the Hölder ball L and some constant $C > 0$ which decreases to zero as K_1 grows. On the event B_n, we thus infer $n\|\underline{Y}_j\|^2 \le 2(C+e)|I_j|$. Therefore, on B_n,

$$\begin{aligned} P^\pi\big(\Theta_n(\beta)^c|Y^n\big) &\le \sum_{j \ge \widetilde{J}_n(\beta)} e^{(G+C+e)|I_j|}\nu_{j,n}\left(\frac{2\pi}{n}\right)^{|I_j|/2} \\ &\le 2e^{-(c-G-e-(1/2)\log 2\pi - C)|I_{\widetilde{J}_n(\beta)}|} \le 2e^{-bK_1 n^{1/(2\beta+1)}} \end{aligned}$$

for some $b > 0$ as soon as $c > G + e + \frac{1}{2}\log 2\pi$ provided we choose K_1 large enough. This proves (6.8). We are ready to complete the proof. For $A_n = \{\theta : (\sum_{j=0}^{\widetilde{J}_n(\beta)} \|\underline{\theta}_j - \underline{\theta}_j^0\|^2)^{1/2} \le Mu_n(\beta)/2\}$,

$$\begin{aligned} &P^\pi\big(\theta : \ell_2(\theta,\theta_0) > Mu_n(\beta)|Y^n\big) \\ &\quad \le P^\pi\big(\{\theta : \ell_2(\theta,\theta_0) > Mu_n(\beta)\} \cap \Theta_n(\beta)|Y^n\big) + P^\pi\big(\Theta_n(\beta)^c|Y^n\big) \\ &\quad \le P^\pi\big(A_n^c|Y^n\big) + P^\pi\big(\Theta_n(\beta)^c|Y^n\big). \end{aligned}$$

We bound the first term of the right-hand side by

$$\begin{aligned} P^\pi\big(A_n^c|Y^n\big) &\le \int_{A_n^c} \prod_{j=0}^{\widetilde{J}_n(\beta)} \frac{e^{-(n/2)\|\underline{\theta}_j - \underline{Y}_j\|^2}(1+\nu_{j,n})\, d\widetilde{\pi}_j(\underline{\theta}_j)}{\int_{\mathbb{R}^{|I_j|}} e^{-(n/2)\|\underline{\theta}_j - \underline{Y}_j\|^2} g_j(\underline{\theta}_j)\, d\underline{\theta}_j} \\ &\le \int_{A_n^c} \prod_{j=0}^{\widetilde{J}_n(\beta)} \nu_{j,n}^{-1} e^{G|I_j|} \frac{e^{-(n/2)\|\underline{\theta}_j - \underline{Y}_j\|^2}(1+\nu_{j,n})\, d\widetilde{\pi}_j(\underline{\theta}_j)}{\int_{[-L_0,L_0]^{|I_j|}} e^{-(n/2)\|\underline{\theta}_j - \underline{Y}_j\|^2} g_j(\underline{\theta}_j)\, d\underline{\theta}_j}. \end{aligned}$$

On B_n, with obvious notation,

$$\int_{[-L_0,L_0]^{|I_j|}} e^{-(n/2)\|\underline{\theta}_j-\underline{Y}_j\|^2} d\underline{\theta}_j \geq \left(\frac{2\pi}{n}\right)^{|I_j|/2} - \Pr\bigl(\exists j \leq \widetilde{J}_n(\beta), k \in I_j : |Y_{j,k}| + n^{-1/2}\bigl|\mathcal{N}(0,1)\bigr| > L_0\bigr).$$

Since $|\theta^0_{j,k}| \leq L$ for all $Y^n \in B_n$, we find $|Y_{j,k}| + n^{-1/2}|\mathcal{N}(0,1)| \leq L + \frac{1}{2} + n^{-1/2}|\mathcal{N}(0,1)|$, and hence

$$\Pr\bigl(\exists j \leq \widetilde{J}_n(\beta), k \in I_j : |Y_{j,k}| + n^{-1/2}\bigl|\mathcal{N}(0,1)\bigr| > L_0\bigr) \leq n \Pr\biggl(\bigl|\mathcal{N}(0,1)\bigr| > \frac{\sqrt{n}}{2}\biggr) \leq 2ne^{-n/8}.$$

It follows that for $Y^n \in B_n$,

$$\int_{[-L_0,L_0]^{|I_j|}} e^{-(n/2)\|\underline{\theta}_j-\underline{Y}_j\|^2} d\underline{\theta}_j \geq \left(\frac{2\pi}{n}\right)^{|I_j|/2} - 2ne^{-n/8} \geq \frac{1}{2}\left(\frac{2\pi}{n}\right)^{|I_j|/2}.$$

We now study the numerator in $P^\pi(A_n^c|Y^n)$. For $Y^n \in B_n$

$$-\|\underline{\theta}_j - \underline{Y}_j\|^2 \leq \|\underline{\theta}_j^0 - \underline{Y}_j\|^2 - \frac{1}{2}\|\underline{\theta}_j - \underline{\theta}_j^0\|^2 \leq \frac{e}{n}|I_j| - \frac{1}{2}\|\underline{\theta}_j - \underline{\theta}_j^0\|^2.$$

On B_n, we can subsequently bound $P^\pi(A_n^c|Y^n)$ by $2e^{\sum_{j=0}^{\widetilde{J}_n(\beta)}(G+e/2)|I_j|}$ times

$$\begin{aligned}
&\int_{A_n^c} \prod_{j=0}^{\widetilde{J}_n(\beta)} \left(\frac{n}{2\pi}\right)^{|I_j|/2} \nu_{j,n}^{-1} e^{-(n/4)\|\underline{\theta}_j-\underline{\theta}_j^0\|^2}(1+\nu_{j,n})\, d\widetilde{\pi}_j(\underline{\theta}_j) \\
&\quad = e^{-(nM^2 u_n(\beta)^2)/32} \\
&\qquad \times \prod_{j=0}^{\widetilde{J}_n(\beta)} \int_{\mathbb{R}^{|I_j|}} \left(\frac{n}{2\pi}\right)^{|I_j|/2} \nu_{j,n}^{-1} e^{-(n/8)\|\underline{\theta}_j-\underline{\theta}_j^0\|^2}(1+\nu_{j,n})\, d\widetilde{\pi}_j(\underline{\theta}_j) \\
&\quad \leq e^{-(nM^2 u_n(\beta)^2)/32} \prod_{j=0}^{\widetilde{J}_n(\beta)} \left(\nu_{j,n}^{-1}\left(\frac{n}{2\pi}\right)^{|I_j|/2} + 2^{|I_j|}e^{G|I_j|}\right) \\
&\quad \leq e^{-(nM^2 u_n(\beta)^2)/32} \prod_{j=0}^{\widetilde{J}_n(\beta)} \bigl(e^{c|I_j|} + 2^{|I_j|}e^{G|I_j|}\bigr).
\end{aligned}$$

Choosing M large enough and using the exponential bound on $P^n_{\theta_0}(B_n^c)$ shows that $E^n_{\theta_0}[P^\pi(A_n^c|Y^n)] \leq e^{-An^{1/(2\beta+1)}}$. This completes the proof of Theorem 3.2.

