The Annals of Statistics
2020, Vol. 48, No. 4, 2155–2179
https://doi.org/10.1214/19-AOS1881

ASYMPTOTIC FREQUENTIST COVERAGE PROPERTIES OF BAYESIAN CREDIBLE SETS FOR SIEVE PRIORS

BY JUDITH ROUSSEAU[1] AND BOTOND SZABO[2]

[1]*Statistics Department, Oxford University, rousseau@ceremade.dauphine.fr*
[2]*Mathematical Institute, Leiden University, b.t.szabo@math.leidenuniv.nl*

We investigate the frequentist coverage properties of (certain) Bayesian credible sets in a general, adaptive, nonparametric framework. It is well known that the construction of adaptive and honest confidence sets is not possible in general. To overcome this problem (in context of sieve type of priors), we introduce an extra assumption on the functional parameters, the so-called "general polished tail" condition. We then show that under standard assumptions, both the hierarchical and empirical Bayes methods, result in honest confidence sets for sieve type of priors in general settings and we characterize their size. We apply the derived abstract results to various examples, including the nonparametric regression model, density estimation using exponential families of priors, density estimation using histogram priors and the nonparametric classification model, for which we show that their size is near minimax adaptive with respect to the considered specific pseudometrics.

1. Introduction. Uncertainty quantification is of key importance in statistical sciences. Estimators without proper uncertainty quantification have only limited practical applicability, since they contain only limited amount of information about their accuracy. In statistics, uncertainty about an estimator is described with the help of confidence sets. Confidence statements are then widely used in statistical practice, for instance, in hypothesis testing. The construction of confidence sets can be, however, very challenging, especially in complex, nonparametric problems.

A very popular aspect of the Bayesian approach is the built-in, straightforward way of quantifying uncertainty, in particular with the help of credible sets, that is, sets with prescribed (typically 95%) posterior probability. By accumulating a large fraction of the posterior mass, these sets describe the remaining uncertainty of the Bayesian procedure. Due to the existing computational machinery of Bayesian techniques (e.g., MCMC, ABC, etc.) these sets are widely used in practice for uncertainty quantification. However, only little is known about their theoretical properties. In parametric models following the celebrated Bernstein–von Mises theorem, credible sets are asymptotically confidence sets as well, laying the base of the practical applicability of the Bayesian approach in simple models; see, for instance, [52].

However, in nonparametric and high-dimensional models the question is still unanswered about how much we can trust Bayesian credible sets as a measure of confidence in the statistical procedure from a frequentist perspective. The first results in the nonparametric paradigm were discouraging, showing that the Bernstein–von Mises theorem does not hold in general, that is, even in the standard Gaussian white noise model using conjugate Gaussian priors the resulting credible sets have frequentist coverage tending to zero; see [16, 17].

Since then the investigation of frequentist coverage properties of Bayesian credible sets has attracted a lot of attention in nonparametric problems. Various approaches were proposed

Received January 2018; revised June 2019.
MSC2020 subject classifications. Primary 62G20, 62G05; secondary 62G08, 62G07.
Key words and phrases. Uncertainty quantification, coverage, posterior contraction rates, adaptation, empirical Bayes, hierarchical Bayes, nonparametric regression, density estimation, classification, sieve prior.

to solve this problem. In [27, 58], the authors verified that by slightly undersmoothing the prior one can still achieve credible sets with good frequentist coverage and minimax size in the same setup as in[16]. Another possibility is to consider weaker, negative Sobolev-norms and derive the Bernstein–von-Mises theorem in the corresponding Sobolev space; see [11, 12, 28].

The preceding results are all based on the knowledge of the regularity of the true underlying function, which is in practice generally not available. A more challenging problem is the construction of Bayesian based confidence sets in the adaptive setting where no information is available on the smoothness of the truth. This, however, turns out to be too much to ask for. In [5, 6, 29, 39], it was shown that it is impossible to construct adaptive confidence sets in general.

More precisely, assume that the true (functional) parameter θ_0 belongs to some regularity or sparsity class Θ^β, indexed by some (unknown) hyper-parameter β belonging to some index set B. When β is unknown, the confidence set $\widehat{C}$ cannot depend on it and it is said to be optimal adaptive if first it has uniform coverage:

$$\liminf_n \inf_{\theta_0 \in \bigcup_{\beta\in B}\Theta^\beta} P_{\theta_0}^{(n)}(\theta_0 \in \widehat{C}) \geq 1-\alpha \tag{1}$$

and second its size is optimal within each parameter class Θ^β, that is, for some universal constant $K > 0$,

$$\liminf_n \inf_{\beta\in B} \inf_{\theta_0\in\Theta^\beta} P_{\theta_0}^{(n)}\Big(\sup_{\theta_1,\theta_2\in\widehat{C}} d(\theta_1,\theta_2) \leq K r_{n,\beta}\Big) \geq 1-\alpha, \tag{2}$$

where α is the prescribed significance level (typically $\alpha = 0.05$), and $r_{n,\beta}$ is the minimax estimation rate within the class Θ^β and with respect to the pseudometric $d(\cdot,\cdot)$.

