The Annals of Statistics
2017, Vol. 45, No. 2, 833–865
DOI: 10.1214/16-AOS1469

ASYMPTOTIC BEHAVIOUR OF THE EMPIRICAL BAYES POSTERIORS ASSOCIATED TO MAXIMUM MARGINAL LIKELIHOOD ESTIMATOR

BY JUDITH ROUSSEAU[1,*,†] AND BOTOND SZABO[2,‡,§]

University Paris Dauphine, CREST-ENSAE,† Budapest University of Technology‡ and Leiden University§*

We consider the asymptotic behaviour of the marginal maximum likelihood empirical Bayes posterior distribution in general setting. First, we characterize the set where the maximum marginal likelihood estimator is located with high probability. Then we provide oracle type of upper and lower bounds for the contraction rates of the empirical Bayes posterior. We also show that the hierarchical Bayes posterior achieves the same contraction rate as the maximum marginal likelihood empirical Bayes posterior. We demonstrate the applicability of our general results for various models and prior distributions by deriving upper and lower bounds for the contraction rates of the corresponding empirical and hierarchical Bayes posterior distributions.

1. Introduction. In the Bayesian approach, the whole inference is based on the posterior distribution, which is proportional to the likelihood times the prior (in case of dominated models). The task of designing a prior distribution Π on the parameter $\theta \in \Theta$ is difficult and in large dimensional models cannot be performed in a fully subjective way. It is therefore common practice to consider a family of prior distributions $\Pi(\cdot|\lambda)$ indexed by a hyper-parameter $\lambda \in \Lambda$ and to either put a hyper-prior on λ (hierarchical approach) or to choose λ depending on the data, so that $\lambda = \hat{\lambda}(\mathbf{x}_n)$ where $\mathbf{x}_n$ denotes the collection of observations. The latter is referred to as an empirical Bayes (hereafter EB) approach, see for instance [17]. There are many ways to select the hyper-parameter λ based on the data, in particular depending on the nature of the hyper-parameter.

Recently, [19] have studied the asymptotic behaviour of the posterior distribution for general empirical Bayes approaches; they provide conditions to obtain consistency of the EB posterior and in the case of parametric models characterized the behaviour of the maximum marginal likelihood estimator $\hat{\lambda}_n \equiv \hat{\lambda}(\mathbf{x}_n)$ (hereafter MMLE), together with the corresponding posterior distribution $\Pi(\cdot|\hat{\lambda}_n; \mathbf{x}_n)$

Received April 2015; revised March 2016.

[1]Supported in part by the ANR IPANEMA, the labex ECODEC.

[2]Supported in part by the labex ECODEC and Netherlands Organization for Scientific Research (NWO). The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

MSC2010 subject classifications. Primary 62G20, 62G05, 60K35; secondary 62G08, 62G07.

Key words and phrases. Posterior contraction rates, adaptation, empirical Bayes, hierarchical Bayes, nonparametric regression, density estimation, Gaussian prior, truncation prior.

on θ. They show that asymptotically the MMLE converges to some oracle value λ_0 which maximizes, in λ, the prior density calculated at the true value θ_0 of the parameter, $\pi(\theta_0|\lambda_0) = \sup\{\pi(\theta_0|\lambda), \lambda \in \Lambda\}$, where the density is with respect to Lebesgue measure. This cannot be directly extended to the nonparametric setup, since in this case, typically the prior distributions $\Pi(\cdot|\lambda)$, $\lambda \in \Lambda$ are not absolutely continuous with respect to a fixed measure. In the nonparametric setup, the asymptotic behaviour of the MMLE and its associated EB posterior distribution has been studied in the (inverse) white noise model under various families of Gaussian prior processes by [3, 9, 14, 28, 29], in the nonparametric regression problem with smoothing spline priors [24] and rescaled Brownian motion prior [26], and in a sparse setting by [13]. In all these papers, the results have been obtained via explicit expression of the marginal likelihood. Interesting phenomena have been observed in these specific cases. In [29], an infinite dimensional Gaussian prior was considered with fixed regularity parameter α and a scaling hyper-parameter τ. Then it was shown that the scaling parameter can compensate for possible mismatch of the base regularity α of the prior distribution and the regularity β of the true parameter of interest up to a certain limit. However, too smooth truth can only be recovered sub-optimally by MMLE empirical Bayes method with rescaled Gaussian priors. In contrast to this in [14], it was shown that by substituting the MMLE of the regularity hyper-parameter into the posterior, then one can get optimal contraction rate (up to a $\log n$ factor) for every Sobolev regularity class, simultaneously.

In this paper, we are interested in generalizing the specific results of [14] (in the direct case), [29] to more general models, shading light on what is driving the asymptotic behaviour of the MMLE in nonparametric or large dimensional models. We also provide sufficient conditions to derive posterior concentration rates for EB procedures based on the MMLE. Finally, we investigate the relationship between the MMLE empirical Bayes and hierarchical Bayes approaches. We show that the hierarchical Bayes posterior distribution (under mild conditions on the hyper-prior distribution) achieves the same contraction rate as the MMLE empirical Bayes posterior distribution. Note that our results do not answer the question whether empirical Bayes and hierarchical Bayes posterior distributions are strongly merging, which is certainly of interest, but would require typically a much more precise analysis of the posterior distributions.

More precisely, set $\mathbf{x}_n$ the vector of observations and assume that conditionally on some parameter $\theta \in \Theta$, $\mathbf{x}_n$ is distributed according to P_θ^n with density p_θ^n with respect to some given measure μ. Let $\Pi(\cdot|\lambda), \lambda \in \Lambda$ be a family of prior distributions on Θ. Then the associated posterior distributions are equal to

$$\Pi(B|\mathbf{x}_n; \lambda) = \frac{\int_B p_\theta^n(\mathbf{x}_n)\, d\Pi(\theta|\lambda)}{\bar{m}(\mathbf{x}_n|\lambda)}, \qquad \bar{m}(\mathbf{x}_n|\lambda) = \int_\Theta p_\theta^n(\mathbf{x}_n)\, d\Pi(\theta|\lambda)$$

for all $\lambda \in \Lambda$ and any Borelian subset B of Θ. The MMLE is defined as

$$\hat{\lambda}_n \in \operatorname*{argmax}_{\lambda \in \Lambda_n} \bar{m}(\mathbf{x}_n|\lambda) \tag{1.1}$$

for some $\Lambda_n \subseteq \Lambda$, and the associated EB posterior distribution by $\Pi(\cdot|\mathbf{x}_n, \hat{\lambda}_n)$. We note that in case there are multiple maximizers one can take an arbitrary one. Furthermore, from practical consideration (both computational and technical) we allow the maximizer to be taken on the subset $\Lambda_n \subseteq \Lambda$.

Our aim is two-fold, first to characterize the asymptotic behaviour of $\hat{\lambda}_n$ and second to derive posterior concentration rates in such models, that is, to determine sequences ε_n going to 0 such that

$$\Pi\big(\theta : d(\theta, \theta_0) \leq \varepsilon_n |\mathbf{x}_n; \hat{\lambda}_n\big) \to 1 \tag{1.2}$$

in probability under $P_{\theta_0}^n$, with $\theta_0 \in \Theta$ and $d(\cdot, \cdot)$ some appropriate positive loss function on Θ [typically a metric or semi-metric, see condition (A2) later for more precise description]. There is now a substantial literature on posterior concentration rates in large or infinite dimensional models initiated by the seminal paper of [11]. Most results, however, deal with fully Bayesian posterior distributions, that is, associated to priors that are not data dependent. The literature on EB posterior concentration rates deals mainly with specific models and specific priors.

Recently, in [8], sufficient conditions are provided for deriving general EB posterior concentration rates when it is known that $\hat{\lambda}_n$ belongs to a well chosen subset Λ_0 of Λ. In essence, their result boils down to controlling $\sup_{\lambda \in \Lambda_0} \Pi(d(\theta, \theta_0) > \varepsilon_n |\mathbf{x}_n, \lambda)$. Hence, either λ has very little influence on the posterior concentration rate and it is not so important to characterize precisely Λ_0 or λ is influential and it becomes crucial to determine properly Λ_0. In [8], the authors focus on the former. In this paper, we are mainly concerned with the latter, with $\hat{\lambda}_n$ the MMLE. Since the MMLE is an implicit estimator (as opposed to the moment estimates considered in [8]) the main difficulty here is to understand what the set Λ_0 is.

We show in this paper that Λ_0 can be characterized roughly as

$$\Lambda_0 = \big\{\lambda : \varepsilon_n(\lambda) \leq M_n \varepsilon_{n,0}\big\}$$

for any sequence M_n going to infinity and with $\varepsilon_{n,0} = \inf\{\varepsilon_n(\lambda); \lambda \in \Lambda_n\}$ and $\varepsilon_n(\lambda)$ satisfying

$$\Pi\big(\|\theta - \theta_0\| \leq K \varepsilon_n(\lambda) |\lambda\big) = e^{-n\varepsilon_n^2(\lambda)}, \tag{1.3}$$

with $(\Theta, \|\cdot\|)$ a Banach space and for some large enough constant K [in the notation we omitted the dependence of $\varepsilon_n(\lambda)$ on K and θ_0]. We then prove that the concentration rate of the MMLE empirical Bayes posterior distribution is of order $O(M_n \varepsilon_{n,0})$. We also show that the preceding rates are sharp, that is, the posterior contraction rate is bounded from below by $\delta_n \varepsilon_{n,0}$ [for arbitrary $\delta_n = o(1)$]. Hence, our results reveal the exact posterior contraction rates for every individual $\theta_0 \in \Theta$. Furthermore, we also show that the hierarchical Bayes method behaves similarly, that is, the hierarchical posterior has the same upper $(M_n \varepsilon_{n,0})$ and lower $(\delta_n \varepsilon_{n,0})$ bounds on the contraction rate for every $\theta_0 \in \Theta$ as the MMLE empirical Bayes posterior.

Our aim is not so much to advocate the use of the MMLE empirical Bayes approach, but rather to understand its behaviour. Interestingly, our results show that it is driven by the choice of the prior family $\{\Pi(\cdot|\lambda), \lambda \in \Lambda)\}$ in the neighbourhood of the true parameter θ_0. This allows to determine a priori which family of prior distributions will lead to well behaved MMLE empirical Bayes posteriors and which will not. In certain cases, however, the computation of the MMLE is very challenging. Therefore, it would be interesting to investigate other type of estimators for the hyper-parameters like the cross validation estimator. At the moment, there is only a limited number of papers on this topic and only for specific models and priors; see, for instance, [26, 27].

These results are summarized in Theorem 2.1, in Corollary 2.1, and in Theorem 2.3, in Section 2. Then three different types of priors on $\Theta = \ell_2 = \{(\theta_j)_{j\in\mathbb{N}}; \sum_j \theta_j^2 < +\infty\}$ are studied, for which upper bounds on $\varepsilon_n(\lambda)$ are given in Section 3.1. We apply these results to three different sampling models: the Gaussian white noise, the regression and the estimation of the density based on i.i.d. data models in Sections 3.5 and 3.6. Proofs are postponed to Section 4, to the Appendix for those concerned with the determination of $\varepsilon_n(\lambda)$ and to the Supplementary Material [23].

1.1. *Notation and setup.* We assume that the observations $\mathbf{x}_n \in \mathcal{X}_n$ (where $\mathcal{X}_n$ denotes the sample space) are distributed according to a distribution P_θ^n (they are not necessarily i.i.d.), with $\theta \in \Theta$, where $(\Theta, \|\cdot\|)$ is a Banach space. We denote by μ a dominating measure and by p_θ^n and E_θ^n the corresponding density and expected value of P_θ^n, respectively. We consider the family of prior distributions $\{\Pi(\cdot|\lambda), \lambda \in \Lambda\}$ on Θ with $\Lambda \subset \mathbb{R}^d$ for some $d \geq 1$ and we denote by $\Pi(\cdot|\mathbf{x}_n; \lambda)$ the associated posterior distributions.

Throughout the paper, $K(\theta_0, \theta)$ denotes the Kullback–Leibler divergence between $P_{\theta_0}^n$ and P_θ^n for all $\theta, \theta_0 \in \Theta$ while $V_2(\theta_0, \theta)$ denotes the centered second moment of the log-likelihood:

$$K(\theta_0, \theta) = \int_{\mathcal{X}_n} p_{\theta_0}^n(\mathbf{x}_n) \log\left(\frac{p_{\theta_0}^n}{p_\theta^n}(\mathbf{x}_n)\right) d\mu(\mathbf{x}_n),$$

$$V_2(\theta_0, \theta) = E_{\theta_0}^n\left(\left|\ell_n(\theta_0) - \ell_n(\theta) - K(\theta_0, \theta)\right|^2\right)$$

with $\ell_n(\theta) = \log p_\theta^n(\mathbf{x}_n)$. As in [12], we define the Kullback–Leibler neighbourhoods of θ_0 as

$$B(\theta_0, \varepsilon, 2) = \left\{\theta; K(\theta_0, \theta) \leq n\varepsilon^2, V_2(\theta_0, \theta) \leq n\varepsilon^2\right\}$$

and note that in the above definition $V_2(\theta_0, \theta) \leq n\varepsilon^2$ can be replaced by $V_2(\theta_0, \theta) \leq Cn\varepsilon^2$ for any positive constant C without changing the results.

For any subset $A \subset \Theta$ and $\varepsilon > 0$, we denote $\log N(\varepsilon, A, d(\cdot, \cdot))$ the ε-entropy of A with respect to the (pseudo) metric $d(\cdot, \cdot)$, that is, the logarithm of the covering number of A by $d(\cdot, \cdot)$ balls of radius ε.

We also write

$$m(\mathbf{x}_n|\lambda) = \frac{\bar{m}(\mathbf{x}_n|\lambda)}{p^n_{\theta_0}(\mathbf{x}_n)} = \frac{\int_\Theta p^n_\theta(\mathbf{x}_n)\, d\Pi(\theta|\lambda)}{p^n_{\theta_0}(\mathbf{x}_n)}.$$

For any bounded function f, $\|f\|_\infty = \sup_x |f(x)|$ and if φ denotes a countable collection of functions $(\varphi_i, i \in \mathbb{N})$, then $\|\varphi\|_\infty = \max_i \|\varphi_i\|_\infty$. If the function is integrable, then $\|f\|_1$ denotes its L_1 norm while $\|f\|_2$ its L_2 norm and if $\theta \in \ell_r = \{\theta = (\theta_i)_{i\in\mathbb{N}}, \sum_i |\theta_i|^r < +\infty\}$, with $r \geq 1$, $\|\theta\|_r = (\sum_i |\theta_i|^r)^{1/r}$.

Throughout the paper, $x_n \lesssim y_n$ means that there exists a constant C such that for n large enough $x_n \leq C y_n$, similarly with $x_n \gtrsim y_n$ and $x_n \asymp y_n$ is equivalent to $y_n \lesssim x_n \lesssim y_n$. For equivalent (abbreviated) notation, we use the symbol $\equiv$.