APPENDIX: ADDITIONAL PROOFS

A.1. Explicit bounds on Ω_n.

PROOF OF (2.3). Since we are on a Hölder space, we can prove the result for $\ell = \ell_\infty$ (see also Section 3). Consider $\theta = (\theta_{j,k})_{(j,k)\in\Lambda} \in \mathcal{H}(\beta, L)$ and pick $J_n(\beta)$ such that

$$\tfrac{1}{2}(L/2M)^{1/\beta}(n/\log n)^{1/(2\beta+1)} \le 2^{J_n(\beta)} \le (L/2M)^{1/\beta}(n/\log n)^{1/(2\beta+1)}.$$

On resolution level $J_n(\beta)$ chose an arbitrary index in Λ, $(J_n(\beta), k^*)$ say. By definition of $\mathcal{H}(\beta, L)$ there exists $\theta' \in \mathcal{H}(\beta, L)$, with $|\theta'_{j,k} - \theta_{j,k}|$ equals $L2^{-J_n(\beta)(\beta+1/2)}$ if $(j, k) = (J_n(\beta), k^*)$ and zero otherwise. Then $\ell_\infty(\theta, \theta') = L2^{-J_n(\beta)\beta} \ge 2\epsilon_n(\theta)$ and $\|\theta - \theta'\|_{L^2} = L2^{-J_n(\beta)(\beta+1/2)} \lesssim \sqrt{\log n/n}$. □

A.2. Derivation of (5.6). We briefly recall the main arguments of Ghosal, Ghosh and van der Vaart [15] leading to inequality (5.6). Their method is based on two assumptions, namely a bound on the local entropy as well as existence of a decomposition $\Theta = \Theta_n \cup (\Theta \setminus \Theta_n)$ such that the prior is uniform on Θ_n (with respect to Kullback–Leibler balls) and assigns negligible mass to $\Theta \setminus \Theta_n$ (cf. [15], equations (2.7), (2.3) and (2.5)). To derive (5.6) only the assumption on the prior needs to be imposed. Recall that

$$P^\pi\big(\theta : \ell(\theta_0, \theta) > \epsilon_n | Y^n\big) = \frac{\int_{\ell(\theta_0,\theta)>\epsilon_n} e^{\mathcal{L}_n(\theta)-\mathcal{L}_n(\theta_0)}\pi(d\theta)}{\int_\Theta e^{\mathcal{L}_n(\theta)-\mathcal{L}_n(\theta_0)}\pi(d\theta)} =: \frac{N_n}{D_n}.$$

Under the imposed conditions, there are constants $c, c' > 0$ such that $P^n_{\theta_0}(D_n \ge \exp(-cnu_n^2)) \ge 1 - e^{-c'nu_n^2}$ (cf. [15], Lemma 8.4). Hence, for any test function Φ_n,

$$\begin{aligned} &E^n_{\theta_0}\big[P^\pi\big(\theta : \ell(\theta_0, \theta) > \epsilon_n | Y^n\big)\big] \\ &\quad \le E^n_{\theta_0}[\Phi_n] + e^{cnu_n^2} E^n_{\theta_0}\bigg[\int_{\ell(\theta_0,\theta)>\epsilon_n} e^{\mathcal{L}_n(\theta)-\mathcal{L}_n(\theta_0)}(1-\Phi_n)\pi(d\theta)\bigg] + e^{-c'nu_n^2} \\ &\quad \le E^n_{\theta_0}[\Phi_n] + e^{cnu_n^2} \sup_{\theta:\ell(\theta_0,\theta)>\epsilon_n} E^n_\theta[1-\Phi_n] + e^{-c'nu_n^2}. \end{aligned}$$

A.3. Proof of Lemma 1.

PROOF OF (6.2). We have

$$\begin{aligned} P^\pi\big(S^c \cap \mathcal{J}_n(\overline{\gamma}) \ne \varnothing | Y^n\big) &\le \sum_{(j,k)\in\mathcal{J}_n(\overline{\gamma})} P^\pi\big(\theta_{j,k} = 0 | Y^n\big) \\ &\le \sum_{(j,k)\in\mathcal{J}_n(\overline{\gamma})} \frac{e^{-(n/2)Y_{j,k}^2}}{w_{j,n}\int_{\mathbb{R}} e^{-(n/2)(Y_{j,k}-\theta)^2} g(\theta)\,d\theta}. \end{aligned}$$

Recall the definition of $\Omega_{n,B}$ in the proof of Theorem 3.1. If $Y^n \in \Omega_{n,B}$ and $\overline{\gamma}$ is large enough, then $|Y_{j,k}| > \frac{1}{2}\overline{\gamma}\sqrt{\log n/n}$. With the same argument as in (6.6),

$$P^\pi\big(S^c \cap \mathcal{J}_n(\overline{\gamma}) \neq \varnothing | Y^n\big) \leq \sum_{j \leq J_n, k \in I_j} \frac{e^{-(nY_{j,k}^2)/2}\sqrt{n}}{w_{j,n} a\sqrt{\pi}} \leq \frac{n^{K+3/2-\overline{\gamma}^2/8}}{a\sqrt{\pi}},$$

and together with (6.5) this completes the proof of (6.2), provided $\overline{\gamma}$ is sufficiently large. □