As mentioned earlier, it is impossible to satisfy both the coverage and the minimax size requirements on the confidence sets in general. To solve this problem, additional assumptions were introduced on the parameter value θ_0 making the construction of adaptive confidence sets possible by discarding certain inconvenient parameters θ_0. A frequently applied assumption is self-similarity where it is assumed that the true parameter has similar "local" and "global" behavior; see, for instance, [4, 15, 21, 32, 35, 47]. Another approach is to discard the parameters which make it impossible to test between the classes Θ^β. This approach was considered in various models in context of regularity classes in [5, 7, 24] and in sparse high-dimensional models [8, 33].

It is known that Bayesian credible balls associated to posterior distributions which concentrate at the minimax rate verify (2); see [25]. The question is then to understand their frequentist coverage and in particular to characterize subsets of $\bigcup_\beta \Theta^\beta$ over which (1) is verified as well.

In [49], the authors have generalized the self-similarity assumption introducing the so-called polished tail assumption, discussed in this article also in more detail. The polished tail (and the stronger self-similarity) assumption was then applied in nonparametric regression with rescaled Brownian motion prior [46] and spline priors [45, 55] and in the context of the Gaussian white noise model with Gaussian priors constructing L_2-, and L_∞-credible sets [23, 48]. Furthermore, an adaptive version of the nonparametric Bernstein–von Mises theorem was given in context of the Gaussian white noise model using conjugate Gaussian priors and spike-and-slab prior [36] under the self-similarity assumption. The polished tail assumption was then slightly extended by the implicit excessive bias assumption introduced in context of the Gaussian white noise model [2] and applied in sparse high-dimensional models with empirical Bayes spike and slab type of priors [3, 14] and with hierarchical and empirical horseshoe prior [51].

All of the above mentioned papers consider specific choices of the model and the prior distribution and use explicit, conjugate computations which obviously have their limitations. Although these papers already shed lights on certain aspects of Bayesian uncertainty quantification, they do not provide a clear understanding of the underlying general phenomena. A general approach for understanding the coverage of credible sets is still missing. Besides for many nonparametric models and priors, no conjugate computation is possible and, therefore, they cannot be handled directly. In this work, we aim to (partially) fill this gap and contribute to the fundamental understanding of this rapidly growing field. We derive abstract results for general choices of models and sieve type of priors, in the spirit of [19, 20, 41].

1.1. *Setup and notation.* We consider observations $\mathbf{Y} \in \mathcal{Y}$ sampled from $P_\theta^{(n)}$, $\theta \in \bar{\Theta}$, which are absolutely continuous with respect to a given measure μ with density $p_\theta^{(n)}$ where n denotes the sample size or signal-to-noise ratio. We denote by $\ell_n(\theta) = \log p_\theta^{(n)}$ the log-likelihood and throughout the paper θ_0 designates the true value of the parameter. We denote by $E_\theta^{(n)}$ and $V_\theta^{(n)}$ the expectation and the variance with respect to $P_\theta^{(n)}$, respectively. For two positive sequences a_n and b_n, we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \le C b_n$ for every $n \in \mathbb{N}$. Furthermore, we denote by $a_n \asymp b_n$ that $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously. For $A, B \in \mathbb{R}^{n \times n}$, the inequality $A \le B$ denotes that $A - B$ is positive semidefinite.

Let us consider a collection of finite dimensional models $\Theta(k)$

$$\Theta = \bigcup_{k \in \mathbb{N}} \Theta(k), \quad \Theta(k) \subset \mathbb{R}^{d_k}, d_k \uparrow \infty, k \in \mathbb{N}, \tag{3}$$

with $d_k \asymp k$. We assume that $\theta_0 \in \bar{\Theta}$ with $\Theta \subseteq \bar{\Theta}$. Note that we do not necessarily assume that θ_0 belongs to any of the models $\Theta(k)$, $k \in \mathbb{N}$, hence we allow for misspecification. These models are very popular and frequently used in practice; see, for instance, [22, 50] for a review.

The parameter k drives the sparsity or the regularity of the model. Finding the model $\Theta(k)$, which is the most appropriate for recovering θ_0, requires additional information about the true parameter (e.g., regularity, sparsity, etc.) which is usually not available. Therefore, a natural approach is to let the data decide about the optimal model $\Theta(k)$. In the Bayesian framework, one can accomplish this by the hierarchical or the empirical Bayes approach. In the hierarchical (or also referred to as full) Bayes approach, one endows the hyperparameter k with a prior distribution π_k and conditionally on k, considers a prior distribution $\pi_{|k}$ on $\theta \in \Theta(k)$, resulting in a two-level prior distribution π on Θ defined by

$$k \sim \pi_k, \qquad \theta | k \sim \pi_{|k}. \tag{4}$$

We denote the posterior distribution on Θ by $\pi(\theta|\mathbf{Y})$ and the conditional distribution of $\theta|(\mathbf{Y}, k)$ by $\pi_{|k}(\theta|\mathbf{Y})$.

In contrast to this in the empirical Bayes approach, one constructs a frequentist estimator $\hat{k}_n$ for the hyperparameter k and plugs it in into the conditional posterior distribution given k, that is,

$$\pi_{|\hat{k}_n}(\theta|\mathbf{Y}) = \pi_{|k}(\theta|\mathbf{Y})|_{k=\hat{k}_n},$$

resulting in the empirical Bayes posterior distribution.