2. Asymptotic behaviour of the MMLE, its associated posterior distribution and the hierarchical Bayes method. Although the problem can be formulated as a classical parametric maximum likelihood estimation problem, since λ is finite dimensional, its study is more involved than the usual regular models due to the complicated nature of the marginal likelihood. Indeed $m(\mathbf{x}_n|\lambda)$ is an integral over an infinite (or large) dimensional space.

For $\theta_0 \in \Theta$ denoting the true parameter, define the sequence $\varepsilon_n(\lambda) \equiv \varepsilon_n(\lambda, \theta_0, K)$ as

$$\Pi\big(\theta : \|\theta - \theta_0\| \leq K\varepsilon_n(\lambda)|\lambda\big) = e^{-n\varepsilon_n(\lambda)^2}, \tag{2.1}$$

for some positive parameter $K > 0$. If the cumulative distribution function of $\|\theta - \theta_0\|$ under $\Pi(\cdot|\lambda)$ is not continuous, then the definition of $\varepsilon_n(\lambda)$ can be replaced by

$$\tilde{c}_0^{-1} n\varepsilon_n(\lambda)^2 \leq -\log \Pi\big(\theta : \|\theta - \theta_0\| \leq K\varepsilon_n(\lambda)|\lambda\big) \leq \tilde{c}_0 n\varepsilon_n(\lambda)^2, \tag{2.2}$$

for some $\tilde{c}_0 \geq 1$ under the assumption that such a sequence $\varepsilon_n(\lambda)$ exists.

Roughly speaking, under the assumptions stated below, $\log m(\mathbf{x}_n|\lambda) \asymp n\varepsilon_n^2(\lambda)$ and $\varepsilon_n(\lambda)$ is the posterior concentration rate associated to the prior $\Pi(\cdot|\lambda)$ and the best possible (oracle) posterior concentration rate over $\lambda \in \Lambda_n$ is denoted

$$\varepsilon^2_{n,0} = \inf_{\lambda\in\Lambda_n} \big\{\varepsilon_n(\lambda)^2 : \varepsilon_n(\lambda)^2 \geq m_n(\log n)/n\big\} \vee m_n(\log n)/n,$$

with any sequence m_n tending to infinity.

With the help of the oracle value $\varepsilon_{n,0}$, we define a set of hyper-parameters with similar properties, as

$$\Lambda_0 \equiv \Lambda_0(M_n) \equiv \Lambda_{0,n}(K, \theta_0, M_n) = \big\{\lambda \in \Lambda_n : \varepsilon_n(\lambda) \leq M_n\varepsilon_{n,0}\big\}, \tag{2.3}$$

with any sequence M_n going to infinity. We show that under general (and natural) assumptions the marginal maximum likelihood estimator $\hat{\lambda}_n$ belongs to the set Λ_0 with probability tending to one, for some constant $K > 0$ large enough. The parameter K provides extra flexibility to the approach and simplifies the proofs of the upcoming conditions in certain examples. In practice (at least in the examples

we have studied), the constant K essentially modifies $\varepsilon_n(\lambda)$ by a multiplicative constant, and thus does not modify the final posterior concentration rate, nor the set Λ_0 since M_n is any sequence going to infinity. Note that our results are only meaningful in cases where $\varepsilon_n(\lambda)$ defined by (2.2) vary with λ.

We now give general conditions under which the MMLE is inside of the set Λ_0 with probability going to 1 under $P_{\theta_0}^n$. Using [8], we will then deduce that the concentration rate of the associated MMLE empirical Bayes posterior distribution is bounded by $M_n\varepsilon_{n,0}$.

Following [19] and [8], we construct for all $\lambda, \lambda' \in \Lambda_n$ a transformation $\psi_{\lambda,\lambda'} : \Theta \mapsto \Theta$ such that if $\theta \sim \Pi(\cdot|\lambda)$ then $\psi_{\lambda,\lambda'}(\theta) \sim \Pi(\cdot|\lambda')$ and for a given sequence $u_n \to 0$ we introduce the notation

$$q_{\lambda,n}^{\theta}(\mathbf{x}_n) = \sup_{\rho(\lambda,\lambda') \le u_n} p_{\psi_{\lambda,\lambda'}(\theta)}^n(\mathbf{x}_n), \tag{2.4}$$

where $\rho : \Lambda_n \times \Lambda_n \to \mathbb{R}^+$ is some loss function and $Q_{\lambda,n}^{\theta}$ the associated measure. Denote by $N_n(\Lambda_0)$, $N_n(\Lambda_n \setminus \Lambda_0)$, and $N_n(\Lambda_n)$ the covering number of Λ_0, $\Lambda_n \setminus \Lambda_0$ and Λ_n by balls of radius u_n, respectively, with respect of the loss function ρ.

We consider the following set of assumptions to bound $\sup_{\lambda \in \Lambda_n \setminus \Lambda_0} m(\mathbf{x}_n|\lambda)$ from above:

- (A1) There exists $N > 0$ such that for all $\lambda \in \Lambda_n \setminus \Lambda_0$ and $n \ge N$, there exists $\Theta_n(\lambda) \subset \Theta$

$$\sup_{\{\|\theta-\theta_0\| \le K\varepsilon_n(\lambda)\} \cap \Theta_n(\lambda)} \frac{\log Q_{\lambda,n}^{\theta}(\mathcal{X}_n)}{n\varepsilon_n(\lambda)^2} = o(1), \tag{2.5}$$

  and such that

$$\int_{\Theta_n(\lambda)^c} Q_{\lambda,n}^{\theta}(\mathcal{X}_n)\, d\Pi(\theta|\lambda) \le e^{-w_n^2 n\varepsilon_{n,0}^2}, \tag{2.6}$$

  for some positive sequence w_n going to infinity.
- (A2) [tests] There exists $0 < \zeta, c_1 < 1$ such that for all $\lambda \in \Lambda_n \setminus \Lambda_0$ and all $\theta \in \Theta_n(\lambda)$, there exist tests $\varphi_n(\theta)$ such that

$$\begin{aligned} E_{\theta_0}^n \varphi_n(\theta) &\le e^{-c_1 n d^2(\theta,\theta_0)}, \\ \sup_{d(\theta,\theta') \le \zeta d(\theta,\theta_0)} Q_{\lambda,n}^{\theta'}\big(1-\varphi_n(\theta)\big) &\le e^{-c_1 n d^2(\theta,\theta_0)}, \end{aligned} \tag{2.7}$$

  where $d(\cdot,\cdot)$ is a semi-metric satisfying

$$\Theta_n(\lambda) \cap \big\{\|\theta-\theta_0\| > K\varepsilon_n(\lambda)\big\} \subset \Theta_n(\lambda) \cap \big\{d(\theta,\theta_0) > c(\lambda)\varepsilon_n(\lambda)\big\} \tag{2.8}$$

  for some $c(\lambda) \ge w_n\varepsilon_{n,0}/\varepsilon_n(\lambda)$ and

$$\log N\big(\zeta u, \{u \le d(\theta,\theta_0) \le 2u\} \cap \Theta_n(\lambda), d(\cdot,\cdot)\big) \le c_1 n u^2/2 \tag{2.9}$$

  for all $u \ge c(\lambda)\varepsilon_n(\lambda)$.

REMARK 2.1. We note that we can weaken (2.5) to

$$\sup_{\{\|\theta-\theta_0\|\le\varepsilon_n(\lambda)\}\cap\Theta_n(\lambda)} Q^{\theta}_{\lambda,n}(\mathcal{X}_n) \le e^{cn\varepsilon_n^2(\lambda)},$$

for some positive constant $c < 1$ in case the cumulative distribution of $\|\cdot - \theta_0\|$ under $\Pi(\cdot|\lambda)$ is continuous, and hence the definition (2.1) is meaningful.

Conditions (2.5) and (2.6) imply that we can control the small perturbations of the likelihood $p^n_{\psi_{\lambda,\lambda'}(\theta)}(\mathbf{x}_n)$ due to the change of measures $\psi_{\lambda,\lambda'}$ and are similar to those used in [8]. They allow us to control $m(\mathbf{x}_n|\lambda)$ uniformly over $\Lambda_n \setminus \Lambda_0$. They are rather weak conditions since u_n can be chosen very small. In [8], the authors show that they hold even with complex priors such as nonparametric mixture models. Assumption (2.7), together with (2.9) have been verified in many contexts, with the difference that here the tests need to be performed with respect to the perturbed likelihoods $q^{\theta}_{\lambda,n}$. Since the u_n—mesh of $\Lambda_n \setminus \Lambda_0$ can be very fine, these perturbations can be well controlled over the sets $\Theta_n(\lambda)$; see, for instance, [8] in the context of density estimation or intensity estimation of Aalen point processes. The interest of the above conditions is that they are very similar to standard conditions considered in the posterior concentration rates literature, starting with [11] and [12], so that there is a large literature on such types of conditions which can be applied in the present setting. Therefore, the usual variations on these conditions can be considered. For instance, an alternative condition to (A2) is:

(A2 bis) There exists $0 < \zeta < 1$ such that for all $\lambda \in \Lambda_n \setminus \Lambda_0$ and all $\theta \in \Theta_n(\lambda)$, there exist tests $\varphi_n(\theta)$ such that (2.7) is verified and for all $j \ge K$, writing

$$B_{n,j}(\lambda) = \Theta_n(\lambda) \cap \{j\varepsilon_n(\lambda) \le \|\theta - \theta_0\| < (j+1)\varepsilon_n(\lambda)\},$$

then

$$B_{n,j}(\lambda) \subset \Theta_n(\lambda) \cap \{d(\theta,\theta_0) > c(\lambda, j)\varepsilon_n(\lambda)\}$$

with

$$\sum_{j\ge K} \exp\left(-\frac{c_1}{2} nc(\lambda, j)^2\varepsilon_n(\lambda)^2\right) \lesssim e^{-nw_n^2\varepsilon_{n,0}^2}$$

and

$$\log N\big(\zeta c(\lambda, j)\varepsilon_n(\lambda), B_{n,j}(\lambda), d(\cdot,\cdot)\big) \le \frac{c_1 nc(\lambda, j)^2\varepsilon_n(\lambda)^2}{2}.$$

Here, the difficulty lies in the comparison between the metric $\|\cdot\|$ of the Banach space and the testing distance $d(\cdot,\cdot)$, in condition (2.8). Outside the white noise model, where the Kullback and other moments of the likelihood ratio are directly linked to the L_2 norm on $\theta - \theta_0$, such comparison may be nontrivial. In van der Vaart and van Zanten [31], the prior had some natural Banach structure and

norm, which was possibly different to the Kullback–Leibler and the testing distance $d(\cdot,\cdot)$, but comparable in some sense. Our approach is similar in spirit. We illustrate this here in the special cases of regression function and density estimation under different families of priors; see Sections 3.5 and 3.6.1. In Section 3.6.2, we use a prior which is not so much driven by a Banach structure and the norm $\|\cdot\|$ is replaced by the Hellinger distance. Hence, in full generality $\|\cdot\|$ could be replaced by any metric, for instance the testing metric $d(\cdot,\cdot)$, as long as the rates $\varepsilon_n(\lambda)$ can be computed.

The following assumption is used to bound from below $\sup_{\lambda\in\Lambda_0} m(\mathbf{x}_n|\lambda)$:

- (B1) There exist $\tilde{\Lambda}_0 \subset \Lambda_0$ and $M_2 \geq 1$ such that for every $\lambda \in \tilde{\Lambda}_0$
$$\big\{\|\theta-\theta_0\| \leq K\varepsilon_n(\lambda)\big\} \subset B\big(\theta_0, M_2\varepsilon_n(\lambda), 2\big),$$
and such that there exists $\lambda_0 \in \tilde{\Lambda}_0$ for which $\varepsilon_n(\lambda_0) \leq M_1\varepsilon_{n,0}$ for some positive M_1.

REMARK 2.2. A variation of (B1) can be considered where $\{\|\theta-\theta_0\| \leq K\varepsilon_n(\lambda)\}$ is replaced by $\{\|\theta-\theta_0\| \leq K\varepsilon_n(\lambda)\} \cap \tilde{\Theta}_n(\lambda)$ where $\tilde{\Theta}_n(\lambda) \subset \Theta$ verifies
$$\Pi\big(\big\{\|\theta-\theta_0\| \leq K\varepsilon_n(\lambda)\big\} \cap \tilde{\Theta}_n(\lambda)|\lambda\big) \gtrsim e^{-K_2 n\varepsilon_n^2(\lambda)},$$
for some $K_2 \geq 1$. This is used in Section 3.6.

2.1. *Asymptotic behaviour of the MMLE and empirical Bayes posterior concentration rate.* We now present the two main results of this section, namely: asymptotic behaviour of the MMLE and concentration rate of the resulting empirical Bayes posterior. We first describe the asymptotic behaviour of $\hat{\lambda}_n$.

THEOREM 2.1. *Assume that there exists $K > 0$ such that conditions* (A1), (A2) *and* (B1) *hold with $w_n = o(M_n)$, then if* $\log N_n(\Lambda_n \setminus \Lambda_0) = o(nw_n^2\varepsilon_{n,0}^2)$,
$$\lim_{n\to\infty} P_{\theta_0}^n(\hat{\lambda}_n \in \Lambda_0) = 1.$$

The proof of Theorem 2.1 is given in Section 4.1.

The above theorem describes the asymptotic behaviour of the MMLE $\hat{\lambda}_n$, via the oracle set Λ_0, in other words it minimizes $\varepsilon_n(\lambda)$. The use of the Banach norm is particularly adapted to the case of priors on parameters $\theta = (\theta_i)_{i\in\mathbb{N}} \in \ell_2$, where the $\theta_i's$ are assumed independent. This type of priors is studied in Section 3.1.

Note that in the definition of $\Lambda_0(M_n)$, M_n can be any sequence going to infinity. In the examples, we have considered in Section 3.1, M_n can be chosen to increase to infinity arbitrarily slowly. If $\varepsilon_n(\lambda)$ is (rate) constant, (2.1) presents no interest since $\Lambda_0 = \Lambda_n$, but if for some $\lambda \neq \lambda'$ the fraction $\varepsilon_n(\lambda)/\varepsilon_n(\lambda')$ either goes to infinity or to 0, then choosing M_n increasing slowly enough to infinity, Theorem 2.1 implies that the MMLE converges to a meaningful subset of Λ_n. In particular, our

results are too crude to be informative in the parametric case. Indeed from [19], in the parametric non-degenerate case $\varepsilon_n(\lambda) \asymp \sqrt{(\log n)/n}$ in definition (2.2) for all λ and $\Lambda_0 = \Lambda_n$. In the parametric degenerative case, where the λ_0 belongs to the boundary of the set Λ then one would have at the limit $\pi(\cdot|\lambda_0) = \delta_{\theta_0}$ corresponding to $\varepsilon_n(\lambda_0) = 0$. So we do recover the oracle parametric value of [19]. However, for the condition $\log N_n(\Lambda_n \setminus \Lambda_0) = o(nw_n^2\varepsilon_{n,0}^2)$ to be valid one would require essentially that Λ_0 is the whole set Λ_n.