PROOF OF (6.3). We have

$$\begin{aligned}
P^\pi\big(S \cap \mathcal{J}_n(\underline{\gamma})^c \neq \varnothing | Y^n\big) &= \sum_{(j,k) \in \mathcal{J}_n(\underline{\gamma})^c} P^\pi\big(\theta_{j,k} \neq 0 | Y^n\big) \\
&\leq \sum_{(j,k) \in \mathcal{J}_n(\underline{\gamma})^c} \frac{w_{j,n} \int_{\mathbb{R}} e^{-(n/2)(\theta - Y_{j,k})^2} g(\theta)\, d\theta}{(1 - w_{j,n}) e^{-(n/2)Y_{j,k}^2}} \\
&\leq 2\sqrt{2\pi} n^{-1/2} \sup_x g(x) \sum_{(j,k) \in \mathcal{J}_n(\underline{\gamma})^c} w_{j,n} e^{(nY_{j,k}^2)/2}.
\end{aligned}$$

If $Y^n \in \Omega_{n,B}$, for any $(j,k) \in \mathcal{J}_n(\underline{\gamma})^c$,

$$\begin{aligned}
nY_{j,k}^2 &\leq \underline{\gamma}^2 \log n + 2\log|I_j| + 2B\log n + 2\underline{\gamma}\sqrt{\log n}\sqrt{2\log|I_j| + 2B\log n} \\
&\leq 2\log|I_j| + \big(2B + \underline{\gamma}^2 + \underline{\gamma}C\big)\log n,
\end{aligned}$$

for some constant C. Hence, whenever $Y^n \in \Omega_{n,B}$, using that $w_{j,n} \leq n^{(\tau\wedge 1)/2-1/4-2B} 2^{-j(1+\tau)}$ with $\tau > 1/2$,

$$\begin{aligned}
P^\pi\big(S \cap \mathcal{J}_n(\underline{\gamma})^c \neq \varnothing | Y^n\big) &\lesssim n^{-3/4+(\tau\wedge 1)/2-B+\underline{\gamma}C/2+\underline{\gamma}^2/2} \sum_{j \leq J_n} |I_j|^2 2^{-j(1+\tau)} \\
&\lesssim n^{1/4-(\tau\wedge 1)/2-B+\underline{\gamma}C/2+\underline{\gamma}^2/2} = O\big(n^{-B}\big),
\end{aligned}$$

where for the last equality, we need that $\underline{\gamma}$ is sufficiently small. The proof of (6.3) follows from (6.5). □

A.4. Proof of Proposition 1. We start with verifying condition (4.3). For $r \geq 1$, there exists an injective mapping $\psi : \mathcal{I}_r \to \mathcal{I}_0$ such that

$$\psi(\theta)_{\mathcal{U}} = \theta^*_{\mathcal{U}} \quad \text{and} \quad \big|\psi(\theta)_{j,k} - \theta_{j,k}\big| \neq \phi_n \qquad \forall (j,k) \notin \mathcal{U}.$$

This implies in particular that $|\mathcal{J}_r| \leq |\mathcal{I}_r| \leq |\mathcal{I}_0| \leq |\mathcal{J}_0|$ and the partition is admissible. Also, by construction of $\mathcal{D}_n$, for every $\theta \in \mathcal{J}_r$ we have that $\ell_2(\theta, \psi(\theta))^2$ takes its values in the lattice $\{\phi_n^2, 2\phi_n^2, 3\phi_n^2, \ldots\}$ and the cardinality of $\{r : u_r^2 = \phi_n^2\}$ is bounded by $2\sum_{j \leq J_n} |I_j| = I$. By induction on $M = 1, 2, \ldots,$ the cardinality of

$\{r : u_r^2 = M\phi_n^2\}$ is further bounded by $\sum_{i=1}^M I^i \le (Cn)^{M+1}$ for some $C > 0$. This implies that for any $K_0 > 0$,

$$\sum_{r=1}^{R_n} e^{-K_0 n u_r^2} \le \sum_{M\ge 1} e^{-K_0 n M\phi_n^2} \big|\{r : u_r^2 = M\phi_n^2\}\big| \le \sum_{M\ge 1} n^{-K_0\phi_0 M}(Cn)^{M+1},$$

which has polynomial decay in n as soon as $\phi_0 > 2/K_0$, and can thus be taken of the form $e^{-K_1 n\Omega(\epsilon_n(\cdot),\mathcal{H}(\beta,L),\ell_\infty)^2}$ for some $K_1 > 0$. This bound is not based on any specific assumption on the experiment $\{P_\theta^n, \theta \in \Theta\}$ and only depends on the set Θ, the loss $\ell = \ell_\infty$, and $d = \ell_2$. It remains to check condition (4.2). We first consider the white noise model. Then

$$\begin{aligned}
&-n^{-1}\big(\mathcal{L}_n(\theta) - \mathcal{L}_n\big(\psi(\theta)\big)\big)\\
&\quad= \frac{\|\theta-\theta_0\|_{L^2}^2 - \|\psi(\theta)-\theta_0\|_{L^2}^2}{2} - \sum_{(j,k)\in\Lambda} \big(Y_{j,k} - \theta_{j,k}^0\big)\big(\theta_{j,k} - \psi(\theta)_{j,k}\big)\\
&\quad= \frac{\|\theta-\psi(\theta)\|_{L^2}^2}{2} + \big\langle \theta - \psi(\theta), \psi(\theta) - \theta_0\big\rangle_{L^2}\\
&\qquad - \sum_{(j,k)\in\Lambda} \big(Y_{j,k} - \theta_{j,k}^0\big)\big(\theta_{j,k} - \psi(\theta)_{j,k}\big).
\end{aligned}$$

The above computation is simply a sequential formulation of the Cameron–Martin formula: Here, we emphasise on the property that $\ell_2(\theta,\theta') = \|\theta - \theta'\|_{L^2}$ is a Hilbert norm associated to the scalar product $\langle\cdot,\cdot\rangle_{L^2}$. The sum in $(j,k) \in \Lambda$ involving $Y_{j,k}$ has to be understood as a limit in $L^2(P_\theta^n)$, and it is well defined since $\theta - \psi(\theta) \in \ell^2(\Lambda)$ and the $Y_{j,k}$ are independent and standard normal under P_θ^n.