Models in the form of (3) are widely used in the Bayesian literature and under nonrestrictive assumptions the posterior distribution can optimally recover the true parameter θ_0. In more detail, it is common to assume that the true parameter belongs to some regularity class $\theta_0 \in \Theta^\beta$ with some unknown regularity hyperparameter β. Then both the hierarchical

and empirical Bayes approaches achieve (nearly) optimal minimax contraction rate around the truth, in a large collection of cases, without using any additional information about its unknown regularity, leading to an adaptive procedure, in the frequentist sense; see [1, 41] and references therein. In this article, our focus is on the quality of Bayesian uncertainty quantification done via credible balls from a frequentist perspective. There are two main properties of interest in a credible set from a frequentist perspective: the frequentist coverage and the expectation of its size under $P_{\theta_0}^{(n)}$, when θ_0 is assumed to be the true value of the parameter. In the literature, the frequentist coverage properties of Bayesian credible sets constructed from sieve posteriors were only investigated for specific choices of priors and likelihoods; see, for instance, [2, 45, 58]. In this article, we present a general approach under which we can simultaneously investigate the frequentist properties of credible sets resulting from different choices of sieve priors and likelihoods.

We introduce some additional notation. Let $B_k(\theta, u, d)$ denote the d-ball in $\Theta(k)$ with center θ and radius u and $B_k^c(\theta, u, d)$ the complement of such a ball. Furthermore, let $\operatorname{diam}(S, d)$ denote the d-diameter of the set S, that is,

$$\operatorname{diam}(S, d) = \sup_{\theta, \theta' \in S} d(\theta, \theta').$$

We define the square distance of the truth from the set $\Theta(k)$ as

$$b(k) = \inf\{d^2(\theta_0, \theta) : \theta \in \Theta(k)\}.$$

For simplicity, we also extend the definition of the function b on $[0, +\infty)$ by $b(x) = b(k)$ for all $x \in [k, k+1)$ and $b(0) = +\infty$. Note that we allow $d(\cdot, \cdot)$ to depend on n, so that in this case $b(k)$ also depends on n. This will be the case in particular in the regression and in the classification examples; see Sections 3.1 and 3.4, respectively. The normalized Kullback–Leibler divergence and variance of the log-likelihood-ratio are denoted by

$$KL(\theta_0, \theta) = \frac{1}{n} E_{\theta_0}^{(n)}\left(\log\left(\frac{p_{\theta_0}^{(n)}}{p_\theta^{(n)}}\right)\right), \qquad V(\theta_0, \theta) = \frac{1}{n} V_{\theta_0}^{(n)}\left(\log\left(\frac{p_{\theta_0}^{(n)}}{p_\theta^{(n)}}\right)\right),$$

respectively. We denote by $N(\varepsilon, A, d)$ the entropy, that is, the number of ε-radius d-balls needed to cover the set A. Throughout the paper, c and C denote global constants whose value may change one line to another.

2. Main results. In this section, we investigate the frequentist properties of Bayesian credible sets resulting from the hierarchical and the empirical Bayes procedures. We consider the general setting described in Section 1.1 and introduce general, abstract conditions under which credible sets have honest frequentist coverage and rate adaptive size. The derived results will be applied in Section 3 for various choices of sampling and prior models.

Using the posterior distribution $\pi(\theta|\mathbf{Y})$, be it hierarchical or empirical, we construct the Bayesian credible sets as balls centered around some estimator $\hat{\theta}$ (typically the posterior mean) $\hat{C}(\alpha) = \{\theta : d(\theta, \hat{\theta}) \le r_\alpha\}$ where $\alpha \in (0, 1)$ and r_α is the radius of the ball satisfying

$$r_\alpha = \arg\inf_{r>0}\{\pi(\theta : d(\theta, \hat{\theta}) \le r|\mathbf{Y}) \ge 1 - \alpha\}. \tag{5}$$

In our analysis, we also introduce some additional flexibility to the credible sets by allowing them to be blown up by a factor $L > 0$ resulting in

$$\widehat{C}(L, \alpha) = \{\theta : d(\theta, \hat{\theta}) \le L r_\alpha\}.$$

We show that these inflated sets (for sufficiently large blow up factor L) have frequentist coverage tending to one and at the same time their size is (nearly) optimal in a minimax sense under some additional restrictions on the parameter θ_0.

In the Gaussian white noise model with Gaussian prior, [49] shows that a key idea to obtain good coverage is that a trade-off between bias and variance is realized, so that the *correct* value of k (or set of values) is selected either under the posterior $\pi_k(k|\mathbf{Y})$ or the empirical estimator $\hat{k}_n$.

To generalize this idea to non-Gaussian setups, let us consider the index set $\mathcal{K} \subseteq \mathbb{N}$ and define for each $\theta_0 \in \bar{\Theta}$,

$$\varepsilon_n^2(k) = b(k) + \frac{k \log n}{n} \quad \text{and} \quad k_n = \inf\{k \in \mathcal{K} : b(k) \leq k(\log n)/n\}, \tag{6}$$

and $\mathcal{K}_n(M) = \{k \in \mathcal{K} : \varepsilon_n(k) \leq M\varepsilon_n(k_n)\}$. Note that in these notation θ_0 is implicit since $b(k)$ depends on θ_0. We would like to note that the $\varepsilon_n(k)$ quantity defined in this paper is related, but different from the $\varepsilon_n(k)$ quantity defined in the paper [41].