Using the above theorem, together with [8], we obtain the associated posterior concentration rate, controlling uniformly $\Pi(d(\theta_0, \theta) \leq \varepsilon_n|\mathbf{x}_n, \lambda)$ over $\lambda \in \Lambda_0$, with $\varepsilon_n = M_n\varepsilon_{n,0}$. To do so, we consider the following additional assumptions:

- (C1) For every $c_2 > 0$ there exists constant $N > 0$ such that for all $\lambda \in \Lambda_0$ and $n \geq N$, there exists $\Theta_n(\lambda)$ satisfying

$$\sup_{\lambda\in\Lambda_0} \int_{\Theta_n(\lambda)^c} \mathcal{Q}_{\lambda,n}^{\theta}(\mathcal{X}_n)\, d\Pi(\theta|\lambda) \leq e^{-c_2 n\varepsilon_{n,0}^2} \tag{2.10}$$

- (C2) There exists $0 < c_1, \zeta < 1$ such that for all $\lambda \in \Lambda_0$ and all $\theta \in \Theta_n(\lambda)$, there exist tests $\varphi_n(\theta)$ satisfying (2.7) and (2.9), where (2.9) is supposed to hold for any $u \geq MM_n\varepsilon_{n,0}$ for some $M > 0$.
- (C3) There exists $C_0 > 0$ such that for all $\lambda \in \Lambda_0$, for all $\theta \in \{d(\theta_0, \theta) \leq M_n\varepsilon_{n,0}\} \cap \Theta_n(\lambda)$,

$$\sup_{\rho(\lambda,\lambda')\leq u_n} d\big(\theta, \psi_{\lambda,\lambda'}(\theta)\big) \leq C_0 M_n\varepsilon_{n,0}.$$

COROLLARY 2.1. *Assume that* $\hat{\lambda}_n \in \Lambda_0$ *with probability going to* 1 *under* $P_{\theta_0}^n$ *and that assumptions* (C1)–(C3) *and* (B1) *are satisfied, then if* $\log N_n(\Lambda_0) \leq O(n\varepsilon_{n,0}^2)$, *there exists* $M > 0$ *such that*

$$E_{\theta_0}^n \Pi\big(\theta : d(\theta, \theta_0) \geq MM_n\varepsilon_{n,0}|\mathbf{x}_n; \hat{\lambda}_n\big) = o(1). \tag{2.11}$$

A consequence of Corollary 2.1 is in terms of frequentist risks of Bayesian estimators. Following [4], one can construct an estimator based on the posterior which converges at the posterior concentration rate: $E_{\theta_0}[d(\hat{\theta}, \theta_0)] = O(M_n\varepsilon_{n,0})$. Similar results can also be derived for the posterior mean in case $d(\cdot, \cdot)$ is convex and bounded, and (2.11) is of order $O(M_n\varepsilon_{n,0})$; see, for instance, [11].

Corollary 2.1 is proved in a similar way to Theorem 1 of [8], apart from the lower bound on the marginal likelihood since here we use the nature of the MMLE which simplifies the computations. The details are presented in Section 4.2. We can refine the condition on tests (C3) by considering slices as in [8].

Next, we provide a lower bound on the contraction rate of the MMLE empirical Bayes posterior distribution. For this, we have to introduce some further assumptions. First of all, we extend assumption (2.5) to the set Λ_0. Let $e : \Theta \times \Theta \to \mathbb{R}^+$

be a pseudo-metric and assume that for all $\lambda \in \Lambda_0$ and some δ_n tending to zero we have

$$\begin{aligned} &\sup_{\{\|\theta-\theta_0\| \le \varepsilon_n(\lambda)\} \cap \Theta_n(\lambda)} \frac{\log Q^{\theta}_{\lambda,n}(\mathcal{X}_n)}{n\varepsilon_n^2(\lambda)} = o(1), \\ &\sup_{\lambda \in \Lambda_0} \frac{n\varepsilon_{n,0}^2}{-\log \Pi(\theta : e(\theta,\theta_0) \le 2\delta_n \varepsilon_{n,0} | \lambda)} = o(1) \end{aligned} \tag{2.12}$$

and consider the modified version of (C3): (C3bis). There exists $C_0 > 0$ such that for all $\lambda \in \Lambda_0$, for all $\theta \in \{e(\theta_0,\theta) \le \delta_n \varepsilon_{n,0}\} \cap \Theta_n(\lambda)$,

$$\sup_{\rho(\lambda,\lambda') \le u_n} d\big(\theta, \psi_{\lambda,\lambda'}(\theta)\big) \le C_0 \delta_n \varepsilon_{n,0}.$$

THEOREM 2.2. *Assume that conditions* (A1)–(C2) *and* (C3bis) *together with assumption* (2.12) *hold. In case* $\log N_n(\Lambda_0) = o(n\varepsilon_{n,0}^2)$ *and* $\varepsilon_{n,0}^2 > m_n(\log n)/n$, *we get that*

$$E^n_{\theta_0} \Pi\big(\theta : e(\theta,\theta_0) \le \delta_n \varepsilon_{n,0} | \hat{\lambda}_n; \mathbf{x}_n\big) = o(1).$$

Typically $e(\cdot,\cdot)$ will be either $d(\cdot,\cdot)$ or $\|\cdot\|$. The lower bound is proved using the same argument as the one used to bound $E^n_{\theta_0}(\Pi(\Theta_n^c | \hat{\lambda}_n, \mathbf{x}_n))$ (see Section 4.1 and 4.2), where $\{d(\theta,\theta_0) \le \delta_n \varepsilon_{n,0}\}$ plays the same role as Θ_n^c. We postpone the details of the proof to Section 1.7 of the Supplementary Material [23].

2.2. *Contraction rate of the hierarchical Bayes posterior.* In this section, we investigate the relation between the MMLE empirical Bayes method and the hierarchical Bayes method. We show that under the preceding assumptions complemented with not too restrictive conditions on the hyper-prior distribution the hierarchical posterior distribution achieves the same convergence rate as the MMLE empirical Bayes posterior. Let us denote by $\tilde{\pi}(\cdot)$ the density function of the hyper-prior, then the hierarchical prior takes the form

$$\Pi(\cdot) = \int_{\Lambda} \Pi(\cdot|\lambda) \tilde{\pi}(\lambda)\, d\lambda.$$

Note that we integrate here over the whole hyper-parameter space Λ, not over the subset $\Lambda_n \subseteq \Lambda$ used in the MMLE empirical Bayes approach.

Intuitively, to have the same contraction rate, one would need that the set of probable hyper-parameter values Λ_0 accumulates enough hyper-prior mass. Let us introduce a sequence $\tilde{w}_n$ tending to infinity and satisfying $\tilde{w}_n = o(M_n \wedge w_n)$ and denote by $\Lambda_0(\tilde{w}_n)$ the set defined in (2.3) with $\tilde{w}_n$.

- (H1) Assume that $\tilde{\Lambda}_0 \subset \Lambda_0(\tilde{w}_n)$ and for some $\bar{c}_0 > 0$ there exists $N > 0$ such that for all $n \ge N$ the hyper-prior satisfies

$$\int_{\tilde{\Lambda}_0} \tilde{\pi}(\lambda)\, d\lambda \gtrsim e^{-n\varepsilon_{n,0}^2}$$

and

$$\int_{\Lambda_n^c} \tilde{\pi}(\lambda)\, d\lambda \le e^{-\bar{c}_0 n \varepsilon_{n,0}^2}.$$

- (H2) Uniformly over $\lambda \in \tilde{\Lambda}_0$ and $\{\theta : \|\theta - \theta_0\| \le K\varepsilon_n(\lambda)\}$ there exists $c_3 > 0$ such that

$$P_{\theta_0}^n \Big\{ \inf_{\lambda' : \rho(\lambda,\lambda') \le u_n} \ell_n\big(\psi_{\lambda,\lambda'}(\theta)\big) - \ell_n(\theta_0) \le -c_3 n \varepsilon_n(\lambda)^2 \Big\} = O\big(e^{-n\varepsilon_{n,0}^2}\big).$$

We can then show that the preceding condition is sufficient for giving upper and lower bounds for the contraction rate of the hierarchical posterior distribution.

THEOREM 2.3. *Assume that the conditions of Theorem* 2.1 *and Corollary* 2.1 *hold alongside with conditions* (H1) *with* $\bar{c}_0 > 2M_2^2 + 1$ *and* (H2). *Then the hierarchical posterior achieves the oracle contraction rate* (*up to a slowly varying term*)

$$E_{\theta_0}^n \Pi\big(\theta : d(\theta, \theta_0) \ge M M_n \varepsilon_{n,0} | \mathbf{x}_n\big) = o(1).$$

Furthermore, *if condition* (2.12) *also holds we have that*

$$E_{\theta_0}^n \Pi\big(\theta : d(\theta, \theta_0) \le \delta_n \varepsilon_{n,0} | \mathbf{x}_n\big) = o(1).$$

The proof of the theorem is given in Section 4.3.

3. Application to sequence parameters and histograms.

3.1. *Sequence parameters*. In this section, we apply Theorem 2.1 and Corollary 2.1 to the case of priors on $(\Theta, \|\cdot\|) = (\ell_2, \|\cdot\|_2)$. We endow the sequence parameter $\theta = (\theta_1, \theta_2, \ldots)$ with independent product priors of the following three types:

(T1) Sieve prior: The hyper-parameter of interest is $\lambda = k$ the truncation: For $2 \le k$,

$$\theta_j \overset{\text{ind}}{\sim} g(\cdot), \qquad \text{if } j \le k, \quad \text{and} \quad \theta_j = 0 \qquad \text{if } j > k.$$

We assume that $\int e^{s_0|x|^{p^*}} g(x)\, dx = a < +\infty$ for some $s_0 > 0$ and $p^* \ge 1$.

(T2) Scale parameter of a Gaussian process prior: let $\tau_j = \tau j^{-\alpha - 1/2}$ and $\lambda = \tau$ with

$$\theta_j \overset{\text{ind}}{\sim} \mathcal{N}(\cdot, \tau_j^2), \qquad 1 \le j \le n, \quad \text{and} \quad \theta_j = 0 \qquad \text{if } j > n.$$

(T3) Rate parameter: same prior as above but this time $\lambda = \alpha$.

REMARK 3.1. Alternatively, one could consider the priors (T2) and (T3) without truncation at level n. The theoretical behaviour of the truncated and nontruncated versions of the priors are very similar, however, from a practical point of view the truncated priors are arguably more natural.

In the hierarchical setup with a prior on k, type (T1) prior has been studied by [1, 25] for generic models, by [22] for density estimation, by [2] for Gaussian white noise model and by [20] for inverse problems. Type (T2) and (T3) priors have been studied with fixed hyper-parameters by [5, 7, 15, 31, 34] or using a prior on $\lambda = \tau$ and $\lambda = \alpha$ in [4, 14, 18, 29]. In the white noise model, using the explicit expressions of the marginal likelihoods and the posterior distributions, [14, 29] have derived posterior concentration rates and described quite precisely the behaviours of the MMLE using type (T3) and (T2) priors, respectively.

In the following, $\Pi(\cdot|k)$ denotes a prior in the form (T1), while $\Pi(\cdot|\tau, \alpha)$ denotes either (T2) or (T3).

3.2. *Deriving $\varepsilon_n(\lambda)$ for priors* (T1)–(T3). It appears from Theorem 2.1 that a key quantity to describe the behaviour of the MMLE is $\varepsilon_n(\lambda)$ defined by (2.1). In the following lemmas, we describe $\varepsilon_n(\lambda) \equiv \varepsilon_n(\lambda, K)$ for any $K > 0$ under the three types of priors above and for true parameters θ_0 belonging to either hyper-rectangles

$$\mathcal{H}_\infty(\beta, L) = \Big\{\theta = (\theta_i)_i : \max_i i^{2\beta+1}\theta_i^2 \le L\Big\}$$

or Sobolev balls

$$\mathcal{S}_\beta(L) = \left\{\theta = (\theta_i)_i : \sum_{i=1}^{\infty} i^{2\beta}\theta_i^2 \le L\right\}.$$

LEMMA 3.1. *Consider priors of type* (T1), *with g positive and continuous on $\mathbb{R}$ and let $\theta_0 \in \ell_2$, then for all $K > 0$ fixed and if $k \in \{2, \dots, \varepsilon n/\log n\}$, with $\varepsilon > 0$ a small enough constant*

$$\varepsilon_n(k)^2 \asymp \sum_{i=k+1}^{\infty} \theta_{0,i}^2 + \frac{k\log n}{n}.$$

Moreover, if $\theta_0 \in \mathcal{H}_\infty(\beta, L) \cup \mathcal{S}_\beta(L)$ with $\beta > 0$ and L any positive constant,

$$\varepsilon_{n,0} \lesssim (n/\log n)^{-\beta/(2\beta+1)}, \tag{3.1}$$

and there exists $\theta_0 \in \mathcal{H}_\infty(\beta, L) \cup \mathcal{S}_\beta(L)$ for which (3.1) is also a lower bound.

The proof of Lemma 3.1 is postponed to Section A.1. We note that it is enough in the above lemma to assume that g is positive and continuous over the set $\{|x| \le M\}$ with $M > 2\|\theta_0\|_\infty$.

REMARK 3.2. One might get rid of the $\log n$ factor in the rate by allowing the density g to depend on n; as in [2, 10], for instance.

Priors of type (T2) and (T3) are Gaussian process priors, thus following [31], let us introduce the so called concentration function

$$\varphi_{\theta_0}(\varepsilon;\alpha,\tau)=\inf_{h\in\mathbb{H}^{\alpha,\tau}:\|h-\theta_0\|_2\le\varepsilon}\|h\|^2_{\mathbb{H}^{\alpha,\tau}}-\log\Pi\big(\|\theta\|_2\le\varepsilon|\alpha,\tau\big), \tag{3.2}$$

where $\mathbb{H}^{\alpha,\tau}$ denotes the Reproducing Kernel Hilbert Space (RKHS) associated to the Gaussian prior $\Pi(\cdot|\alpha,\tau)$

$$\mathbb{H}^{\alpha,\tau}=\left\{\theta=(\theta_i)_{i\in\mathbb{N}};\sum_{i=1}^{n}i^{2\alpha+1}\theta_i^2<+\infty,\theta_i=0\text{ for }i>n\right\}=\mathbb{R}^n,$$

with for all $\theta\in\mathbb{H}^{\alpha,\tau}$

$$\|\theta\|^2_{\mathbb{H}^{\alpha,\tau}}=\tau^{-2}\sum_{i=1}^{n}i^{2\alpha+1}\theta_i^2.$$

Then from Lemma 5.3 of [32],

$$\varphi_{\theta_0}(K\varepsilon;\alpha,\tau)\le-\log\Pi\big(\|\theta-\theta_0\|_2\le K\varepsilon|\alpha,\tau\big)\le\varphi_{\theta_0}(K\varepsilon/2;\alpha,\tau). \tag{3.3}$$

We also have that

$$\tilde{c}_1^{-1}(K\varepsilon/\tau)^{-1/\alpha}\le-\log\Pi\big(\|\theta\|_2\le K\varepsilon|\alpha,\tau\big)\le\tilde{c}_1(K\varepsilon/\tau)^{-1/\alpha}, \tag{3.4}$$

for some $\tilde{c}_1\ge1$; see, for instance, Theorem 4 of [16]. This leads to the following two lemmas.