Recall the definition of $\mathcal{U}$ in (4.7). For $(j,k) \in \mathcal{U}$, we have by construction $|\theta_{j,k}^0 - \psi(\theta)_{j,k}| \le \phi_n/4$ and for $(j,k) \in \mathcal{U}^c$, $|\theta_{j,k}^0 - \psi(\theta)_{j,k}| \le 3\phi_n/4$. In the latter case, we also know that $|\theta_{j,k} - \psi(\theta)_{j,k}| \ne \phi_n$ but has values in $\{0, 2\phi_n, 3\phi_n, \dots\}$. Therefore,

$$\mathcal{L}_n(\theta) - \mathcal{L}_n\big(\psi(\theta)\big) \le -\frac{n\|\theta-\psi(\theta)\|_{L^2}^2}{8} + n\sum_{(j,k)\in\Lambda} \big(Y_{j,k} - \theta_{j,k}^0\big)\big(\theta_{j,k} - \psi(\theta)_{j,k}\big).$$

Introduce the event $\Omega_n = \{\max_{j\le J_n, k\in I_j} |Y_{j,k} - \theta_{j,k}^0|\sqrt{n} \le 2\sqrt{\log n}\}$. For $Y^n \in \Omega_n$,

$$\begin{aligned}
&\Big|\sum_{(j,k)\in\Lambda} \big(Y_{j,k} - \theta_{j,k}^0\big)\big(\theta_{j,k} - \psi(\theta)_{j,k}\big)\Big|\\
&\qquad= \Big|\sum_{(j,k)\in\Lambda} \mathbf{1}_{|\theta_{j,k}-\psi(\theta)_{j,k}|\ge\phi_n}\big(Y_{j,k} - \theta_{j,k}^0\big)\big(\theta_{j,k} - \psi(\theta)_{j,k}\big)\Big|\\
&\qquad\le 2\phi_0^{-1}\|\theta - \psi(\theta)\|_{L^2}^2
\end{aligned}$$

due to $|Y_{j,k} - \theta^0_{j,k}| \le 2\sqrt{\log n}/\sqrt{n} \le 2/\phi_0\phi_n$. Picking ϕ_0 large enough,

$$\mathcal{L}_n(\theta) - \mathcal{L}_n(\psi(\theta)) \le -\frac{n\|\theta - \psi(\theta)\|^2_{L^2}}{16} \qquad \text{on } \Omega_n.$$

Since $P^n_{\theta_0}(\Omega^c_n) \le 2n^{-1}$, condition (4.2) is satisfied. This completes the proof of Proposition 1.

A.5. Proof of Proposition 3. We start with the construction of the prior π_n. Contrariwise to the white noise model, we truncate $j \le \overline{J}_n$ with $\sqrt{n}/\log n < 2^{\overline{J}_n} \le 2\sqrt{n}/\log n$. Set $\Theta = \bigcup_{\beta\in[\beta_1,\beta_2]} \mathcal{H}'(\beta, L)$. Recall the definition of $\mathcal{D}_n$ in (4.5) with $\phi_n = \phi_0\sqrt{\log n/n}$ and consider

$$\mathcal{D}'_n = \{\theta \in \mathcal{D}_n : \exists\theta' \in \Theta \text{ such that } |\theta_{j,k} - \theta'_{j,k}| \le \phi_n, \forall j \le \overline{J}_n, k \in I_j\}$$

as set of nonnormalised test densities. By construction $\sum_{j,k} \theta_{j,k}\Psi_{j,k} \ge c/2$, $\forall\theta \in \mathcal{D}'_n$ and, therefore, $\sqrt{f_\theta} = \|\theta\|^{-1}_{L^2} \sum_{j,k} \theta_{j,k}\Psi_{j,k}$ is well-defined, that is, f_θ is a density [note that this definition extends (4.9) in a consistent way]. The set $\mathcal{D}'_n$ constitutes the sieve and the prior is given by $\pi_n \propto \sum_{\theta\in\mathcal{D}'_n} \delta_{\theta/\|\theta\|_{L^2}}$.

For the subsequent analysis, we need some inequalities for the elements in $\mathcal{D}'_n$, which are derived next. Due to $\beta_1 > 1/2$, the coefficients of the parameter vectors are absolutely summable and

$$\overline{L} = \max\Big(\sup_{\theta\in\Theta\cup\mathcal{D}'_n} \ell_\infty(\theta/\|\theta\|_{L^2}, 0) + \sum_{(j,k)\in\Lambda} |\theta_{j,k}|, 1\Big) < \infty.$$

Let $\theta \in \mathcal{D}'_n$. By construction, there exists a $\theta' \in \Theta$ with $\|\theta'\|_{L^2} = 1$ and $|\theta_{j,k} - \theta'_{j,k}| \le \phi_n$ for all $(j,k) \in \Lambda$. With $\|\theta\|^2_{L^2} = \langle\theta + \theta', \theta - \theta'\rangle_{L^2} + 1$ we find

$$\text{(A.9)} \quad \tfrac{1}{2} \le \|\theta\|_{L^2} \le 2, \qquad \big|\|\theta\|_{L^2} - 1\big| \le 4\overline{L}\phi_n \quad \text{and} \quad \left|\frac{1}{\|\theta\|_{L^2}} - 1\right| \le 8\overline{L}\phi_n.$$

Next, let us construct an admissible partition. Notice that there is a finite J_0, such that

$$\text{(A.10)} \qquad \sup_{\theta\in\Theta\cup\mathcal{D}'_n} \max_{j>J_0, k\in I_j} |\theta_{j,k}| + \sum_{j>J_0}^{\infty} \sum_{k\in I_j} \theta^2_{j,k} < 2^{-21}\frac{1}{\overline{L}^3}.$$

Let $Q = \lceil\overline{L}^2 2^{11}\rceil$. For every (j,k), we can define an equivalence relation $\simeq$ via $\theta_{j,k} \simeq \theta'_{j,k}$ iff $\theta_{j,k} = \theta'_{j,k}$ or $\theta_{j,k}, \theta'_{j,k} \in (\theta^0_{j,k} - q_{j,k}(\theta_0)\phi_n, \theta^0_{j,k} + q_{j,k}(\theta_0)\phi_n)$, where

$$q_{j,k}(\theta_0) = \begin{cases} Q, & \text{if } j \le J_0, \\ 1, & \text{if } j > J_0, |\theta^0_{j,k}| > 2^{-9}\phi_n, \\ 0, & \text{if } j > J_0, |\theta^0_{j,k}| \le 2^{-9}\phi_n. \end{cases}$$