To control the frequentist coverage of $\hat{C}(L, \alpha)$, we need to restrict ourselves to a subset of $\bar{\Theta}$, in a manner similar to [49], generalizing their idea outside the white noise model with empirical Gaussian process prior. We introduce below the general polished tail condition which determines the subclass of functions for which frequentist coverage can be obtained.

DEFINITION 1. Let $\theta \in \bar{\Theta}$, we say that θ (or equivalently its associated bias function $b(\cdot)$) satisfies the general polished tail condition associated to the pseudometric $d(\cdot, \cdot)$ if there exist integers k_0, $R_0 > 1$ and a real $0 < \tau < 1$ such that, for $n > 0$,

$$b(kR_0) \leq \tau b(k) \quad \forall k \in \{k_0, \ldots, k_n\}. \tag{7}$$

For given k_0, R_0 and τ, we denote by $\Theta_{0,n}(R_0, k_0, \tau)$ the class of $\theta \in \bar{\Theta}$ satisfying (7).

We note that in the case where $d(\cdot, \cdot)$ is the ℓ_2-norm, for instance in the Gaussian white noise model, the bias function is $b(k) = \sum_{j=k+1}^{\infty} \theta_{0,j}^2$. The polished tail condition in [49] reads as

$$\sum_{j=N+1}^{\infty} \theta_{0,j}^2 \leq L \sum_{j=N+1}^{\rho N} \theta_{0,j}^2 \quad \forall N \geq N_0, \tag{8}$$

for some $N_0, L, \rho > 0$, which is equivalent to our definition of $\Theta_{0,n}(R_0, k_0, \tau)$ (with $k_0 = N_0$, $\tau = L/(L+1)$ and $R_0 = \rho$). Our new definition, however, extends also to the case where the pseudometric $d(\cdot, \cdot)$ is substantially different from the ℓ_2-norm.

The generalization of the usual bias and variance trade-off is by obtaining a trade-off between the bias (or more precisely the approximation error) $nb(k)$ and a prior penalization term $k \log n$ induced by the prior mass of small neighborhoods: $\pi_{|k}(\theta : d(\theta_{[k]}^o, \theta) \leq u_n)$, where $u_n = o(1)$ and $\theta_{[k]}^o \in \Theta(k)$ can be viewed as a projection of θ_0 onto $\Theta(k)$, typically with respect to the pseudometric d or the KL-divergence. Then typically if $u_n \asymp n^{-H}$ for some $H > 0$, then $\log \pi_{|k}(\theta : d(\theta_{[k]}^o, \theta) \leq u_n) \asymp -k \log n$, so that the set $\mathcal{K}_n(M)$ corresponds to values of k for which this trade-off is achieved.

LEMMA 1. *For any $\theta_0 \in \bar{\Theta}$ and $k \in \mathcal{K}_n(M)$, we have that $k \leq 2M^2 k_n$. Furthermore, for any $\theta_0 \in \Theta_{0,n}(R_0, k_0, \tau)$ let us assume that there exists an $A_0 > 1$ such that*

$$\begin{aligned} &\text{for all } k < k_0 \text{ there exists } k' \in \{k_0, k_0 + 1, \ldots, A_0 k_0\}, \\ &\text{such that } b(k) \geq b(k'). \end{aligned} \tag{9}$$

Then for every $k \in \mathcal{K}_n(M)$ we have $k \geq ck_n$, with $c = R_0^{-m}(2R_0 \vee k_0 A_0)^{-1}$, where $m > 0$ is the smallest integer satisfying $\tau^m \leq (8M^2 R_0)^{-1}$.

The proof of the lemma is deferred to Section B.4 in the Supplementary Material [42].

REMARK 1. Condition (9) is very mild. It is easy to see that it holds automatically for nested sets $\Theta(k)$, where the bias function $k \mapsto b(k)$ is monotone nonincreasing. Furthermore, it can also be verified for models where nestedness occurs only on given geometric subsequences $\Theta(k) \subset \Theta(A_0 k) \subset \Theta(A_0^2 k) \subset ...$, for instance, histograms with regular bins; see Section 3.2.

We will show in Section 2.1 that in the hierarchical Bayes approach the posterior distribution concentrates on $\mathcal{K}_n(M)$ for M large enough if the true parameter satisfies the general polished tail condition (7). A similar phenomenon occurs for the empirical Bayes method, that is, the maximum marginal likelihood estimator $\hat{k}_n$ belongs to the set $\mathcal{K}_n(M)$ with high probability; see Section 2.2.

In the hierarchical prior case, we also consider the following condition on the prior on k:

H The prior on k satisfies

$$e^{-c_2 k \log(k)} \lesssim \pi_k(k) \lesssim e^{-c_1 k}, \quad k \in \mathcal{K}, \tag{10}$$

for some positive constants $c_1, c_2 > 0$.