LEMMA 3.2. *In the case of type* (T2) *and* (T3) *priors, with* $\theta_0\in\mathcal{S}_\beta(L)\cup\mathcal{H}_\infty(\beta,L)$:

- *If* $\beta\neq\alpha+1/2$

$$\begin{aligned}&\frac{\|\theta_0\|_2}{\sqrt{n\tau^2}}\mathbb{1}_{n\tau^2>1}+n^{-\frac{\alpha}{2\alpha+1}}\tau^{\frac{1}{2\alpha+1}}\\&\lesssim\varepsilon_n(\lambda)\lesssim n^{-\frac{\alpha}{2\alpha+1}}\tau^{\frac{1}{2\alpha+1}}+\left(\frac{a(\alpha,\beta)}{n\tau^2}\right)^{\frac{\beta}{2\alpha+1}\wedge\frac12},\end{aligned} \tag{3.5}$$

  where $a(\alpha,\beta)=L^{\frac{\alpha+1/2}{\beta}}/|2\alpha-2\beta+1|$ *if* $\theta_0\in\mathcal{H}_\infty(\beta,L)$ *while* $a(\alpha,\beta)=L^{\frac{\alpha+1/2}{\beta}}$ *if* $\theta_0\in\mathcal{S}_\beta(L)$. *The constants depend possibly on* K *but neither on* n, τ *or* α.
- *If* $\beta=\alpha+1/2$ *then*

$$\begin{aligned}&\frac{\|\theta_0\|_2}{\sqrt{n\tau^2}}\mathbb{1}_{n\tau^2>1}+n^{-\frac{\alpha}{2\alpha+1}}\tau^{\frac{1}{2\alpha+1}}\\&\lesssim\varepsilon_n(\lambda)\lesssim n^{-\frac{\alpha}{2\alpha+1}}\tau^{\frac{1}{2\alpha+1}}+\left(\frac{\log(n\tau^2)}{n\tau^2}\right)^{\frac12}\mathbb{1}_{n\tau^2>1},\end{aligned} \tag{3.6}$$

  where the term $\log(n\tau^2)$ *can be eliminated in the case where* $\theta_0\in\mathcal{S}_\beta(L)$.

LEMMA 3.3. *In the case of prior type* (T2) (*with* $\lambda = \tau$):

- *If* $\alpha + 1/2 < \beta$, *then for all* $\theta_0 \in \mathcal{H}_\infty(\beta, L) \cup \mathcal{S}_\beta(L)$

$$\varepsilon_{n,0} \lesssim n^{-(2\alpha+1)/(4\alpha+4)}, \tag{3.7}$$

  and for all $\theta_0 \in \ell_2(L)$ *satisfying* $\|\theta_0\|_2 \geq c$ *for some fixed* $c > 0$, (3.7) *is also a lower bound.*
- *If* $\alpha + 1/2 > \beta$, *then*

$$\varepsilon_{n,0} \lesssim n^{-\beta/(2\beta+1)}. \tag{3.8}$$

- *If* $\alpha + 1/2 = \beta$, *then*

$$\begin{aligned} &\varepsilon_{n,0} \lesssim n^{-\beta/(2\beta+1)} \log n^{1/(2\beta+1)}, \qquad \text{if } \theta_0 \in \mathcal{H}_\infty(\beta, L), \\ &\varepsilon_{n,0} \lesssim n^{-\beta/(2\beta+1)}, \qquad \text{if } \theta_0 \in \mathcal{S}_\beta(L), \end{aligned} \tag{3.9}$$

  and there exists $\theta_0 \in \mathcal{H}_\infty(\beta, L)$ *for which the upper bound* (3.9) *is also a lower bound.*

In the case of prior type (T3) (*with* $\lambda = \alpha$),

$$\varepsilon_{n,0} \lesssim n^{-\beta/(2\beta+1)}, \qquad \text{if } \theta_0 \in \mathcal{S}_\beta(L) \cup \mathcal{H}_\infty(\beta, L). \tag{3.10}$$

We note that for the scaling prior (T2) in the case $\alpha + 1/2 < \beta$ Lemma 3.3 provides us the sub-optimal rate $\varepsilon_{n,0} \asymp n^{-(2\alpha+1)/(4\alpha+4)}$. Therefore, under condition (2.12) [verified in the supplementary material for prior (T2)] in all three types of examples studied in this paper (white noise, regression and estimation of density models), we get that for all $\theta_0 \neq 0$ with $\alpha + 1/2 < \beta$, the type (T2) prior leads to sub-optimal posterior concentration rates [and in case $\theta_0 \in \mathcal{H}_\infty(\beta, L)$, $\beta = \alpha + 1/2$ as well].

An important tool to derive posterior concentration rates in the case of empirical Bayes procedures is the construction of the change of measure $\psi_{\lambda,\lambda'}$. We present in the following section how these changes of measures can be constructed in the context of priors (T1)–(T3).

3.3. *Change of measure.* In the case of prior (T1), there is no need to construct $\psi_{\lambda,\lambda'}$ due to the discrete nature of the hyper-parameter $\lambda = k$ the truncation threshold.

In the case of prior (T2) if $\tau, \tau' > 0$, then define for all $i \in \mathbb{N}$

$$\psi_{\tau,\tau'}(\theta_i) = \frac{\tau'}{\tau}\theta_i \tag{3.11}$$

so that $\psi_{\tau,\tau'}(\theta) = (\psi_{\tau,\tau'}(\theta_i), i \in \mathbb{N}) = \theta\tau'/\tau$ and if $\theta \sim \Pi(\cdot|\tau, \alpha)$, then $\psi_{\tau,\tau'}(\theta) \sim \Pi(\cdot|\tau', \alpha)$.

Similarly, in the case of type (T3) prior,

$$\psi_{\alpha,\alpha'}(\theta_i) = i^{\alpha-\alpha'}\theta_i \tag{3.12}$$

so that $\psi_{\alpha,\alpha'}(\theta) = (\psi_{\alpha,\alpha'}(\theta_i), i \in \mathbb{N})$ and if $\theta \sim \Pi(\cdot|\tau,\alpha)$, then $\psi_{\alpha,\alpha'}(\theta) \sim \Pi(\cdot|\tau,\alpha')$. Note in particular that if $\alpha' \geq \alpha$ and $\sum_i \theta_i^2 < +\infty$ hold then $\sum_i \psi_{\alpha,\alpha'}(\theta_i)^2 < \infty$. This will turn out to be useful in the sequel.

3.4. *Choice of the hyper-prior.* In this section, we give sufficient conditions on the hyper-priors in the case of the prior distribution (T1)–(T3), such that condition (H1) is satisfied. The proofs are deferred to Section 3 of the Supplementary Material [23].

LEMMA 3.4. *In case of prior* (T1) *assume that* $\theta_0 \in S_\beta(L) \cup \mathcal{H}_\infty(\beta, L)$ *for some* $\beta \geq \beta_1 > \beta_0 \geq 0$. *Then for any hyper-prior satisfying*

$$k^{-c_2 k} \lesssim \tilde{\pi}(k) \lesssim e^{-c_1 k^{1/(1+2\beta_0)}}, \tag{3.13}$$

for some $c_1, c_2 > 0$, *assumption* (H1) *holds*.

Note that the hypergeometric and the Poisson distribution satisfies the above conditions.

LEMMA 3.5. *Consider the prior* (T2) *then for any hyper-prior satisfying*

$$e^{-c_1 \tau^{\frac{2}{1+2\alpha}}} \lesssim \tilde{\pi}(\tau) \lesssim \tau^{-c_2} \qquad \text{for } \tau \geq 1 \text{ with some } c_1 > 0 \text{ and } c_2 > 1 + 1/c_0,$$

$$e^{-c_3 \tau^{-2}} \lesssim \tilde{\pi}(\tau) \lesssim \tau^{c_4} \qquad \text{for } \tau \leq 1 \text{ with some } c_2 > 0 \text{ and } c_4 > 1/c_0 - 1,$$

for some $c_0 > 0$, *assumption* (H1) *holds*.

Note that for instance the inverse gamma and Weibull distributions satisfy this assumption.

REMARK 3.3. To obtain the polynomial upper bound of the hyper-prior densities $\tilde{\pi}(\tau)$ in Lemma 3.5, the set Λ_n is taken to be larger than it is necessary in the empirical Bayes method to achieve adaptive posterior contraction rates; see, for instance, Propositions 3.2 and 3.4. Nevertheless, the conditions on the hyper-entropy are still satisfied, that is, by taking $u_n = e^{-2c_0\bar{c}_0 \bar{w}_n^2 n\varepsilon_{n,0}^2}$ on $\Lambda \setminus \Lambda_0$ and $u_n = n^{-d}$ (for any $d > 0$) on Λ_0 we get that $\log N_n(\Lambda_n) = o(w_n^2 n\varepsilon_{n,0}^2)$ and $\log N_n(\Lambda_0) = o(n\varepsilon_{n,0}^2)$.

LEMMA 3.6. *Consider the prior* (T3) *and assume that* $\theta_0 \in S_\beta(L) \cup \mathcal{H}_\infty(\beta, L)$ *for some* $\beta > \beta_0 > 0$. *Then for any hyper-prior satisfying*

$$e^{-c_2\alpha} \lesssim \tilde{\pi}(\alpha) \lesssim e^{-c_0 \alpha^{1/c_1}}, \qquad \text{for } \alpha > 0$$

and for some $c_0, c_1, c_2 > 0$, *assumption* (H1) *holds*.

In the following sections, we prove that in the Gaussian white noise, regression and density estimation models the MMLE empirical Bayes posterior concentration rate is bounded from above by $M_n\varepsilon_{n,0}$ and from below by $\delta_n\varepsilon_{n,0}$, where $\varepsilon_{n,0}$ is given in Lemma 3.3 under priors (T1)–(T3) and M_n, respectively δ_n, tends to infinity, respectively 0, arbitrary slowly.

3.5. *Application to the nonparametric regression model.* In this section, we show that our results apply to the nonparametric regression model. We consider the fixed design regression problem, where we assume that the observations $\mathbf{x}_n = (x_1, x_2, \ldots, x_n)$ satisfy

$$x_i = f_0(t_i) + Z_i, \qquad i = 1, 2, \ldots, n, \tag{3.14}$$

where $Z_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ random variables (with known σ^2 for simplicity) and $t_i = i/n$.

Let us denote by $\theta_0 = (\theta_{0,1}, \theta_{0,2}, \ldots)$ the Fourier coefficients of the regression function $f_0 \in L_2(M)$: $f_0(t) = \sum_{j=1}^{\infty} \theta_{0,j} e_j(t)$, so that $(e_j(\cdot))_j$ is the Fourier basis. We note that following from Lemma 1.7 in [30] and Parseval's inequality we have that

$$\|f_0\|_2 = \|\theta_0\|_2 = \|f_0\|_n,$$

where $\|f_0\|_n$ denotes the L_2-metric associated to the empirical norm.

First, we deal with the random truncation prior (T1) where applying Theorem 2.1, Corollary 2.1 and Theorem 2.3 combined with Lemma 3.1 we get that both the MMLE empirical Bayes and hierarchical Bayes posteriors are rate adaptive (up to a $\log n$ factor). The following proposition is proved in Section 1.1 of the Supplementary Material [23].

PROPOSITION 3.1. *Assume that $f_0 \in \mathcal{H}_\infty(\beta, L) \cup S_\beta(L)$ and consider a type* (T1) *prior. Let $\Lambda_n = \{2, \ldots, k_n\}$ with $k_n = \varepsilon n/\log n$ for some small enough constant $\varepsilon > 0$. Then, for any M_n tending to infinity and $K > 0$ the MMLE estimator $\hat{k}_n \in \Lambda_0 = \{k : \varepsilon_n(k) \le M_n\varepsilon_{n,0}\}$ with probability going to* 1 *under $P^n_{\theta_0}$, where $\varepsilon_n(k)$ and $\varepsilon_{n,0}$ are given in Lemma* 3.1.

Furthermore, we also have the following contraction rates: for all $0 < \beta_1 \le \beta_2 < +\infty$, uniformly over $\beta \in (\beta_1, \beta_2)$

$$\sup_{f_0 \in \mathcal{H}_\infty(\beta,L)\cup S_\beta(L)} E^n_{f_0} \Pi\big(f : \|f_0 - f\|_2 \ge M_n (n/\log n)^{-\frac{\beta}{2\beta+1}} | \mathbf{x}_n; \hat{k}_n\big) = o(1),$$

$$\sup_{f_0 \in \mathcal{H}_\infty(\beta,L)\cup S_\beta(L)} E^n_{f_0} \Pi\big(f : \|f_0 - f\|_2 \ge M_n (n/\log n)^{-\frac{\beta}{2\beta+1}} | \mathbf{x}_n\big) = o(1),$$

where the latter is satisfied if the hyper prior on k satisfies (3.13).

Finally, we note that the above bounds are sharp in the sense that both the MMLE empirical and the hierarchical Bayes posterior contraction rates are

bounded from below by $\delta_n(n/\log n)^{-\beta/(2\beta+1)}$ *with* $P^n_{f_0}$*-probability tending to one, for any* $\delta_n = o(1)$ *and some* $f_0 \in \mathcal{H}_\infty(\beta, L) \cup S_\beta(L)$.

Next, we consider the priors (T2) and (T3). As a consequence of Theorem 2.1, Corollary 2.1, Theorem 2.3 and Lemma 3.3 we can show that both the hierarchical Bayes and the MMLE empirical Bayes method for the rescaled Gaussian prior (T2) is optimal only in a limited range of regularity classes $\mathcal{S}_\beta(L) \cup \mathcal{H}_\infty(\beta, L)$ satisfying $\beta < \alpha + 1/2$, else the posterior achieves a sub-optimal contraction rate $n^{-(2\alpha+1)/(4\alpha+4)}$. However, by taking the MMLE of the regularity hyper-parameter α or endowing it with a hyper-prior distribution in the Gaussian prior (T3), the posterior achieves the minimax contraction rate $n^{-\beta/(1+2\beta)}$. Similar results were derived in [29] and [14] in the context of the (inverse) Gaussian white noise model using semi-explicit computations. We note that our implicit (and general) approach not just reproduces the previous findings in the direct (noninverse problem) case, but also improves on the posterior contraction rate in case of the prior (T3), where in [14] an extra $\log n$ factor was present.