This induces an equivalence relation on the nonnormalised densities $\mathcal{D}'_n$ via $\theta \simeq \theta'$ iff $\theta_{j,k} \simeq \theta'_{j,k}$ for all (j,k), $j \leq \overline{J}_n$. By construction, there exists $\theta^* \in \mathcal{D}'_n$ such that $|\theta^0_{j,k} - \theta^*_{j,k}| \leq \frac{1}{2}\phi_n$ for all (j,k). Denote by $\mathcal{I}_r$, $r = 0, 1, \dots$ the equivalence classes of $\mathcal{D}'_n$ and let $\mathcal{I}_0$ be the equivalence class of θ^*. Define $J_n(\beta)$ as in (4.6), replacing $\phi_n/4$ by $2^{-9}\phi_n$ in the first condition. Using (A.9), there exists a constant $A = A(\beta, \overline{L}, \phi_0, Q)$ such that, for all $\theta \in \mathcal{I}_0$,

$$\begin{aligned}\ell_\infty\big(\theta_0, \theta/\|\theta\|_{L^2}\big) &\leq \ell_\infty(\theta_0, \theta) + 4\overline{L}\phi_n \ell_\infty\big(\theta/\|\theta\|_{L^2}, 0\big)\\ &\leq 2Q\phi_n \sum_{j=0}^{J_n(\beta)} 2^{j/2} + \epsilon_n(\beta) + 4\overline{L}^2\phi_n\\ &\leq A\epsilon_n(\beta).\end{aligned}$$

For this A, we define $\mathcal{J}_0 = \{\theta \in \mathcal{D}'_n : \ell_\infty(\theta_0, \theta/\|\theta\|_{L^2}) \leq A\epsilon_n(\beta)\}$, $\mathcal{J}_r = \mathcal{I}_r \cap \mathcal{J}_0^c$. Now, for any $r \geq 1$, we construct an injective map $\psi : \mathcal{J}_r \to \mathcal{J}_0$ and verify that for this map the properties (4.2) and (4.3) hold. To this end, define $\iota(\theta_{j,k})$ as $\lceil \theta^0_{j,k}\phi_n^{-1}\rceil\phi_n$ if $\theta_{j,k} > \theta^0_{j,k}$ and $\lfloor \theta^0_{j,k}\phi_n^{-1}\rfloor\phi_n$ otherwise. If $(j,k) \in \mathcal{J}_r$, $r \neq 0$,

$$\psi(\theta)_{j,k} = \begin{cases} \theta_{j,k}, & \text{if } |\theta_{j,k} - \theta^0_{j,k}| < q_{j,k}(\theta_0)\phi_n,\\ \iota(\theta_{j,k}), & \text{if } |\theta_{j,k} - \theta^0_{j,k}| \geq q_{j,k}(\theta_0)\phi_n, q_{j,k}(\theta_0) > 0,\\ 0, & \text{if } q_{j,k}(\theta_0) = 0.\end{cases}$$

It is not difficult to see that $\psi : \mathcal{J}_r \to \mathcal{J}_0$ is injective. This completes the proof of the admissible part. By following the same arguments as in the proof of Proposition 1, condition (4.3) can be verified.

Therefore, it remains to check (4.2). For $u > 0$, we have $\log u = 2\log(\sqrt{u}) \leq 2(\sqrt{u} - 1)$ and, therefore,

$$\mathcal{L}_n(\theta) - \mathcal{L}_n\big(\psi(\theta)\big) \leq 2\sum_{i=1}^n \bigg(\frac{\sqrt{f_\theta}(Y_i)}{\sqrt{f_{\psi(\theta)}}(Y_i)} - 1\bigg). \tag{A.11}$$

We further decompose the right-hand side using

$$\begin{aligned}\frac{x}{y} - 1 = \frac{x-y}{z} &+ \frac{(x-y)(z-y)}{z^2}\\ &+ \frac{(x-y)(z-y)^2}{z^3} + \frac{(x-y)(z-y)^3}{z^3 y}\end{aligned} \tag{A.12}$$

with $x = \sqrt{f_\theta}(Y_i)$, $y = \sqrt{f_{\psi(\theta)}}(Y_i)$, and $z = \sqrt{f_{\theta_0}}(Y_i)$. In the sequel, we control the large deviations behaviour of the terms on the right-hand side separately [denoting the single steps by (I)–(IV)]. The key ingredient is the following well-known version of Bernstein's inequality: If $X_1, \dots, X_n$ are i.i.d., centered and $|X_i| \leq M$, then $\forall t > 0$, $\mathbb{P}(|\sum_{i=1}^n X_i| > t) \leq 2\exp(-\frac{1}{2}t^2/(n\mathbb{E}[X_1^2] + Mt/3))$.

(I) Define $\Omega_{n,1}(\tau)$ as the event

$$\left\{\left|\sum_{i=1}^{n} \frac{\Psi_{j,k}(Y_i)}{\sqrt{f_{\theta_0}}(Y_i)} - n\int \Psi_{j,k}(u)\sqrt{f_{\theta_0}}(u)\,du\right| \le \tau\sqrt{n\log n}, \forall j \le \overline{J}_n, k \in I_j\right\}.$$

Observe that the random variables $\Psi_{j,k}(Y_i)/\sqrt{f_{\theta_0}}(Y_i)$, $i = 1, \dots, n$, are i.i.d., bounded in absolute value by a multiple of $n^{1/4}$ and their second moment is one. Thus, by a union bound and Bernstein's inequality, $P_{\theta_0}^n(\Omega_{n,1}(\tau)^c) \lesssim n^{-1}$, provided that τ is large enough. On $Y^n \in \Omega_{n,1}(\tau)$,

$$\begin{aligned}\sum_{i=1}^{n} \frac{\sqrt{f_\theta}(Y_i) - \sqrt{f_{\psi(\theta)}}(Y_i)}{\sqrt{f_{\theta_0}}(Y_i)} &\le n\int \big(\sqrt{f_\theta}(u) - \sqrt{f_{\psi(\theta)}}(u)\big)\sqrt{f_{\theta_0}}(u)\,du \\ &\quad + \tau\sqrt{n\log n}\sum_{(j,k)}\left|\frac{\theta_{j,k}}{\|\theta\|_{L^2}} - \frac{\psi(\theta)_{j,k}}{\|\psi(\theta)\|_{L^2}}\right|.\end{aligned}$$

Using the inequalities (A.9), we can bound the second term on the right-hand side by $\tau\phi_0^{-1}(1 + 16\overline{L}^2)n\|\theta - \psi(\theta)\|_{L^2}^2$, and thus, making ϕ_0 large enough we obtain on $Y^n \in \Omega_{n,1}(\tau)$,

$$\begin{aligned}&\sum_{i=1}^{n} \frac{\sqrt{f_\theta}(Y_i) - \sqrt{f_{\psi(\theta)}}(Y_i)}{\sqrt{f_{\theta_0}}(Y_i)} \\ &\quad \le n\int \big(\sqrt{f_\theta}(u) - \sqrt{f_{\psi(\theta)}}(u)\big)\sqrt{f_{\theta_0}}(u)\,du + 2^{-9}n\|\theta - \psi(\theta)\|_{L^2}^2.\end{aligned}$$