In order to bound from below the frequentist coverage of $\widehat{C}(L, \alpha)$, we restrict ourselves to a subset of parameters $\Theta_{0,n} \subseteq \Theta_{0,n}(R_0, k_0, \tau)$ for some positive R_0, k_0, τ on which we consider the following assumptions, used both for the empirical Bayes and for the hierarchical Bayes approaches:

A0 The centering point $\hat{\theta}$ satisfies that for all $\varepsilon > 0$ there exists $M_\varepsilon > 0$,

$$\sup_{\theta_0 \in \Theta_{0,n}} P_{\theta_0}^{(n)}\big(d(\theta_0, \hat{\theta}) \le M_\varepsilon \varepsilon_n(k_n)\big) \ge 1 - \varepsilon. \tag{11}$$

A1 There exist $c_3, c_4, C > 0$ such that for all $\theta_0 \in \Theta_{0,n}$,

$$\pi_{|k_n}\big(KL(\theta_0, \theta) \le c_3 \varepsilon_n^2(k_n), V(\theta_0, \theta) \le C\varepsilon_n^2(k_n)\big) \ge C^{-1} e^{-c_4 k_n \log n}.$$

A2 For given sequence $\bar{K}_n$, assume that there exist constants $J_0, J_1, c_5 > 0$, $c_6 \in (0, 1)$ such that the following conditions hold for all $k \le \bar{K}_n$:

(i) There exist sets $\Theta_n(k) \subset \Theta(k)$ satisfying

$$\pi_{|k}\big(\Theta_n(k)^c\big) \le C e^{-(c_2+c_3+c_4+2) n \varepsilon_n^2(k_n)}.$$

(ii) There exist measurable (in $\mathbf{Y}$) functions $\varphi_n(\theta) \in [0, 1]$ such that

$$\sup_{\theta \in \Theta_n(k)} E_{\theta_0}^{(n)}\big(\varphi_n(\theta)\big) \le e^{-c_5 n d^2(\theta_0, \theta)},$$

$$\sup_{\theta \in \Theta_n(k)} \sup_{\substack{\theta' \in \Theta_n(k): \\ d(\theta', \theta) \le c_6 d(\theta_0, \theta)}} E_{\theta'}^{(n)}\big(1 - \varphi_n(\theta)\big) \le e^{-c_5 n d^2(\theta_0, \theta)}.$$

(iii) For all $u \ge J(k) := \max(J_0 \varepsilon_n(k_n), J_1 \sqrt{k(\log n)/n})$,

$$\log N\big(c_6 u, \Theta_n(k) \cap \{u \le d(\theta_0, \theta) \le 2u\}, d\big) \le c_5 n u^2/2.$$

A3 For all $\gamma > 0$, there exists $M_0 > 0$ such that for all $M_0 k_n \le k \le \bar{K}_n$,

$$\pi_{|k}\big(B_k\big(\theta_0, J_1\sqrt{k(\log n)/n}, d\big) \cap \Theta_n(k)\big) \le e^{-(c_2+c_3+c_4+\gamma) n \varepsilon_n^2(k_n)}.$$

A4 Assume that for all $M, \varepsilon > 0$, there exist $c_7, c_8, c_9, c_{10}, \delta_0, B_\varepsilon > 0$ and $r \ge 2$ such that the following conditions hold for all $k \in \mathcal{K}_n(M)$:

(i) There exists a parameter $\theta^o_{[k]} \in \Theta(k)$ satisfying

$$B_k\big(\theta^o_{[k]}, \sqrt{k/n}, d\big) \cap \Theta_n(k) \subset S_n(k, c_7, c_8, r),$$

where

$$S_n(k, c_7, c_8, r) = \Big\{\theta \in \Theta(k) : E^{(n)}_{\theta_0} \log \frac{p^{(n)}_{\theta^o_{[k]}}}{p^{(n)}_\theta} \le c_7 k,\ E^{(n)}_{\theta_0}\Big(\log \frac{p^{(n)}_{\theta^o_{[k]}}}{p^{(n)}_\theta} - E^{(n)}_{\theta_0} \log \frac{p^{(n)}_{\theta^o_{[k]}}}{p^{(n)}_\theta}\Big)^r \le c_8 k^{r/2}\Big\}.$$

(ii) Let $\bar{B} = \Theta_n(k) \cap B_k(\theta_0, (M_\varepsilon + 1)\varepsilon_n(k_n), d)$. Then for every $\theta_0 \in \Theta_{0,n}$,

$$P_{\theta_0}\big(\max_{k \in \mathcal{K}_n(M)} \sup_{\theta \in \bar{B}} \big(\ell_n(\theta) - \ell_n\big(\theta^o_{[k]} - B_\varepsilon k\big) \le 0\big) \ge 1 - \varepsilon.$$

(iii) For every $\delta_{n,k} \le \delta_0$,

$$\sup_{\theta \in \bar{B}} \frac{\pi_{|k}(B_k(\theta, \delta_{n,k}\sqrt{k/n}, d) \cap \Theta_n(k))}{\pi_{|k}(B_k(\theta^o_{[k]}, \sqrt{k/n}, d))} \le c_{10} e^{c_9 k \log(\delta_{n,k})}.$$

REMARK 2. The parameter $\bar{K}_n$ in Assumptions A2 and A3 is chosen to be $Ak_n \log n$ (for some large enough constant $A > 0$) for the hierarchical Bayes method. In case of the empirical Bayes method, it is the upper bound of the interval where the maximum marginal likelihood estimator is taken, that is, $\hat{k}_n \in \{1, 2, \ldots, \bar{K}_n\}$; see (16). In this case, $\bar{K}_n$ is typically taken to be n^H for some $H \in (0, 1/2)$.