PROPOSITION 3.2. *Assume that* $f_0 \in \mathcal{S}_\beta(L) \cup \mathcal{H}_\infty(\beta, L)$ *for some* $\beta > 0$ *and consider type* (T2) *and* (T3) *priors with* $\alpha > 0$. *Furthermore, take* $\Lambda_n(\tau) = [n^{-1/(4\alpha)}, n^{\alpha/2}]$ *and* $\Lambda_n(\alpha) = (0, c_0 n^{c_1}]$, *respectively, for some* $c_0, c_1 > 0$. *Then* $\hat{\lambda}_n \in \Lambda_0$ *with* $P^n_{f_0}$*-probability tending to* 1. *Furthermore, both in the case of the MMLE empirical Bayes and hierarchical Bayes approach we have for any* M_n *going to infinity with hyper-priors satisfying* (H1) (*see, for instance, Lemma 3.5 and Lemma 3.6*) *that*:

- *For the multiplicative scaling prior* (T2)
  - *If* $\beta > \alpha + 1/2$, *the posterior concentration rate is bounded from above by*

$$M_n \varepsilon_{n,0} \asymp M_n n^{-(2\alpha+1)/(4\alpha+4)},$$

    and for $\delta_n = o(1)$ *and* $\|f_0\|_2 \geq c$ (*for some positive constant* c) *it is bounded from below by*

$$\delta_n \varepsilon_{n,0} \asymp \delta_n n^{-(2\alpha+1)/(4\alpha+4)}.$$

  - *If* $\beta < \alpha + 1/2$, *the posterior concentration rate is bounded by*

$$M_n \varepsilon_{n,0} \lesssim M_n n^{-\beta/(2\beta+1)},$$

    with an extra $\log n$ *term if* $\beta = \alpha + 1/2$ *and* $f_0 \in \mathcal{H}_\infty(\beta, L)$.
- *For the regularity prior* (T3), *the posterior contraction rate is also*

$$M_n \varepsilon_{n,0} \lesssim M_n n^{-\beta/(2\beta+1)}.$$

Proposition 3.2 is proved in Section 1.2 of the Supplementary Material [23].

REMARK 3.4. In fact, our results are stronger than the minimax results presented in Propositions 3.1 and 3.2. From Theorem 2.1 and Corollary 2.1, it follows that for both the MMLE empirical Bayes and the hierarchical Bayes methods the posterior contracts around the truth for every $\theta_0 \in \Theta$ with rate $M_n \varepsilon_{n,0}(\theta_0)$, which is more informative than a statement on the worst case scenario over some regularity class, that is, the minimax result.

REMARK 3.5. We note that in the case of the Gaussian white noise model the same posterior contraction rate results (both for the empirical Bayes and hierarchical Bayes approaches) hold for the priors (T1)–(T3) as in the nonparametric regression model. The proof of this statement can be easily derived as a special case of the results on the nonparametric regression; see the end of the proofs of Propositions 3.1 and 3.2.

3.6. *Application to density estimation.* In this section, we consider the density estimation problem on $[0, 1]$, that is, the observations $\mathbf{x}_n = (x_1, \ldots, x_n)$ are independent and identically distributed from a distribution with density f with respect to Lebesgue measure. We consider two families of priors on the set of densities $\mathcal{F} = \{f : [0, 1] \to \mathbb{R}^+; \int_0^1 f(x)\,dx = 1\}$. In the first case, we parameterize the densities as

$$
\begin{aligned}
f(x) = f_\theta(x) &= \exp\left(\sum_{j=1}^{\infty} \theta_j \varphi_j(x) - c(\theta)\right), \\
e^{c(\theta)} &= \int_0^1 \exp\left(\sum_{j=1}^{\infty} \theta_j \varphi_j(x)\right) dx,
\end{aligned}
\tag{3.15}
$$

where $(\varphi_j)_{j\in\mathbb{N}}$ forms an orthonormal basis with $\varphi_0 = 1$ and $\theta = (\theta_j)_{j\in\mathbb{N}} \in \ell_2$. Hence, (3.15) can be seen either as a log-linear model or as an infinite dimensional exponential family; see, for instance, [1, 21, 22, 31, 33].

In the second case, we consider random histograms to parameterize $\mathcal{F}$.

3.6.1. *Log-linear model.* We study priors based on the parameterization (3.15) and we assume that the true density has the form $f_0 = f_{\theta_0}$ for some $\theta_0 \in \ell_2$ and throughout the section we will assume that f_0 verifies $\|\log f_0\|_\infty < +\infty$ and that $\theta_0 \in \mathcal{S}_\beta(L)$ for some $L > 0$. We study the MMLE empirical Bayes and hierarchical Bayes methods based on priors of type (T1), (T2) and (T3) in this model. We consider the usual metric in the context of density estimation, namely the Hellinger metric $h(f_1, f_2)^2 = \int_0^1 (\sqrt{f_1}(x) - \sqrt{f_2}(x))^2\,dx$.

First, we consider the type (T1) prior where $\lambda = k$. We show that Theorems 2.1, 2.3 and Corollary 2.1 can be applied so that the MMLE empirical Bayes and hierarchical posterior rates are minimax adaptive over a collection of Sobolev classes.

PROPOSITION 3.3. *Assume that $\theta_0 \in \mathcal{S}_\beta(L)$ with $\beta > 1/2$, consider a type* (T1) *prior, and let $\Lambda_n = \{2, \dots, k_n\}$ with $k_n = k_0\sqrt{n}/(\log n)^3$. Then, for any M_n going to infinity and $K > 0$, if $\hat{k}_n$ is the MMLE over Λ_n, with probability going to* 1 *under $P^n_{\theta_0}$, $\hat{k}_n \in \Lambda_0 = \{k; \varepsilon_n(k) \le M_n\varepsilon_{n,0}\}$, where $\varepsilon_n(k)$ and $\varepsilon_{n,0}$ are given in Lemma* 3.1 *and for all $1/2 < \beta_1 \le \beta_2 < +\infty$*

$$\sup_{\beta\in(\beta_1,\beta_2)} \sup_{\theta_0\in\mathcal{S}_\beta(L)} E^n_{\theta_0}\big\{\Pi\big(h(f_{\theta_0}, f_\theta) \ge M_n(n/\log n)^{-\frac{\beta}{2\beta+1}} |\mathbf{x}_n; \hat{k}_n\big)\big\} = o(1).$$

Similarly, the hierarchical posterior distribution with hyper-prior satisfying the conditions of Lemma 3.4 *also achieves the* (*nearly*) *minimax contraction rate*

$$\sup_{\beta\in(\beta_1,\beta_2)} \sup_{\theta_0\in\mathcal{S}_\beta(L)} E^n_{\theta_0}\big\{\Pi\big(h(f_{\theta_0}, f_\theta) \ge M_n(n/\log n)^{-\frac{\beta}{2\beta+1}} |\mathbf{x}_n\big)\big\} = o(1).$$

Moreover, there exists $\theta_0 \in \mathcal{S}_\beta(L)$ for which $\delta_n(n/\log n)^{-\beta/(2\beta+1)}$ is a lower bound on the posterior concentration rate for both the empirical and the hierarchical Bayes methods.

The proof of Proposition 3.3 is presented in Section 1.3 of the Supplementary Material [23].

We now apply Theorems 2.1, 2.3, and Corollary 2.1 to priors (T2) and (T3) and derive similar concentration rates as in the case of the regression model. Let

$$\bar{\tau}_n = n^{\alpha/2-1/4}, \qquad \underline{\tau}_n = n^{-1/4+1/(8\alpha)}.$$

PROPOSITION 3.4. *Assume that $\theta_0 \in \mathcal{S}_\beta(L)$ with $\beta > 1/2$ and consider a type* (T2) *prior with $\alpha > 1/\sqrt{2}$ and $\Lambda_n = (\underline{\tau}_n, \bar{\tau}_n)$. Then $\hat{\lambda}_n \in \Lambda_0$ with probability going to* 1 *under $P^n_{\theta_0}$ and the same conclusions as in Proposition* 3.2 *hold.*

The constraint $\alpha > 1/\sqrt{2}$ is to ensure that for all $\beta \le \alpha + 1/2$, $\tau = n^{-(\beta-\alpha)/(2\beta+1)}$ which corresponds to the minimizer of $\varepsilon_n(\tau)$ (up to a multiplicative constant) belongs to the set $(\underline{\tau}_n, \bar{\tau}_n)$.

PROPOSITION 3.5. *Assume that $\theta_0 \in \mathcal{S}_\beta(L)$ with $\beta > 1/2$ and consider a type* (T3) *prior with $\alpha > 1/2$ and $\Lambda_n = [1/2 + 1/n^{1/4}, \bar{\lambda}_n]$, with $\bar{\lambda}_n = \log n/(16 \log\log n)$. Then for any M_n going to infinity the MMLE empirical Bayes posterior achieves the minimax contraction rate*

$$M_n\varepsilon_{n,0} \lesssim M_n n^{-\beta/(2\beta+1)}.$$

Furthermore, the hierarchical posterior also achieves the minimax contraction rate for hyper-priors satisfying (H1).

The proofs of Propositions 3.4 and 3.5 are presented in Sections 1.4 and 1.5 of the Supplementary Material [23].

We now consider the second family of priors.

3.6.2. *Random histograms.* In this section, we parameterize $\mathcal{F}$ using piecewise constant functions, as in [6] for instance. In other words, we define

$$f_\theta(x) = k\sum_{j=1}^{k}\theta_j \mathbb{1}_{I_j}, \qquad I_j = \big((j-1)/k, j/k\big], \tag{3.16}$$
$$\sum_{j=1}^{k}\theta_j = 1, \qquad \theta_j \geq 0,$$

and we consider a Dirichlet prior on $\theta = (\theta_1, \ldots, \theta_k)$ with parameter $(\alpha, \ldots, \alpha)$. The hyper-parameter on which maximization is performed is $\lambda = k$, as in the case of the truncation prior (T1). We define the sequence $\varepsilon_n(k)$ in terms of the Hellinger distance, that is, it satisfies (2.1) with $h(f_0, f_\theta)$ replacing $\|\theta - \theta_0\|$.

We then have the following result.

PROPOSITION 3.6. *Assume that f_0 is continuous and bounded from above and below by C_0 and c_0, respectively. If $\Lambda = \{1, \ldots, k_n\}$, with $k_n = O((n/\log n))$ and if $\alpha \leq A$ for some constant A independent on k, then for all $k \in \Lambda$*

$$b(k)^2 + \frac{k\log(n/k)}{n} \lesssim \varepsilon_n(k)^2 \lesssim b(k)^2 + \frac{k\log n}{n}, \tag{3.17}$$

with

$$b(k)^2 = \sum_{j=1}^{k}\int_{I_j}(\sqrt{f_0} - \tilde{\eta}_j k)^2\,dx, \qquad \tilde{\eta}_j = \int_{I_j}\sqrt{f_0}(x)\,dx.$$

Now suppose that $f_0 \in \mathcal{H}_\beta(L)$, i.e. f_0 is Holder with smoothness parameter $\beta > 0$ and radius parameter $L > 0$. The MMLE empirical Bayes posterior achieves the minimax contraction rate (up to a $\log n$ term), that is, for all $M_n \to +\infty$

$$M_n\varepsilon_{n,0} \lesssim M_n(n/\log n)^{-\beta/(2\beta+1)}$$

and

$$E^n_{f_0}\Pi\big(h(f_0, f_\theta) \geq M_n\varepsilon_{n,0}|\mathbf{x}^n_n, \hat{k}\big) = o(1).$$

Equation (3.17) of Proposition 3.6 is proved in Section A.4, while the rest of the proof is given in Section 1.6 of the Supplementary Material [23].

4. Proofs.

4.1. *Proof of Theorem 2.1.* Following from the definition of $\hat{\lambda}_n$ given in (1.1), we have that $m(\mathbf{x}_n|\lambda) \leq m(\mathbf{x}_n|\hat{\lambda}_n)$ for all $\lambda \in \Lambda_n$. Therefore, to prove our statement

it is sufficient to show that with $P_{\theta_0}^n$-probability tending to one we have

$$\sup_{\lambda\in\Lambda_n\setminus\Lambda_0} m(\mathbf{x}_n|\lambda) < m(\mathbf{x}_n|\lambda_0) \le \sup_{\lambda\in\Lambda_0} m(\mathbf{x}_n|\lambda),$$

where λ_0 is some hyper-parameter belonging to Λ_0 (possibly dependent on n).

We proceed in two steps. First, we show that there exists a constant $C > 0$ such that with $P_{\theta_0}^n$-probability tending to one we have

$$m(\mathbf{x}_n|\lambda_0) \ge e^{-Cn\varepsilon_{n,0}^2}. \tag{4.1}$$

Then we complete the proof by showing that for any sequence $w_n' = o(M_n^2 \wedge w_n^2)$ going to infinity

$$P_{\theta_0}^n\Big(\sup_{\lambda\in\Lambda_n\setminus\Lambda_0} m(\mathbf{x}_n|\lambda) > e^{-nw_n'\varepsilon_{n,0}^2}\Big) = o(1). \tag{4.2}$$

We prove the first inequality (4.1) using the standard technique for lower bounds of the likelihood ratio (e.g., Lemma 10 of [12]). Without loss of generality, we can assume that there exists $\lambda \in \Lambda_n$ such that $\varepsilon_n(\lambda) \ge \varepsilon_{n,0}$. Then take an arbitrary $\lambda_0 \in \tilde{\Lambda}_0$ such that $\varepsilon_n(\lambda_0) \le M_1\varepsilon_{n,0}$ for an arbitrary $M_1 > 1$. Then we have from the assumption (B1) and the definition of $\varepsilon_n(\lambda)$ given in (2.2) that with $P_{\theta_0}^n$-probability tending to one the following inequality holds:

$$\begin{aligned} m(\mathbf{x}_n|\lambda_0) &\ge \int_{\theta\in B_n(\theta_0, M_2\varepsilon_n(\lambda_0),2)} e^{\ell_n(\theta)-\ell_n(\theta_0)}\,d\Pi(\theta|\lambda_0) \\ &\ge \Pi\big(B_n\big(\theta_0, M_2\varepsilon_n(\lambda_0), 2\big)|\lambda_0\big)e^{-2n\varepsilon_n^2(\lambda_0)M_2^2} \\ &\ge e^{-(\tilde{c}_0+2M_2^2)M_1 n\varepsilon_{n,0}^2}. \end{aligned} \tag{4.3}$$

We now prove (4.2). Split $\Lambda_n \setminus \Lambda_0$ into balls of size $u_n/2$ and choose for each ball a point in $\Lambda_n \setminus \Lambda_0$. We denote by $(\lambda_i)_{i=1}^{N_n(\Lambda_n\setminus\Lambda_0)}$ these points. Consider the set $\Theta_n(\lambda_i)$ defined in (2.6) and divide it into sieves

$$S_{n,j}^{(i)} = \big\{\theta \in \Theta_n(\lambda_i);\, j\varepsilon_n(\lambda_i)c(\lambda_i) \le d(\theta,\theta_0) \le (j+1)\varepsilon_n(\lambda_i)c(\lambda_i)\big\}.$$