(II) Similar as for (I), set $\Omega_{n,2}$ for the event

$$\left\{\left|\sum_{i=1}^{n} \frac{\Psi_{j,k}(Y_i)\Psi_{j',k'}(Y_i)}{f_{\theta_0}(Y_i)} - n\delta_{(j,k),(j',k')}\right| \le n^{3/4}\log n, \quad \forall j, j' \le \overline{J}_n, k \in I_j, k' \in I_{j'}\right\},$$

with $\delta_{(j,k),(j',k')}$ the Kronecker delta. Now, $\Psi_{j,k}(Y_i)\Psi_{j',k'}(Y_i)/f_{\theta_0}(Y_i)$, $i = 1, \dots, n$, are i.i.d. and bounded in absolute value by a multiple of $\sqrt{n}$. The second moment is also smaller than const. $\times\, n^{1/2}$. Using a union bound and Bernstein's inequality, $P_{\theta_0}^n(\Omega_{n,2}^c) \lesssim n^{-1}$ for n large enough. On $Y^n \in \Omega_{n,2}$, we see that $\sum_{i=1}^n(\sqrt{f_\theta}(Y_i) - \sqrt{f_{\psi(\theta)}}(Y_i))(\sqrt{f_{\theta_0}}(Y_i) - \sqrt{f_{\psi(\theta)}}(Y_i))/f_{\theta_0}(Y_i)$ can be bounded by its expectation plus

$$\sum_{j,k}\left|\frac{\theta_{j,k}}{\|\theta\|_{L^2}} - \frac{\psi(\theta)_{j,k}}{\|\psi(\theta)\|_{L^2}}\right| \sum_{j',k':\Psi_{j,k}\Psi_{j',k'}\neq 0}\left|\theta_{j,k}^0 - \frac{\psi(\theta)_{j,k}}{\|\psi(\theta)\|_{L^2}}\right| n^{3/4}\log n.$$

Due to the compact support of Ψ, there are of the order of $\log n$ index pairs (j', k') with $j' \le \overline{J}_n$ and $\Psi_{j,k}\Psi_{j',k'} \neq 0$. Using that $\psi(\theta) \in \mathcal{J}_0$, together with the inequalities (A.9), yields $|\theta^0_{j,k} - \psi(\theta)_{j,k}/\|\psi(\theta)\|_{L^2}| \le (Q + 4\overline{L})\phi_n$. Because of $|\theta_{j,k} - \psi(\theta)_{j,k}|\mathbf{1}_{\theta_{j,k}\neq\psi(\theta)_{j,k}} \ge \phi_n$, the expression in the last display is smaller than $2^{-9}n\|\theta - \psi(\theta)\|^2_{L^2}$, for sufficiently large n and, therefore, on $Y^n \in \Omega_{n,2}$,

$$\sum_{i=1}^n \frac{(\sqrt{f_\theta}(Y_i) - \sqrt{f_{\psi(\theta)}}(Y_i))(\sqrt{f_{\theta_0}}(Y_i) - \sqrt{f_{\psi(\theta)}}(Y_i))}{f_{\theta_0}(Y_i)}$$
$$\le n\int \big(\sqrt{f_\theta}(u) - \sqrt{f_{\psi(\theta)}}(u)\big)$$
$$\times \big(\sqrt{f_{\theta_0}}(u) - \sqrt{f_{\psi(\theta)}}(u)\big)\,du + 2^{-9}n\|\theta - \psi(\theta)\|^2_{L^2}.$$

(III) This case works similar as (II) and is therefore only sketched here. In fact, we need to consider $\Omega_{n,3}$ which is the same event as $\Omega_{n,2}$ but applied to the random variables $\Psi_{j_1,k_1}(Y_i)\Psi_{j_2,k_2}(Y_i)\Psi_{j_3,k_3}(Y_i)/f^{3/2}_{\theta_0}(Y_i)$ (and the $n^{3/4}$ should be exchanged with n). Since these random variables are bounded in absolute value by a constant times $n^{3/4}$ and have second moment smaller than a constant times n, we obtain $P^n_{\theta_0}(\Omega^c_{n,3}) \lesssim n^{-1}$. Using the inequalities (A.9) again, on $Y^n \in \Omega_{n,3}$,

$$\sum_{i=1}^n \frac{(\sqrt{f_\theta}(Y_i) - \sqrt{f_{\psi(\theta)}}(Y_i))(\sqrt{f_{\theta_0}}(Y_i) - \sqrt{f_{\psi(\theta)}}(Y_i))^2}{f^{3/2}_{\theta_0}(Y_i)}$$
$$\le n\int \big(\sqrt{f_\theta}(u) - \sqrt{f_{\psi(\theta)}}(u)\big)\big(\sqrt{f_{\theta_0}}(u) - \sqrt{f_{\psi(\theta)}}(u)\big)^2 \frac{du}{\sqrt{f_{\theta_0}}(u)}$$
$$+ 2^{-10}n\|\theta - \psi(\theta)\|^2_{L^2}.$$

The first term on the right-hand side can be further bounded by a constant times

$$n\sum_{j,k}\left|\frac{\theta_{j,k}}{\|\theta\|_{L^2}} - \frac{\psi(\theta)_{j,k}}{\|\psi(\theta)\|_{L^2}}\right|\int |\Psi_{j,k}(u)|\big(\sqrt{f_{\theta_0}}(u) - \sqrt{f_{\psi(\theta)}}(u)\big)^2\,du.$$

Expanding $\sqrt{f_{\theta_0}}(u) - \sqrt{f_{\psi(\theta)}}(u)$ and using the compactness of Ψ as well as $|\theta^0_{j,k} - \psi(\theta)_{j,k}/\|\psi(\theta)\|_{L^2}| \le (Q + 4\overline{L})\phi_n$ and (A.9), we find that the last display can be further bounded by $O(n\|\theta - \psi(\theta)\|^2_{L^2}\phi_n \log n)$, and so, on $Y^n \in \Omega_{n,3}$,

$$\sum_{i=1}^n \frac{(\sqrt{f_\theta}(Y_i) - \sqrt{f_{\psi(\theta)}}(Y_i))(\sqrt{f_{\theta_0}}(Y_i) - \sqrt{f_{\psi(\theta)}}(Y_i))^2}{f^{3/2}_{\theta_0}(Y_i)} \le 2^{-9}n\|\theta - \psi(\theta)\|^2_{L^2}.$$