REMARK 3. One can also handle dimensions d_k which do not grow linearly with k. Then the definition of $\varepsilon_n(k)$ has to be modified to $\varepsilon_n(k) = b(k) + d_k(\log n)/n$ and the conditions in A1–A4 have to be given with k replaced by d_k.

A brief explanation of the above conditions is in order. Assumptions A1, A2 are the standard prior small ball probability, remaining mass, testing and entropy conditions, routinely used in the literature for determining the contraction rates of the posteriors; see, for instance, [20]. Assumption A3 is commonly considered when upper bounding marginal likelihoods; see, for instance, [9, 30, 40, 41]. These conditions are used to describe the behavior of the posterior distribution on the hyperparameter k and derive upper bounds for the posterior contraction rates. Proving A3 is quite simple in case the pseudometric d is locally equivalent to the ℓ_2-norm; however, it is quite challenging in the context of mixture models, where the geometry of the L_1 metric is complex.

Assumption A4 gathers three conditions, which are required to hold only over the models $k \in \mathcal{K}_n(M)$. This assumption is used to derive lower bounds for the radius of the credible sets.

Assumption A4(i) requires that locally the (slightly modified) Kullback– Leibler divergence can be bounded by the distance $d(\cdot, \cdot)$ (up to a multiplicative constant). Note that due to the model misspecification, that is, typically $\theta_0 \notin \Theta(k)$, we consider a projection $\theta^o_{[k]}$ of θ_0 onto $\Theta(k)$ for controlling the prior penalization term; see the discussion below (6). This is a rather mild assumption, the main difficulty here lies in obtaining a sharp upper bound on $KL(\theta_0, \theta) - KL(\theta_0, \theta^o_{[k]})$ and not only on $KL(\theta_0, \theta)$. It can be weakened by considering c_7

going to infinity, this would, however, induce a bigger inflation of the radius of the credible ball $\hat{C}_n(L,\alpha)$.

In Assumption A4(ii), the log-likelihood ratio is uniformly controlled in a neighborhood of θ_0 with high probability. This is not such a stringent condition since the required control is not sharp at all. Indeed it is required that the log-likelihood ratio $\ell_n(\theta) - \ell_n(\theta^o_{[k]})$ is bounded from above by $O(k)$, but note that under P_{θ_0} its expectation is equal to $-n(KL(\theta_0,\theta) - KL(\theta_0,\theta^o_{[k]})) \le 0$, which acts as a pull back force; see the proof of Propositions 4, 5 and 6 in the Supplementary Material [42].

In condition A4(iii), note that since $\theta \in B_k(\theta_0,(M_\varepsilon+1)\varepsilon_n(k_n),d)$ and since (typically) $\theta^o_{[k]} \in B_k(\theta_0,\varepsilon_n(k),d)$, $d(\theta,\theta^o_{[k]}) \le (M_\varepsilon+1)\varepsilon_n(k_n) + \varepsilon_n(k) \le (M+M_\varepsilon+1)\varepsilon_n(k_n)$ for $k \in \mathcal{K}_n(M)$, so that condition A4(iii) requires that in case the ball around any point in the vicinity of $\theta^o_{[k]}$ has a substantially smaller radius than a $\sqrt{k/n}$ ball centered around $\theta^o_{[k]}$ then the prior mass of the ball is also substantially smaller. This is verified in particular when the distance $d(\cdot,\cdot)$ behaves locally like the Euclidean distance and the prior densities are bounded from below and above, locally. The intuition behind this condition is the following. To achieve high frequentist coverage for the credible set, the prior cannot put substantially more mass around the centering point than on a small neighborhood of the truth, else the posterior would be even more concentrated around the centering point resulting in overly confident uncertainty statements. Since the centering point is random, but living in a close neighborhood of the truth we require this condition to hold uniformly over the ball $B_k(\theta_0,(M_\varepsilon+1)\varepsilon_n(k_n),d)$. Conditions A3 and A4(iii) are the most demanding assumptions, because they require nontrivial upper bounds on prior masses of d-balls.

Assumption A0 is on the centering point and is satisfied typically for usual estimates such as the posterior mean; see the examples in Section 3.

In the literature, different variations were considered of the standard conditions A1 and A2; see, for instance, [19, 20]. Here, we consider another version of A2(iii) (based on slicing of the sets $\Theta_n(k)$), which will be applied in the density estimation example with exponential families of priors. Define $J_0(k) = J_0 \vee J_1\sqrt{k\log n/n}\varepsilon_n(k_n)^{-1}$.