We have following from assumption (2.9) that for all j

$$\log N\big(\zeta j\varepsilon_n(\lambda_i)c(\lambda_i), S_{n,j}^{(i)}, d(\cdot,\cdot)\big) \le c_1 nj^2\varepsilon_n(\lambda_i)^2c(\lambda_i)^2/2 \tag{4.4}$$

and constructing a net of $S_{n,j}^{(i)}$ with radius $\zeta j\varepsilon_n(\lambda_i)c(\lambda_i)$ we have following from assumption (2.7) that there exist tests $\varphi_{n,j}^{(i)}$ satisfying

$$\begin{aligned} E_{\theta_0}^n\big(\varphi_{n,j}^{(i)}\big) &\le e^{-c_1 nj^2\varepsilon_n(\lambda_i)^2c(\lambda_i)^2}, \\ \int_{S_{n,j}^{(i)}} Q_{\lambda_i,n}^{\theta}\big(1-\varphi_{n,j}^{(i)}\big)\,d\Pi(\theta|\lambda_i) &\le e^{-c_1 nj^2\varepsilon_n(\lambda_i)^2c(\lambda_i)^2}\Pi\big(S_{n,j}^{(i)}|\lambda_i\big). \end{aligned} \tag{4.5}$$

Let us take the test $\varphi_{n,i} = \max_j \varphi_{n,j}^{(i)}$ and for convenience introduce the notation $B_n(\lambda) = \Theta_n(\lambda) \cap \{\theta : \|\theta - \theta_0\| \le K\varepsilon_n(\lambda)\}$. Then using the chaining argument, Markov's inequality, Fubini's theorem and (2.7) we get that

$$
\begin{aligned}
&P_{\theta_0}^n\Big(\sup_{\lambda\in\Lambda_n\setminus\Lambda_0} m(\mathbf{x}_n|\lambda) > e^{-nw_n'\varepsilon_{n,0}^2}\Big)\\
&\quad\le \sum_{i=1}^{N_n(\Lambda_n\setminus\Lambda_0)} P_{\theta_0}^n\Big(\sup_{\rho(\lambda_i,\lambda)\le u_n} m(\mathbf{x}_n|\lambda) > e^{-nw_n'\varepsilon_{n,0}^2}\Big)\\
&\quad\le \sum_{i=1}^{N_n(\Lambda_n\setminus\Lambda_0)} E_{\theta_0}^n[\varphi_{n,i}]\\
&\qquad + e^{nw_n'\varepsilon_{n,0}^2}\Bigg\{\sum_{i=1}^{N_n(\Lambda_n\setminus\Lambda_0)} E_{\theta_0}^n\Big(\sup_{\rho(\lambda_i,\lambda)\le u_n}\int_{\psi_{\lambda_i,\lambda}^{-1}\{B_n(\lambda_i)\}} e^{\ell_n(\theta)-\ell_n(\theta_0)}\,d\Pi(\theta|\lambda)\Big)\\
&\qquad + \sum_{i=1}^{N_n(\Lambda_n\setminus\Lambda_0)} E_{\theta_0}^n\Big(\sup_{\rho(\lambda_i,\lambda)\le u_n}\int_{\psi_{\lambda_i,\lambda}^{-1}\{\Theta_n(\lambda_i)\cap B_n(\lambda_i)^c\}} e^{\ell_n(\theta)-\ell_n(\theta_0)}\\
&\qquad \times (1-\varphi_{n,i})\,d\Pi(\theta|\lambda)\Big)\\
&\qquad + \sum_{i=1}^{N_n(\Lambda_n\setminus\Lambda_0)} E_{\theta_0}^n\Big(\sup_{\rho(\lambda_i,\lambda)\le u_n}\int_{\psi_{\lambda_i,\lambda}^{-1}\{\Theta_n(\lambda_i)^c\}} e^{\ell_n(\theta)-\ell_n(\theta_0)}\,d\Pi(\theta|\lambda)\Big)\Bigg\}\\
&\quad\le N_n(\Lambda_n\setminus\Lambda_0)2e^{-c_1 n\inf_i \varepsilon_n(\lambda_i)^2 c(\lambda_i)^2}\\
&\qquad + e^{nw_n'\varepsilon_{n,0}^2}\Bigg\{\sum_{i=1}^{N_n(\Lambda_n\setminus\Lambda_0)}\int_{B_n(\lambda_i)} Q_{\lambda_i,n}^{\theta}(\mathcal{X}_n)\,d\Pi(\theta|\lambda_i)\\
&\qquad + \sum_{i=1}^{N_n(\Lambda_n\setminus\Lambda_0)}\int_{\Theta_n(\lambda_i)\cap B_n(\lambda_i)^c} Q_{\lambda_i,n}^{\theta}(1-\varphi_{n,i})\,d\Pi(\theta|\lambda_i)\\
&\qquad + \sum_{i=1}^{N_n(\Lambda_n\setminus\Lambda_0)}\int_{\Theta_n(\lambda_i)^c} Q_{\lambda_i,n}^{\theta}(\mathcal{X}_n)\,d\Pi(\theta|\lambda_i)\Bigg\}.
\end{aligned}
\tag{4.6}
$$

Next, we deal with each term on the right-hand side of (4.6) separately and show that all of them tend to zero. One can easily see that since $\lambda_i \in \Lambda_n \setminus \Lambda_0$ and following the definition of $c(\lambda_i)$ given below (2.8), we have that

$$
N_n(\Lambda_n \setminus \Lambda_0)e^{-(c_1/2)n\inf_i \varepsilon_n(\lambda_i)^2 c(\lambda_i)^2} \le N_n(\Lambda_n \setminus \Lambda_0)e^{-(c_1/2)w_n^2 n\varepsilon_{n,0}^2} = o(1).
$$

For the second term, we have following from assumption (2.5), the definitions of $\varepsilon_n(\lambda)$ and the set Λ_0 given in (2.2) and (2.3), respectively, that

$$\begin{aligned} &e^{nw_n'\varepsilon_{n,0}^2} \sum_{i=1}^{N_n(\Lambda_n\setminus\Lambda_0)} \int_{B_n(\lambda_i)} Q_{\lambda_i,n}^{\theta}(\mathcal{X}_n)\, d\Pi(\theta|\lambda_i) \\ &\quad\le \sum_{i=1}^{N_n(\Lambda_n\setminus\Lambda_0)} e^{nw_n'\varepsilon_{n,0}^2} e^{o(1)n\varepsilon_n^2(\lambda_i)} \Pi\big(B_n(\lambda_i)|\lambda_i\big) \\ &\quad\le e^{-nM_n^2\varepsilon_{n,0}^2(\tilde{c}_0^{-1}+o(1))} = o(1). \end{aligned}$$

Next, following from (4.5) we have that

$$\begin{aligned} &e^{nw_n'\varepsilon_{n,0}^2} \sum_{i=1}^{N_n(\Lambda_n\setminus\Lambda_0)} \int_{\Theta_n(\lambda_i)\cap B_n(\lambda_i)^c} Q_{\lambda_i,n}^{\theta}(1-\varphi_n)\, d\Pi(\theta|\lambda_i) \\ &\quad\le e^{nw_n'\varepsilon_{n,0}^2} \sum_{i=1}^{N_n(\Lambda_n\setminus\Lambda_0)} e^{-c_1 n\varepsilon_n(\lambda_i)^2 c(\lambda_i)^2} \\ &\quad\le e^{-c_1 nw_n^2\varepsilon_{n,0}^2(1+o(1))} = o(1). \end{aligned}$$

Finally, we have following assumption (2.6) that the fourth term on the right-hand side of (4.6) can be bounded from above by

$$\begin{aligned} e^{nw_n'\varepsilon_{n,0}^2} \sum_{i=1}^{N_n(\Lambda_n\setminus\Lambda_0)} \int_{\Theta_n(\lambda_i)^c} Q_{\lambda_i,n}^{\theta}\big(\mathcal{X}^{(n)}\big)\, d\Pi(\theta|\lambda_i) &\le N_n(\Lambda_n\setminus\Lambda_0) e^{-(w_n^2-w_n')n\varepsilon_{n,0}^2} \\ &\le e^{-nw_n^2\varepsilon_{n,0}^2(1+o(1))} = o(1). \end{aligned}$$

4.2. *Proof of Corollary* 2.1. The proof of Corollary 2.1, follows the same lines of reasoning as Theorem 1 in [8], with the adding remark that

$$m(\mathbf{x}_n|\hat{\lambda}_n) \ge m(\mathbf{x}_n|\lambda), \qquad \forall\lambda\in\Lambda_n,$$

so that no uniform lower bound in the form $\inf_{\lambda\in\Lambda_0} m(\mathbf{x}_n|\lambda)$ is required. We have

$$\begin{aligned} &E_{\theta_0}^n \Pi\big(d(\theta,\theta_0) > MM_n\varepsilon_{n,0}|\mathbf{x}_n;\hat{\lambda}_n\big) \\ &\quad= E_{\theta_0}^n\left(\frac{\int_{d(\theta,\theta_0)>MM_n\varepsilon_{n,0}} e^{\ell_n(\theta)-\ell_n(\theta_0)}\, d\Pi(\theta|\hat{\lambda}_n)}{\int_{\Theta} e^{\ell_n(\theta)-\ell_n(\theta_0)}\, d\Pi(\theta|\hat{\lambda}_n)}\right) \equiv E_{\theta_0}^n\left(\frac{H_n(\hat{\lambda}_n)}{m(\mathbf{x}_n|\hat{\lambda}_n)}\right). \end{aligned}$$

We construct $\varphi_n = \max_{\lambda_i}\max_j \max_l \varphi_n^{(i)}(\theta_{j,l})$, with $(\lambda_i)_{i\le N_n(\Lambda_0)}$ a net of Λ_0 with radius u_n, and for all $j \ge MM_n$, $(\theta_{j,l})_{l\le N_{n,j}}$ a $\zeta j\varepsilon_n(\lambda_i)$ net of $\bar{S}_{n,j} = \{\theta,\, j\varepsilon_{n,0} \le$

$d(\theta,\theta_0) \le (j+1)\varepsilon_{n,0}\} \cap \Theta_n(\lambda_i)$. By assumption (C2), $\log N_{n,j} \le c_1 n j^2 \varepsilon_n^2/2$ and $\log N_n(\Lambda_0) \le c_3 n\varepsilon_{n,0}^2$ (for some $c_3 > 0$). Then we have for any $c_2 > 0$,

$$E_{\theta_0}^n\left(\frac{H_n(\hat{\lambda}_n)}{m(\mathbf{x}_n|\hat{\lambda}_n)}\right) \le P_{\theta_0}^n(\hat{\lambda}_n \notin \Lambda_0) + E_{\theta_0}^n(\varphi_n) + P_{\theta_0}^n\big[m(\mathbf{x}_n|\hat{\lambda}_n) < e^{-c_2 n\varepsilon_{n,0}^2}\big] + e^{c_2 n\varepsilon_{n,0}^2} E_{\theta_0}^n\Big[(1-\varphi_n)\sup_{\lambda\in\Lambda_0} H_n(\lambda)\Big]. \tag{4.7}$$

We assumed that the first term tends to zero (see Theorem 2.1 for verification of this condition in case of MMLE). Furthermore, by construction

$$E_{\theta_0}^n(\varphi_n) \le N_n(\Lambda_0)\sup_i \sum_{j\ge MM_n} e^{c_1 n j^2 \varepsilon_n^2(\lambda_i)/2} e^{-c_1 n j^2 \varepsilon_n^2(\lambda_i)} \lesssim e^{-nc_1 M_n^2 \varepsilon_{n,0}^2/4}.$$

Also,

$$P_{\theta_0}^n\big[m(\mathbf{x}_n|\hat{\lambda}_n) < e^{-c_2 n\varepsilon_{n,0}^2}\big] \le P_{\theta_0}^n\big[m(\mathbf{x}_n|\lambda_0) < e^{-c_2 n\varepsilon_{n,0}^2}\big] = o(1)$$

following from (4.1) with $c_2 \ge c_3 + M_1(\tilde{c}_0 + 2M_2^2 + 2)$. The control of the last term of (4.7) follows from the proof of Theorem 1 of [8].

4.3. *Proof of Theorem 2.3.* As a first step for notational convenience, let us denote by B_n^c the sets $\{\theta : d(\theta,\theta_0) \ge MM_n\varepsilon_{n,0}\}$ or $\{\theta : d(\theta,\theta_0) \le \delta_n\varepsilon_{n,0}\}$

$$\begin{aligned}\Pi(B_n^c|\mathbf{x}_n) &= \int_{\Lambda_0(M_n)} \Pi(B_n^c|\mathbf{x}_n;\lambda)\tilde{\pi}(\lambda|\mathbf{x}_n)\,d\lambda \\ &\quad + \int_{\Lambda_0(M_n)^c} \Pi(B_n^c|\mathbf{x}_n;\lambda)\tilde{\pi}(\lambda|\mathbf{x}_n)\,d\lambda \\ &\le \sup_{\lambda\in\Lambda_0(M_n)} \Pi(B_n^c|\mathbf{x}_n;\lambda) + \int_{\Lambda_0(M_n)^c} \tilde{\pi}(\lambda|\mathbf{x}_n)\,d\lambda.\end{aligned} \tag{4.8}$$

Then from the proofs of Theorem 1 of [8] and Theorem 2.2 follows that the expected value of the first term on the right-hand side of the preceding display tends to zero. We note that assumption (H2) is needed to deal with the denominator in the posterior, unlike in Corollary 2.1, where weaker assumptions were sufficient following from the definition of the maximum marginal likelihood estimator $\hat{\lambda}_n$.

Hence, it remained to deal with the second term on the right-hand side of (4.8). The hyper-posterior takes the form

$$\pi(\lambda|\mathbf{x}_n) \propto m(\mathbf{x}_n|\lambda)\tilde{\pi}(\lambda)$$

and from the proof of Theorem 1 of [8] (pages 10–11) and (4.6) in the proof of Theorem 2.1 we have with $P_{\theta_0}^n$-probability tending to one that

$$\begin{aligned} m(\mathbf{x}_n|\lambda) &\ge e^{-(\tilde{c}_0+2M_2^2)\tilde{w}_n^2 n\varepsilon_{n,0}^2} \qquad \text{for } \lambda \in \Lambda_0(\tilde{w}_n), \quad \text{and} \\ m(\mathbf{x}_n|\lambda) &\le e^{-w_n' n\varepsilon_{n,0}^2} \qquad \text{for } \lambda \in \Lambda_n \setminus \Lambda_0(M_n), \end{aligned}$$

for any $w_n' = o(M_n^2 \wedge w_n^2)$, hence there exists w_n', which also satisfies $\tilde{w}_n = o(w_n')$. Therefore, with $P_{\theta_0}^n$-probability tending to one we also have that

$$\int_{\Lambda_n \setminus \Lambda_0(M_n)} \pi(\lambda|\mathbf{x}_n)\,d\lambda \le \frac{e^{-w_n' n\varepsilon_{n,0}^2}}{e^{-(\tilde{c}_0+2M_2^2)\tilde{w}_n^2 n\varepsilon_{n,0}^2}\int_{\Lambda_0(\tilde{w}_n)}\tilde{\pi}(\lambda)\,d\lambda} = o(1).$$

Finally, similarly to the preceding display we have that

$$\begin{aligned} E_{\theta_0}^n \int_{\Lambda\setminus\Lambda_n} \pi(\lambda|\mathbf{x}_n)\,d\lambda &\le \frac{\int_{\Lambda\setminus\Lambda_n} E_{\theta_0}^n m(\mathbf{x}_n|\lambda)\tilde{\pi}(\lambda)\,d\lambda}{e^{-(\tilde{c}_0+2M_2^2)\tilde{w}_n^2 n\varepsilon_{n,0}^2}\int_{\Lambda_0(\tilde{w}_n)}\tilde{\pi}(\lambda)\,d\lambda} + o(1) \\ &\lesssim e^{(\tilde{c}_0+2M_2^2+1)\tilde{w}_n^2 n\varepsilon_{n,0}^2}\int_{\Lambda\setminus\Lambda_n}\tilde{\pi}(\lambda)\,d\lambda + o(1) = o(1), \end{aligned}$$

completing the proof.