(IV) For this term, no exponential inequality is needed, and a deterministic bound can be obtained as follows. Observe that there is a constant $c(\Psi)$, such that $|\sqrt{f_{\theta_0}}(Y_i) - \sqrt{f_{\psi(\theta)}}(Y_i)| \le c(\Psi)Q\phi_n\sum_{j=0}^{\overline{J}_n} 2^{j/2} \lesssim 2^{\overline{J}_n/2}\phi_n$. This shows that

$$\sum_{i=1}^n \big(\sqrt{f_\theta}(Y_i) - \sqrt{f_{\psi(\theta)}}(Y_i)\big)\big(\sqrt{f_{\theta_0}}(Y_i) - \sqrt{f_{\psi(\theta)}}(Y_i)\big)^3/\big(f^{3/2}_{\theta_0}(Y_i)\sqrt{f_{\psi(\theta)}}(Y_i)\big)$$

can be bounded by a constant times

$$n2^{2\overline{J}_n}\phi_n^3 \sum_{j,k}\left|\frac{\theta_{j,k}}{\|\theta\|_{L^2}} - \frac{\psi(\theta)_{j,k}}{\|\psi(\theta)\|_{L^2}}\right|.$$

Using the definition of $\overline{J}_n$ and (A.9), we find that this term is of negligible order $O(n\|\theta - \psi(\theta)\|_2^2/\log n)$, uniformly over θ.

Now, we are ready to complete the proof. Since $P_{\theta_0}^n((\Omega_{n,1}(\tau)\cap\Omega_{n,2}\cap\Omega_{n,3})^c) \lesssim n^{-1}$, we can throughout the following assume that $Y^n \in \Omega_{n,1}(\tau)\cap\Omega_{n,2}\cap\Omega_{n,3}$. It is then enough to prove $\frac{1}{n}(\mathcal{L}_n(\theta) - \mathcal{L}_n(\psi(\theta))) \le -K_0\|\theta - \psi(\theta)\|_{L^2}^2$ for some positive constant K_0. Combining the estimates in (I)–(IV), with (A.11) and (A.12), we find, for sufficiently large n,

$$\begin{aligned}
&\frac{1}{n}\big(\mathcal{L}_n(\theta) - \mathcal{L}_n\big(\psi(\theta)\big)\big)\\
&\quad\le \int \big(\sqrt{f_\theta}(u) - \sqrt{f_{\psi(\theta)}}(u)\big)\big(2\sqrt{f_{\theta_0}}(u) - \sqrt{f_{\psi(\theta)}}(u)\big)\,du + 2^{-7}\|\theta - \psi(\theta)\|_{L^2}^2\\
&\quad= \left(\frac{1}{2} - \|\psi(\theta)\|_{L^2}\right)\sum_{j,k}\left(\frac{\theta_{j,k}}{\|\theta\|_{L^2}} - \frac{\psi(\theta)_{j,k}}{\|\psi(\theta)\|_{L^2}}\right)^2\\
&\qquad + 2\sum_{j,k}\left(\frac{\theta_{j,k}}{\|\theta\|_{L^2}} - \frac{\psi(\theta)_{j,k}}{\|\psi(\theta)\|_{L^2}}\right)\big(\theta_{j,k}^0 - \psi(\theta)_{j,k}\big) + 2^{-7}\|\theta - \psi(\theta)\|_{L^2}^2
\end{aligned}$$

using that

$$\int \big(\sqrt{f_\theta}(u) - \sqrt{f_{\psi(\theta)}}(u)\big)\sqrt{f_{\psi(\theta)}}(u)\,du = -\frac{1}{2}\int \big(\sqrt{f_\theta}(u) - \sqrt{f_{\psi(\theta)}}(u)\big)^2\,du.$$

If $|\theta_{j,k}^0| > 2^{-9}\phi_n$, then by construction of $\psi(\theta)_{j,k}$, we have

$$\big(\theta_{j,k} - \psi(\theta)_{j,k}\big)\big(\theta_{j,k}^0 - \psi(\theta)_{j,k}\big) \le 0.$$

Otherwise, if $|\theta_{j,k}^0| \le 2^{-9}\phi_n$, then $\psi(\theta)_{j,k} = 0$ and so

$$\frac{2}{\|\psi(\theta)\|_{L^2}}\sum_{j,k}\big(\theta_{j,k} - \psi(\theta)_{j,k}\big)\big(\theta_{j,k}^0 - \psi(\theta)_{j,k}\big) \le 2^{-7}\|\theta - \psi(\theta)\|_{L^2}^2.$$

Since also $\sum_{j,k}\theta_{j,k}(\theta_{j,k}^0 - \psi(\theta)_{j,k}) \le \overline{L}Q\phi_n$, $\frac{1}{2} - \|\psi(\theta)\|_{L^2} \le -1/4$ and

$$-\sum_{j,k}\left(\frac{\theta_{j,k}}{\|\theta\|_{L^2}} - \frac{\psi(\theta)_{j,k}}{\|\psi(\theta)\|_{L^2}}\right)^2 \le -\frac{1}{4}\|\theta - \psi(\theta)\|_{L^2}^2 + 2\left(\frac{1}{\|\theta\|_{L^2}} - \frac{1}{\|\psi(\theta)\|_{L^2}}\right)^2,$$

we obtain that $n^{-1}(\mathcal{L}_n(\theta) - \mathcal{L}_n(\psi(\theta)))$ is less than

$$-\frac{3}{64}\|\theta - \psi(\theta)\|_{L^2}^2 + \frac{1}{2}\left(\frac{1}{\|\theta\|_{L^2}} - \frac{1}{\|\psi(\theta)\|_{L^2}}\right)^2 + 2\overline{L}Q\phi_n\left|\frac{1}{\|\theta\|_{L^2}} - \frac{1}{\|\psi(\theta)\|_{L^2}}\right|.$$