A2 (iiib) There exist a (possibly infinite) cover $B_{n,j}(k)$, of the set $\Theta_n(k) \cap \{\theta : d(\theta,\theta_0) \ge J(k)\varepsilon_n(k_n)\}$ for $k \le \bar{K}_n$, such that for all j there exists $c(k,j) \ge J(k)$ satisfying

$$B_{n,j}(k) \subset \Theta_n(k) \cap \{d(\theta,\theta_0) > c(k,j)\varepsilon_n(k_n)\} \quad \text{with} \tag{12}$$

$$\sum_j \exp\left(-\frac{c_5}{2}nc(k,j)^2\varepsilon_n(k_n)^2\right) \lesssim e^{-(c_2+c_3+c_4+2)n\varepsilon_n^2(k_n)}, \tag{13}$$

where c_2, c_3, c_4 are defined in Assumptions H and A1 and

$$\log N\big(c_6c(k,j)\varepsilon_n(k_n), B_{n,j}(k), d\big) \le \frac{c_5c(k,j)^2n\varepsilon_n(k_n)^2}{2}. \tag{14}$$

In the next subsections, we show that under the above assumptions together with the general polished tail restriction the credible sets resulting both from the hierarchical and the empirical Bayes procedures have optimal size and high frequentist coverage.

2.1. *Hierarchical Bayes approach.* In this section, we present the results for the hierarchical prior defined by (4) satisfying Assumption H. We show that under the general polished tail condition and the assumptions introduced in the preceding section the inflated credible set $\hat{C}(L_n,\alpha)$ with $L_n \gtrsim \sqrt{\log n}$ has good frequentist properties, that is, it has good frequentist coverage and we can characterize their size on $\Theta_{0,n} = \Theta_{0,n}(R_0,k_0,\tau)$, $R_0 > 1$, $k_0 \ge 1$ and $\tau < 1$.

THEOREM 1. *Assume that conditions* H, A0–A4 *and* (9) *hold, with* $\bar{K}_n = Ak_n \log n$ *and* $A = c_2 + c_3 + c_4 + 1$ *in Assumptions* A2 *and* A3, *then for every* $\varepsilon > 0$ *there exists a constant* $L_{\varepsilon,\alpha} > 0$ *such that*

$$\liminf_n \inf_{\theta_0 \in \Theta_{0,n}} P_{\theta_0}^{(n)}\big(\theta_0 \in \widehat{C}(L_{\varepsilon,\alpha}\sqrt{\log n}, \alpha)\big) > 1 - \varepsilon. \tag{15}$$

REMARK 4. In Theorem 1 (and also in Theorem 2 below), the inflation of the radius is of order $\sqrt{\log n}$, which is an unpleasant feature of the result. We believe that this is a necessary inflation, at least for centering points $\hat{\theta}$ leaving in the bulk of the posterior distribution, like the posterior mean. Indeed as appears in the proof (see also Lemmas 2 and 3), the posterior mass essentially lives on the sets of k that achieves the balance $k \log n \asymp n\varepsilon_n^2(k)$, while an optimal behavior would be to achieve the balance $k \asymp n\varepsilon_n^2(k)$. This is a typical feature of hierarchical (or empirical) Bayesian approaches with a hyperprior on the model k and is strongly related to the $\log n$ penalty induced by the marginal likelihood, as expressed in the reknown BIC approximation. This results in having the posterior distribution concentrate on values of k that are too small, so that the bias $b(k)$ dominates the statistical error within each model $\Theta(k)$ which is $O(k/n)$. The necessity of the $\sqrt{\log n}$ factor is demonstrated in the context of the nonparametric regression model. In Propositions 2 and 3, it is shown that without a $\sqrt{\log n}$ blow up the credible sets have coverage tending to zero for certain representative (typical) elements of the polished tail class. There are two ways to temper this. One can either follow [18] using a block prior on the components of θ which groups together models in blocks and within each block shrinks very strongly the coefficients to 0 to ensure that the selected models under the posterior have a large enough number of components. An alternative method could be to find a centering point $\hat{\theta}$ which is rougher than the posterior; see Lemma B.1 in the Supplementary Material.

The proof of Theorem 1 is deferred to Section 5. A key step in the proof is understanding the asymptotic behavior of $\pi_k(k|\mathbf{Y})$. In particular, we show that the posterior distribution accumulates most of its mass on $\mathcal{K}_n(M)$, where a trade-off between bias and prior-penalization or complexity (equivalent to the variance term in the Gaussian setup) is achieved. This is presented in the following lemma.

LEMMA 2. *Take any* $\varepsilon > 0$ *and assume that conditions* H *and* A1–A3 *hold. Then there exists a large enough constant* $\bar{M}_\varepsilon > 0$ *such that*

$$\sup_{\theta_0 \in \bar{\Theta}} E_{\theta_0}^{(n)}\big(\pi_k\big(k \notin \mathcal{K}_n(\bar{M}_\varepsilon)|\mathbf{Y}\big)\big) \le \varepsilon.$$

The proof is presented in Section B.1 of the Supplementary Material. Lemma 2 is similar in spirit to Theorem 2.1 of [41], however, the definition of $\varepsilon_n(k)$ being different in both papers; the proofs of both results are significantly different.

The following lemma states that $\varepsilon_n(k_n)$ corresponds to the posterior concentration rates, hence $\hat{\theta}$ can be any random point of the posterior distribution or depending on $d(\cdot,\cdot)$ the posterior mean or other posterior summary.