APPENDIX: PROOF OF THE LEMMAS ABOUT THE RATE $\varepsilon_n(\lambda)$

A.1. Proof of Lemma 3.1. We have $\|\theta-\theta_0\|_2^2 = \sum_{j=1}^k(\theta_j-\theta_{0,j})^2 + \sum_{j=k+1}^\infty \theta_{0,j}^2$ so that $\|\theta-\theta_0\|_2^2 \le K^2\varepsilon^2$ if and only if $\sum_{j=1}^k(\theta_j-\theta_{0,j})^2 \equiv \|\theta-\theta_{0,[k]}\|_2^2 \le \delta^2$, with $\delta^2 = K^2\varepsilon^2 - \sum_{j=k+1}^\infty \theta_{0,j}^2$, and $\theta_{0,[k]} = (\theta_{0,j},\, j\le k)$. Then

$$\begin{aligned} \int_{\theta\in\mathbb{R}^k} g(\theta)\mathbb{1}\{\|\theta-\theta_{0,[k]}\|_2 \le \delta\}\,d\theta &\le \|g\|_\infty^k \frac{\pi^{k/2}\delta^k}{\Gamma(k/2+1)} \\ &\ge \underline{g}^k \frac{\pi^{k/2}\delta^k}{\Gamma(k/2+1)} \end{aligned}$$

with $\underline{g} = \inf_{B_k(\delta)} g(x)$ where $B_k(\delta) = \{x; \min_{i\le k}|x-\theta_{0,i}| \le \delta\}$. The Stirling formula implies that both the lower and upper bounds have the form $\exp\{k\log(C\delta/\sqrt{k})\}$ and since $\delta = o(1)$ this is equivalent to $\exp\{k\log(\delta/\sqrt{k})(1+o(1))\}$. We thus have

$$\varepsilon_n(k) > \Bigg(\sum_{i>k}\theta_{0,i}^2\Bigg)^{1/2}\Big/K \quad\text{and}\quad n\varepsilon_n^2(k) = k\log(\sqrt{k}/s_n)\big(1+o(1)\big),$$

with $s_n^2 = K^2\varepsilon_n^2(k) - \sum_{j=k+1}^\infty \theta_{0,j}^2$. In other words, $s_n > 0$ and

$$s_n^2 + \sum_{j=k+1}^\infty \theta_{0,j}^2 = \frac{K^2k}{n}\log\Big(\frac{\sqrt{k}}{s_n}\Big)\big(1+o(1)\big). \tag{A.1}$$

Also, if $\sum_{j=k+1}^\infty \theta_{0,j}^2 = o(k\log n/n)$, then (A.1) implies that

$$s_n^2 = \frac{K^2k}{n}\log\Big(\frac{\sqrt{k}}{s_n}\Big)\big(1+o(1)\big) \quad\Rightarrow\quad s_n^2 = \frac{K^2k}{2n}\log\big(2n/K^2\big)\big(1+o(1)\big).$$

Now take $\theta_0 \in \mathcal{H}_\infty(\beta, L) \cup \mathcal{S}_\beta(L)$, since $\sum_{i>k} \theta_{0,i}^2 \lesssim k^{-2\beta}$, choosing $k = \lfloor (n/\log n)^{1/(2\beta+1)} \rfloor$ leads to $\varepsilon_{n,0} \lesssim (n/\log n)^{-\beta/(2\beta+1)}$. Finally, considering $\theta_{0,i}^2 = (1+i)^{-2\beta-1}$ for $\mathcal{H}_\infty(\beta, L)$ implies that this is also a lower bound in this case. Furthermore, for all $\delta_n = o(1/M_n)$ and for all k such that

$$k^{-2\beta} + \frac{k \log n}{n} \leq M_n^2 (n/\log n)^{-2\beta/(2\beta+1)} \quad \Rightarrow \quad k \lesssim M_n^2 (n/\log n)^{1/(2\beta+1)}$$

and $\delta_n^2(k^{-2\beta} + k \log n/n) = o(k^{-2\beta}) = o(\sum_{i>k} \theta_{0,i}^2)$ so that

$$\Pi\big(\|\theta - \theta_0\|_2 \leq \delta_n \varepsilon_n(k) | k\big) = 0$$

and condition (2.12) is verified.

A.2. Proof of Lemma 3.2. We need to study

$$\inf_{h \in \mathbb{H}^{\alpha,\tau} : \|h - \theta_0\|_2 \leq \varepsilon_n} \|h\|_{\mathbb{H}^{\alpha,\tau}}^2 .$$

Let us distinguish three cases $\beta > \alpha + 1/2$, $\beta < \alpha + 1/2$ and $\beta = \alpha + 1/2$, and note that the following computations hold both for the truncated and nontruncated versions of the priors (T2) and (T3).

In the case $\beta > \alpha + 1/2$ and if $\theta_{0,i}^2 \leq L i^{-2\beta-1}$ for all i, then

$$\inf_{h \in \mathbb{H}^{\alpha,\tau} : \|h - \theta_0\|_2 \leq \varepsilon} \|h\|_{\mathbb{H}^{\alpha,\tau}}^2 \leq \tau^{-2} L \sum_{i=1}^{\infty} i^{2\alpha - 2\beta} \lesssim \frac{L \tau^{-2}}{\beta - \alpha - 1/2}$$

while when $\theta_0 \in \mathcal{S}_\beta(L)$ $\inf_{h \in \mathbb{H}^{\alpha,\tau} : \|h - \theta_0\|_2 \leq \varepsilon} \|h\|_{\mathbb{H}^{\alpha,\tau}}^2 \leq \tau^{-2} L$. Also,

$$n^{-\frac{\alpha}{2\alpha+1}} \tau^{\frac{1}{2\alpha+1}} \lesssim \varepsilon_n(\alpha, \tau) \lesssim n^{-\frac{\alpha}{2\alpha+1}} \tau^{\frac{1}{2\alpha+1}} + \left(\frac{1}{n \tau^2 (\beta - \alpha - 1/2)} \right)^{1/2}$$

if $\theta_0 \in \mathcal{H}_\infty(\beta, L)$, while

$$n^{-\frac{\alpha}{2\alpha+1}} \tau^{\frac{1}{2\alpha+1}} \lesssim \varepsilon_n(\alpha, \tau) \lesssim n^{-\frac{\alpha}{2\alpha+1}} \tau^{\frac{1}{2\alpha+1}} + \left(\frac{1}{n \tau^2} \right)^{1/2}$$

if $\theta_0 \in \mathcal{S}_\beta(L)$. Now, if $0 < \beta < \alpha + 1/2$, with $\theta_0 \in \mathcal{H}_\infty(\beta, L)$

$$\inf_{h \in \mathbb{H}^{\alpha,\tau} : \|h - \theta_0\|_2 \leq \varepsilon} \|h\|_{\mathbb{H}^{\alpha,\tau}}^2 \leq \tau^{-2} L \sum_{i=1}^{(\frac{L}{2\beta})^{\frac{1}{2\beta}} \varepsilon_n^{-\frac{1}{\beta}}} i^{2\alpha - 2\beta} \lesssim L^{\frac{2\alpha+1}{2\beta}} \frac{\tau^{-2} \varepsilon^{-\frac{2\alpha-2\beta+1}{\beta}}}{2\alpha + 1 - 2\beta}$$

and when $\theta_0 \in \mathcal{S}_\beta(L)$

$$\inf_{h \in \mathbb{H}^{\alpha,\tau} : \|h - \theta_0\|_2 \leq \varepsilon} \|h\|_{\mathbb{H}^{\alpha,\tau}}^2 \leq \tau^{-2} L^{\frac{2\alpha+1}{2\beta}} \varepsilon^{-(2\alpha-2\beta+1)/\beta}.$$

If $\beta = \alpha + 1/2$, the same result holds for $\theta_0 \in \mathcal{S}_\beta(L)$, but it becomes

$$\inf_{h \in \mathbb{H}^{\alpha,\tau} : \|h-\theta_0\|_2 \le \varepsilon} \|h\|^2_{\mathbb{H}^{\alpha,\tau}} \le \frac{\tau^{-2} L}{\beta} \big|\log(\varepsilon)\big| \big(1 + o(1)\big)$$

when $\theta_0 \in \mathcal{H}_\infty(\beta, L)$. These lead to the upper bound in (3.5) and (3.6).

Furthermore, for every $\theta_0 \in \mathcal{S}_\beta(L) \cup \mathcal{H}_\infty(\beta, L)$ satisfying $\|\theta_0\|_2 > 2\varepsilon$, when $\|h - \theta_0\|_2 \le \varepsilon$ then $\|h\|_2 > \|\theta_0\|_2/2$, hence

$$\inf_{h \in \mathbb{H}^{\alpha,\tau} : \|h-\theta_0\|_2 \le \varepsilon} \|h\|^2_{\mathbb{H}^{\alpha,\tau}} \ge \tau^{-2} \inf_{h \in \mathbb{H}^{\alpha,\tau} : \|h-\theta_0\|_2 \le \varepsilon} \|h\|_2^2 \gtrsim \|\theta_0\|_2^2 \tau^{-2}.$$

Hence, if $\|\theta_0\|_2 > 2\varepsilon_n(\alpha, \tau)$ for $a(\alpha, \beta)$ defined in Lemma 3.2,

$$\varepsilon_n(\lambda) \gtrsim \frac{\|\theta_0\|_2}{\sqrt{n\tau^2}} + n^{-\alpha/(2\alpha+1)} \tau^{1/(2\alpha+1)}$$

and for all $\tau^2 n$ lower bounded by a positive constant the above inequality remains valid when $\|\theta_0\|_2 \le 2\varepsilon_n(\lambda)$, providing us the lower bound in (3.5) and (3.6).

A.3. Proof of Lemma 3.3. The proof is based on minimizing the upper bounds obtained in Lemmas 3.1 and 3.2.

- First, consider $\lambda = \tau$. When $\beta > \alpha + 1/2$, note that for all $\tau \ge n^{-1/(4\alpha+4)}$

$$\left(\frac{1}{n\tau^2}\right)^{1/2} \lesssim n^{-\alpha/(2\alpha+1)} \tau^{1/(2\alpha+1)}$$

so that $\varepsilon_n(\tau) \asymp n^{-\alpha/(2\alpha+1)} \tau^{1/(2\alpha+1)}$ which is minimized at $\tau \asymp n^{-1/(4\alpha+4)}$ so that (3.7) is verified. Following from (3.5), the lower bound is obtained with every $\|\theta_0\|_2 \ge c > 0$, for any arbitrary positive constant c. Indeed in this case, we have $\varepsilon_n(\tau) \gtrsim (n\tau^2)^{-1/2}$ which implies that the lower bound is the same as the upper bound (3.7). Furthermore, we note that the lower bound

$$\varepsilon_{n,0} \gtrsim n^{-(2\alpha+1)/(4\alpha+4)} \tag{A.2}$$

holds for every $\theta_0 \ne 0$ (and large enough n). Therefore, we also have for every τ_0 satisfying $\varepsilon_n(\tau_0) \lesssim \varepsilon_{n,0}$ that $\tau_0 \gtrsim n^{-1/(4\alpha+4)}$.

When $\beta < \alpha + 1/2$, we have for all $\tau \ge n^{-(\beta-\alpha)/(2\beta+1)}$ that $\varepsilon_n(\tau) \asymp n^{-\frac{\alpha}{2\alpha+1}} \times \tau^{\frac{1}{2\alpha+1}}$, which is minimized at $\tau \asymp n^{-(\beta-\alpha)/(2\beta+1)}$, leading to the upper bound (3.8). The upper bound is obtained choosing for instance $\theta_{0,i} = \sqrt{L} i^{-\beta-1/2}$ for all $i \le K_n$, for some sequence K_n going to infinity, so that

$$\begin{aligned}
\inf_{\|h-\theta\|_2 \le \varepsilon_n(\tau)} \|h\|^2_{\mathbb{H}^{\alpha,\tau}} &\ge \tau^{-2} \sum_{i=1}^{K_n} i^{2\alpha+1} \big[\theta_{0,i}^2 - 2\theta_{0,i}(\theta_{0,i} - h_i)\big] \\
&\gtrsim \tau^{-2} \big(L K_n^{2\alpha-2\beta+1} - 2\sqrt{L}\varepsilon_n(\tau) K_n^{2\alpha-\beta+1}\big) \\
&\gtrsim \tau^{-2} K_n^{2\alpha-2\beta+1}
\end{aligned}$$

and $K_n \leq k_0 \varepsilon_n(\tau)^{-1/\beta}$. This leads to $\varepsilon_n(\tau) \geq (n\tau^2)^{-\beta/(2\alpha+1)}$, with an extra $\log n$ term in the case $\alpha + 1/2 = \beta$ and $\theta_0 \in \mathcal{H}_\infty(\beta, L)$ so that the lower bound is of the same order as the upper bound (3.6) which in terms implies that the lower bound is the same as the upper bound (3.8).

We now consider the case $\lambda = \alpha$, then we have a generic upper bound for $\varepsilon_n(\alpha)$ in the form $n^{-(\alpha\wedge\beta)/(2\alpha+1)}$ following from (3.5) and $\theta_0 \in \mathcal{H}_\infty(\beta, L) \cup \mathcal{S}_\beta(L)$, while the lower bound is a multiple of $n^{-\alpha/(2\alpha+1)}$. We thus have $\varepsilon_{n,0} \lesssim n^{-\beta/(2\beta+1)}$ for all $\theta_0 \in \mathcal{H}_\infty(\beta, L) \cup S_\beta(L)$ and the constant depends only on β and L.