Now, we need to distinguish two cases. First, assume that there is a (j,k) with $j \le J_0$ and $\psi(\theta)_{j,k} \ne \theta_{j,k}$. In this case, $\|\theta - \psi(\theta)\|_{L^2}^2 \ge Q^2\phi_n^2$. By (A.9) and the choice of Q,

$$\begin{aligned}\frac{1}{n}\big(\mathcal{L}_n(\theta) - \mathcal{L}_n\big(\psi(\theta)\big)\big) &\le -\frac{1}{64}\|\theta - \psi(\theta)\|_{L^2}^2 + \big(2^7\overline{L}^2 + 2^5\overline{L}^2 Q - 2^{-5}Q^2\big)\phi_n^2 \\ &\le -\frac{1}{64}\|\theta - \psi(\theta)\|_{L^2}^2.\end{aligned}$$

Now, suppose the opposite, that is, whenever $\psi(\theta)_{j,k} \ne \theta_{j,k}$ then $j > J_0$. By construction of J_0 [see (A.10)],

$$\begin{aligned}\left|\frac{1}{\|\theta\|_{L^2}} - \frac{1}{\|\psi(\theta)\|_{L^2}}\right| &\le 4\big|\langle\psi(\theta) + \theta, \psi(\theta) - \theta\rangle\big| \\ &\le 4 \max_{j > J_0, k \in I_j} |\theta_{j,k} + \psi(\theta)_{j,k}|\phi_n^{-1}\|\theta - \psi(\theta)\|_{L^2}^2 \\ &\le 2^{-7}(\overline{L}Q)^{-1}\phi_n^{-1}\|\theta - \psi(\theta)\|_{L^2}^2.\end{aligned}$$

Similar, we obtain

$$\left(\frac{1}{\|\theta\|_{L^2}} - \frac{1}{\|\psi(\theta)\|_{L^2}}\right)^2 \le 2^{-5}\|\theta - \psi(\theta)\|_{L^2}^2,$$

by using the Cauchy–Schwarz inequality instead. Therefore, in this case, we also get $\frac{1}{n}(\mathcal{L}_n(\theta) - \mathcal{L}_n(\psi(\theta))) \le -\frac{1}{64}\|\theta - \psi(\theta)\|_{L^2}^2$. This completes the proof.

Acknowledgements. The comments and suggestions of two anonymous referees and an Associate Editor that helped to improve considerably a previous version of this work are gratefully acknowledged.

REFERENCES

[1] ARBEL, J., GAYRAUD, G. and ROUSSEAU, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scand. J. Stat.* **40** 549–570. MR3091697

[2] BARRON, A. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical report, Univ. Illinois at Urbana-Campaign.

[3] BELITSER, E. and GHOSAL, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.* **31** 536–559. MR1983541

[4] BROWN, L. D. and LOW, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24** 2524–2535. MR1425965

[5] CAI, T. T. (2008). On information pooling, adaptability and superefficiency in nonparametric function estimation. *J. Multivariate Anal.* **99** 421–436. MR2396972

[6] CAI, T. T. and LOW, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** 1805–1840. MR2102494

[7] CAI, T. T. and LOW, M. G. (2005). Adaptive estimation of linear functionals under different performance measures. *Bernoulli* **11** 341–358. MR2132730
[8] CAI, T. T. and LOW, M. G. (2006). Adaptation under probabilistic error for estimating linear functionals. *J. Multivariate Anal.* **97** 231–245. MR2208851
[9] CAI, T. T., LOW, M. G. and ZHAO, L. H. (2007). Trade-offs between global and local risks in nonparametric function estimation. *Bernoulli* **13** 1–19. MR2307391
[10] CASTILLO, I., KERKYACHARIAN, G. and PICARD, D. (2014). Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields* **158** 665–710. MR3176362
[11] CHOUDHURI, N., GHOSAL, S. and ROY, A. (2004). Bayesian estimation of the spectral density of a time series. *J. Amer. Statist. Assoc.* **99** 1050–1059. MR2109494
[12] COHEN, A. (2003). *Numerical Analysis of Wavelet Methods*. North-Holland, Amsterdam. MR1990555
[13] COHEN, A., DAUBECHIES, I. and VIAL, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** 54–81. MR1256527
[14] DONOHO, D. and LIU, R. G. (1991). Geometrizing rates of convergence III. *Ann. Statist.* **19** 668–701.
[15] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007
[16] GHOSAL, S. and VAN DER VAART, A. (2006). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274
[17] GINÉ, E. and NICKL, R. (2011). Rates on contraction for posterior distributions in L^r-metrics, $1 \le r \le \infty$. *Ann. Statist.* **39** 2883–2911. MR3012395
[18] HOFFMANN, M. and NICKL, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409. MR2906872
[19] KRUIJER, W., ROUSSEAU, J. and VAN DER VAART, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.* **4** 1225–1257. MR2735885
[20] LEPSKIĬ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466. MR1091202
[21] LE CAM, L. and YANG, G. L. (2000). *Asymptotics in Statistics: Some Basic Concepts*, 2nd ed. Springer, New York. MR1784901
[22] LOW, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. MR1604412
[23] ROBERT, C. (2004). *The Bayesian Choice*. Springer, New York.
[24] ROUSSEAU, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.* **38** 146–180. MR2589319
[25] ROUSSEAU, J., CHOPIN, N. and LISEO, B. (2012). Bayesian nonparametric estimation of the spectral density of a long or intermediate memory Gaussian process. *Ann. Statist.* **40** 964–995. MR2985940
[26] SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26. MR0184378
[27] SCRICCIOLO, C. (2014). Adaptive Bayesian density estimation in L^p-metrics with Pitman–Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Anal.* **9** 475–520. MR3217004
[28] SHEN, W., TOKDAR, S. and GHOSAL, S. (2012). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. Technical report.
[29] TANG, Y. and GHOSAL, S. (2007). Posterior consistency of Dirichlet mixtures for estimating a transition density. *J. Statist. Plann. Inference* **137** 1711–1726. MR2323858
[30] 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

[31] ZHAO, L. H. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.* **28** 532–552. MR1790008

M. HOFFMANN
J. ROUSSEAU
CEREMADE
UNIVERSITÉ PARIS DAUPHINE
PLACE DU MARÉCHAL DE LATTRE DE TASSIGNY
75016 PARIS
FRANCE
E-MAIL: hoffmann@ceremade.dauphine.fr
rousseau@ceremade.dauphine.fr

MATHEMATICAL INSTITUTE
UNIVERSITY OF LEIDEN
NIELS BOHRWEG 1
2333 CA LEIDEN
THE NETHERLANDS
E-MAIL: schmidthieberaj@math.leidenuniv.nl