A.4. Proof of equation (3.17) in Proposition 3.6. We prove the first part of proposition, namely the bounds on $\varepsilon_n(k)$. Denote by g_0 the function

$$g_0(x) = k \sum_{j=1}^{k} \tilde{\eta}_j \mathbb{1}_{I_j}(x),$$

then g_0 is the projection of $\sqrt{f_0}$ on the set of piecewise constant functions on a k regular grid and for any $\theta \in \mathcal{S}_k$ the k-dimensional simplex,

$$h^2(f_0, f_\theta) = h^2\big(f_0, g_0^2\big) + \sum_{j=1}^{k} (\sqrt{\theta_j} - \tilde{\eta}_j \sqrt{k})^2 \geq h^2\big(f_0, g_0^2\big) = b(k)^2.$$

Define $\bar{\theta}_{j,k} = (\tilde{\eta}_j \sqrt{k})^2 / \sum_l \tilde{\eta}_l^2 k$ and for some $v_n = o(1)$ consider $\theta = (\theta_1, \ldots, \theta_k) \in \mathcal{S}_k$ satisfying $|\theta_j - \bar{\theta}_{j,k}| \leq \bar{\theta}_{j,k} v_n$ for $j \leq k-1$. Then $|\theta_k - \bar{\theta}_{k,k}| \leq \sum_{j=1}^{k-1} \bar{\theta}_{j,k} v_n \leq v_n$. Note that $b(k)^2 = 1 - \sum_{j=1}^{k} \tilde{\eta}_j^2 k$, so that

$$\begin{aligned}
\sum_{j=1}^{k} (\sqrt{\theta_j} - \tilde{\eta}_j \sqrt{k})^2 &= \sum_{j=1}^{k} \Bigg(\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}} \sqrt{\sum_l \tilde{\eta}_l^2 k} \Bigg)^2 \\
&\leq 2 \sum_{j=1}^{k} (\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}})^2 + 2 \sum_{j=1}^{k} \bar{\theta}_{j,k} \Bigg(\sqrt{\sum_l \tilde{\eta}_l^2 k} - 1 \Bigg)^2 \\
&\leq 2 v_n^2 + 2 b(k)^2,
\end{aligned}$$

which implies that for such θ, $h^2(f_0, f_\theta) \leq 3b(k)^2 + 2v_n^2$. Since $c_0 \leq f_0 \leq C_0$, $c_0/k \leq \bar{\theta}_{j,k} \leq C_0/k$ and we also have, as in the proof of Lemma 6.1 of [11], that if $v_n \leq c_0/(2k)$, then $v_n \leq \bar{\theta}_{k,k}/2$ and

$$\begin{aligned}
\pi\big(|\theta_j - \bar{\theta}_{j,k}| \leq \bar{\theta}_{j,k} v_n, \forall j \leq k-1\big) &\gtrsim \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k} \bar{\theta}_{k,k}^{\alpha-1} \prod_{j \leq k-1} \int_{\bar{\theta}_{j,k}(1-v_n)}^{\bar{\theta}_{j,k}(1+v_n)} x^{\alpha-1} \, dx \\
&\gtrsim \frac{(C_1 v_n)^k \Gamma(k\alpha)}{(\alpha\Gamma(\alpha))^{k-1} \Gamma(\alpha)} \prod_{j \leq k-1} \bar{\theta}_{j,k}^{\alpha} \\
&\gtrsim \frac{(C_2 v_n)^k \Gamma(k\alpha) k^{-k\alpha}}{(\alpha\Gamma(\alpha))^{k-1} \Gamma(\alpha)},
\end{aligned}$$

for some constant $C_1, C_2 > 0$. Since $\alpha \le A$, if $v_n = n^{-h}$ for some $h > 0$,

$$\pi\big(|\theta_j - \bar{\theta}_{j,k}| \le \bar{\theta}_{j,k} v_n, \forall j \le k-1\big) \gtrsim e^{-ck\log n},$$

which implies that for all k such that $b(k)^2 \lesssim k\log n/n$ we have $\varepsilon_n(k)^2 \lesssim b(k)^2 + k\log n/n$. We now bound from below $\varepsilon_n(k)$. Since $h^2(f_0, f_\theta) = b(k)^2 + \sum_{j=1}^k (\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}}\sqrt{1-b(k)^2})^2$, on the set $h^2(f_0, f_\theta) \le \varepsilon_n^2$, $b(k)^2 \le \varepsilon_n^2$ and $\sum_{j=1}^k (\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}}\sqrt{1-b(k)^2})^2 \le \varepsilon_n^2$. Using elementary algebra and the Cauchy–Schwarz inequality, we have if ε_n is small, $b(k)$ is small and

$$\begin{aligned}
&\sum_{j=1}^k \big(\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}}\sqrt{1-b(k)^2}\big)^2 \\
&\quad \ge \sum_{j=1}^k (\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}})^2 + \frac{b^4(k)}{4} - 2b^2(k)\sqrt{\sum_{j=1}^k (\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}})^2} \\
&\quad = \left(\sqrt{\sum_{j=1}^k (\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}})^2} - \frac{b(k)^2}{2}\right)^2.
\end{aligned}$$

Over the set $\sum_{j=1}^k (\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}})^2 \ge \varepsilon_n^2/2$, then

$$\sqrt{\sum_{j=1}^k (\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}})^2} \ge b(k)^2$$

and

$$\sum_{j=1}^k \big(\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}}\sqrt{1-b(k)^2}\big)^2 \ge \frac{1}{4}\sum_{j=1}^k (\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}})^2,$$

so that if $h(f_0, f_\theta) \le \varepsilon_n$ small enough, then

$$h^2(f_0, f_\theta) \ge b(k)^2 + \frac{1}{4}\sum_{j=1}^k (\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}})^2.$$

Hence,

$$\begin{aligned}
&\Pi\big\{h^2(f_0, f_\theta) \le K\varepsilon_n(k)^2\big\} \\
&\quad \le \Pi\left(\sum_{j=1}^k (\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}})^2 \le K\varepsilon_n(k)^2 - b(k)^2\right),
\end{aligned}$$

with $b(k)^2 < K\varepsilon_n(k)^2$. Set $s_n^2 = K\varepsilon_n(k)^2 - b(k)^2$. On the set

$$\sum_{j=1}^{k}(\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}})^2 \leq s_n^2,$$

we split $\{1,\dots,k-1\}$ into $|\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}}| \leq 1/\sqrt{k}$ and $|\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}}| > 1/\sqrt{k}$. The cardinality of the latter is bounded from above by $s_n^2 k$. Moreover, if $|\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}}| \leq 1/\sqrt{k}$ then by triangle inequality $\sqrt{\theta_j} \lesssim 1/\sqrt{k}$ else $\sqrt{\theta_j} \lesssim s_n$. We have

$$\begin{aligned}
&\Pi\left(\sum_{j=1}^{k}(\theta_j^{1/2} - \bar{\theta}_{j,k}^{1/2})^2 \leq s_n^2\right) \\
&\quad\leq \frac{\pi^{k/2}\Gamma(\alpha k)s_n^k}{\Gamma(\alpha)^k\Gamma(k/2+1)} \sum_{l=0}^{\lfloor s_n^2 k\rfloor} \binom{k}{l} s_n^{(2\alpha-1)l} k^{-(k-l)(\alpha-1/2)} \\
&\quad\leq \frac{\pi^{\frac{k}{2}}\Gamma(\alpha k)s_n^k}{\Gamma(\alpha)^k\Gamma(k/2+1)}\Bigg(k^{-k(\alpha-1/2)} \\
&\qquad + \sum_{l\leq s_n^2 k} Ce^{l\log(k)+2l-(k-l)(\alpha-1/2)\log(k)+2l(\alpha-1/2)\log(s_n)}\Bigg) \\
&\quad\lesssim \exp\Bigg\{\alpha k\log(k) - k\log\Gamma(\alpha) - \frac{k}{2}\log(k) + k\log(s_n) \\
&\qquad - k\left(\alpha - \frac{1}{2}\right)\log k + O(k)\Bigg\} \\
&\quad\lesssim \exp\big(k\log(s_n) + O(k)\big)
\end{aligned}$$

if $\alpha \geq 1/2$. If $\alpha < 1/2$, for each θ split $\{1,\dots,k-1\}$ into the set S of indices where $\theta_i \geq \rho_n/k$ and its complement, with $\rho_n = o(1)$. The number of indices such that $\theta_i < \rho_n/k$ is bounded by $O(s_n^2 k)$ on the set $\sum_{j=1}^{k}(\sqrt{\theta_j} - \sqrt{\bar{\theta}_{j,k}})^2 \leq s_n^2$, so that

$$\begin{aligned}
&\Pi\left(\sum_{j=1}^{k}(\sqrt{\theta} - \sqrt{\bar{\theta}_{j,k}})^2 \leq s_n^2\right) \\
&\quad\leq \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k} \\
&\qquad\times \sum_{S\subset\{1,\dots,k\}} \int_{\sum_{i\in S}(\theta_i^{1/2} - \bar{\theta}_{i,k}^{1/2})^2 \leq s_n^2} \mathbb{1}_{\substack{\forall i\in S \\ \theta_i \geq \frac{\rho_n}{k}}} \prod_{i\in S}\theta_i^{\alpha-1}\,d\theta_i \int \mathbb{1}_{\substack{\forall i\in S^c \\ \theta_i < \frac{\rho_n}{k}}} \prod_{i\in S^c}\theta_i^{\alpha-1}\,d\theta_i \\
&\quad\leq \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k}
\end{aligned}$$

$$\times \sum_{S\subset\{1,\dots,k\}} \left(\int_{\sum_{i\in S}(u_i-\bar\theta_{i,k}^{1/2})^2\le s_n^2} \mathbb{1}_{\substack{\forall i\in S\\ u_i\ge(\frac{\rho_n}{k})^{\frac12}}} \prod_{i\in S} u_i^{2\alpha-1}\,du_i \right) \left(\frac{\rho_n}{k}\right)^{|S^c|\alpha} \alpha^{-|S^c|}$$

$$\le \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k} \sum_{l\ge k(1-s_n^2)} \left(\frac{\rho_n}{k}\right)^{(k-l)\alpha} \alpha^{-(k-l)} \left(\frac{\rho_n}{k}\right)^{l(\alpha-1/2)} \frac{\sqrt{\pi}^l s_n^l}{\Gamma(l/2+1)} \binom{k}{l}$$

$$\le \frac{\Gamma(k\alpha)}{(\alpha\Gamma(\alpha))^k} (\rho_n/k)^{k\alpha} \sum_{l\ge k(1-s_n^2)}^{k} \alpha^l e^{l\log(\frac{k^{1/2}Cs_n}{\sqrt{l\rho_n}})+k\log k-l\log l-(k-l)\log(k-l)+O(k)}$$

$$\le \exp\{k\alpha\log(\rho_n)+k\log(s_n/\sqrt{\rho_n})+O(k)\} \le e^{k\log s_n-k(1/2-\alpha)\log\rho_n+O(k)}.$$

Hence, choosing $|\log\rho_n| = o(|\log s_n|)$ leads to

$$\Pi\left(\sum_{j=1}^{k}(\sqrt{\theta}-\sqrt{\bar\theta_{j,k}})^2 \le s_n^2\right) \le e^{k(1+o(1))\log s_n},$$

so that $s_n^2|\log s_n| \ge k/n$ and $s_n^2 \gtrsim k/n\log(n/k)$.

Acknowledgements. The authors would like to thank the Associate Editor and the referees for their useful comments which lead to an improved version of the manuscript.

SUPPLEMENTARY MATERIAL

Supplement to "Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator" (DOI: 10.1214/16-AOS1469SUPP; .pdf). This is the supplementary material associated to the present paper. We provide here the proofs of Propositions 3.1–3.6, together with some technical Lemmas used in the context of priors (T2) and (T3) and some technical Lemmas used in the study of the hierarchical Bayes posteriors. Finally some Lemmas used in the regression and density estimation problems are given.

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] BABENKO, A. and BELITSER, E. (2010). Oracle convergence rate of posterior under projection prior and Bayesian model selection. *Math. Methods Statist.* **19** 219–245. MR2742927

[3] BELITSER, E. and ENIKEEVA, F. (2008). Empirical Bayesian test of the smoothness. *Math. Methods Statist.* **17** 1–18. MR2400361

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

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

[6] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** 2353–2383. MR3405597

[7] COX, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923. MR1232525

[8] DONNET, S., RIVOIRARD, V., ROUSSEAU, J. and SCRICCIOLO, C. (2014). Posterior concentration rates for empirical Bayes procedures, with applications to Dirichlet Process mixtures. Available at arXiv:1406.4406v1.

[9] FLORENS, J.-P. and SIMONI, A. (2012). Regularized posteriors in linear ill-posed inverse problems. *Scand. J. Stat.* **39** 214–235. MR2927022

[10] GAO, C. and ZHOU, H. H. (2016). Rate exact Bayesian adaptation with modified block priors. *Ann. Statist.* **44** 318–345. MR3449770

[11] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[12] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274

[13] JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135

[14] KNAPIK, B. T., SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2016). Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Related Fields* **164** 771–813. MR3477780

[15] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39** 2626–2657. MR2906881

[16] KUELBS, J. and LI, W. V. (1993). Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal.* **116** 133–157. MR1237989

[17] LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York. MR1639875

[18] LIAN, H. (2014). Adaptive rates of contraction of posterior distributions in Bayesian wavelet regression. *J. Statist. Plann. Inference* **145** 92–101. MR3125351

[19] PETRONE, S., ROUSSEAU, J. and SCRICCIOLO, C. (2014). Bayes and empirical Bayes: Do they merge? *Biometrika* **101** 285–302. MR3215348

[20] RAY, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat.* **7** 2516–2549. MR3117105

[21] RIVOIRARD, V. and ROUSSEAU, J. (2012). Bernstein–von Mises theorem for linear functionals of the density. *Ann. Statist.* **40** 1489–1523. MR3015033

[22] RIVOIRARD, V. and ROUSSEAU, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Anal.* **7** 311–333. MR2934953

[23] ROUSSEAU, J. and SZABO, B. (2016). Supplement to "Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator." DOI:10.1214/16-AOS1469SUPP.

[24] SERRA, P. and KRIVOBOKOVA, T. (2014). Adaptive empirical Bayesian smoothing splines. Available at arXiv:1411.6860.

[25] SHEN, W. and GHOSAL, S. (2015). Adaptive Bayesian procedures using random series priors. *Scand. J. Stat.* **42** 1194–1213. MR3426318

[26] SNIEKERS, S. and VAN DER VAART, A. (2015). Adaptive Bayesian credible sets in regression with a Gaussian process prior. *Electron. J. Stat.* **9** 2475–2527. MR3425364

[27] SZABÓ, B., VAN DER VAART, A. and VAN ZANTEN, H. (2015). Honest Bayesian confidence sets for the L^2-norm. *J. Statist. Plann. Inference* **166** 36–51. MR3390132

[28] SZABÓ, B., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** 1391–1428. MR3357861

[29] SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2013). Empirical Bayes scaling of Gaussian priors in the white noise model. *Electron. J. Stat.* **7** 991–1018. MR3044507

[30] TSYBAKOV, A. B. (2004). *Introduction à L'estimation Non-paramétrique*. *Mathématiques & Applications* (*Berlin*) [*Mathematics & Applications*] **41**. Springer, Berlin. MR2013911

[31] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663

[32] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics*: *Contributions in Honor of Jayanta K. Ghosh*. *Inst. Math. Stat. Collect.* **3** 200–222. IMS, Beachwood, OH. MR2459226

[33] VERDINELLI, I. and WASSERMAN, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.* **26** 1215–1241. MR1647645

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

CEREMADE
UNIVERSITY PARIS DAUPHINE
PLACE DU MARÉCHAL DELATTRE DE TASSIGNY
75016 PARIS
FRANCE
E-MAIL: rousseau@ceremade.dauphine.fr

MATHEMATICAL INSTITUTE
LEIDEN UNIVERSITY
NIELS BOHRWEG 1
LEIDEN, 2333 CA
THE NETHERLANDS
E-MAIL: b.t.szabo@math.leidenuniv.nl