LEMMA 3. *Assume that conditions* H *and* A1–A3 *hold. Then for every* $\varepsilon > 0$ *there exists* $C_\varepsilon > 0$ *such that*

$$\sup_{\theta_0 \in \bar{\Theta}} E_{\theta_0}^{(n)}\big(\pi\big(d(\theta,\theta_0) \ge C_\varepsilon \varepsilon_n(k_n)|\mathbf{Y}\big)\big) \le \varepsilon.$$

The proof of Lemma 3 is presented in Section B.2 in the Supplementary Material.

Finally, we show that the radius of the credible set is bounded from above by a multiple of $\varepsilon_n(k_n)$.

COROLLARY 1. *Under the assumptions of Lemma 3 and* A0, *there exists $C_\varepsilon > 0$ large enough such that*

$$\inf_{\theta_0 \in \bar{\Theta}} P_{\theta_0}^{(n)}\big(\operatorname{diam}\big(\widehat{C}(1,\alpha), d\big) \le C_\varepsilon \varepsilon_n(k_n)\big) \ge 1 - \varepsilon.$$

The lemma is a straightforward consequence of Assumption A0 and Lemma 3.

2.2. *Empirical Bayes approach.* An alternative approach to endow the hyperparameter k by a prior is to estimate it from the data directly and plug in this estimator into the posterior distribution. One of the most commonly used approaches is the maximum marginal likelihood empirical Bayes method, where one estimates the hyperparameter with the maximizer of the marginal likelihood function

$$\hat{k}_n = \arg\max_{k \le \bar{K}_n} \int_{\Theta(k)} e^{\ell_n(\theta)} \pi_{|k}(\theta)\, d\theta, \tag{16}$$

where $\ell_n(\theta)$ denotes the log-likelihood function. This empirical Bayes technique is closely related to the hierarchical Bayes approach [41], however, in certain situations they can have substantially different behavior [10, 34].

In the empirical Bayes approach, we construct the (inflated) credible set similar to the hierarchical Bayes case, that is, we consider a d-ball around the centering point $\hat{\theta}$ (typically the empirical Bayes posterior mean)

$$\widehat{C}_{\hat{k}_n}(L,\alpha) = \big\{\theta \in \Theta(\hat{k}_n) : d(\theta, \hat{\theta}) \le L_n r_\alpha(\hat{k}_n)\big\}, \tag{17}$$

where $L_n > 0$ is a blow up factor and the radius $r_\alpha(\hat{k}_n)$ is defined as

$$r_\alpha(\hat{k}_n) = \arg\inf_{r>0} \big\{\pi_{|\hat{k}_n}\big(d(\theta,\hat{\theta}) \le r_\alpha(\hat{k}_n) | \mathbf{Y}\big) \ge 1 - \alpha\big\}, \tag{18}$$

for a typically small $\alpha \in (0,1)$. We show that these sets have similar size as the hierarchical Bayes credible sets and good frequentist coverage under the general polished tail condition (7) for appropriately large blow up factor L_n of order $\sqrt{\log n}$.

THEOREM 2. *Assume that conditions* A0--A4 *and* (9) *hold with $\bar{K}_n \le n^H$ for some $H \ge 0$. Then for every $\varepsilon, \alpha \in (0,1)$ there exists a large enough constant $L_{\varepsilon,\alpha}$ such that*

$$\liminf_n \inf_{\theta_0 \in \Theta_{0,n}} P_{\theta_0}^{(n)}\big(\theta_0 \in \widehat{C}_{\hat{k}_n}(L_{\varepsilon,\alpha}\sqrt{\log n}, \alpha)\big) \ge 1 - \varepsilon.$$

Furthermore, there exists $C_\varepsilon > 0$ such that

$$\inf_{\theta_0 \in \bar{\Theta}} P_{\theta_0}^{(n)}\big(\operatorname{diam}\big(\widehat{C}_{\hat{k}_n}(1,\alpha), d\big) \le C_\varepsilon \varepsilon_n(k_n)\big) \ge 1 - \varepsilon.$$

The proof is deferred to Section B.3 in the Supplementary Material.

3. Application to various models. In this section, we apply our abstract results (Theorems 1 and 2 and Corollary 1) in four examples: nonparametric regression, density estimation with histogram priors and with exponential family priors and nonparametric classification. To prove the contraction rate and coverage results in all of the examples, we have shown that the considered pseudometrics are locally equivalent to the ℓ_2 norm on the parameter space $\bar{\Theta}$. These results are of interest on their own right. Besides it also results in (nearly) optimal posterior contraction rates and coverage of the credible sets in terms of the ℓ_2 and other related norms. Condition A4(ii) requires uniform and sharp control on the likelihood ratio. In the following examples, we give a general strategy to prove such kind of statements, which can come handy in other nonparametric problems as well.

3.1. *Application to nonparametric regression.* In this section, we consider the fixed design regression model and investigate the behavior of Bayesian credible sets based on sieve priors. Assume that we observe the sequence $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ satisfying

$$Y_i = f_0(x_i) + \sigma Z_i, \quad x_i \in [0,1], i = 1, 2, \ldots, n, \tag{19}$$

where Z_i are i.i.d. standard normal random variables, $\sigma = 1$ for simplicity and $x_1, x_2, \ldots, x_n$ are fixed (or random) design points.

Next, we consider the basis $\phi_1(x), \phi_2(x), \ldots$ in $L_2[0,1]$. Note that every $f \in L_2[0,1]$ can be written in the form $f(x) = f_\theta(x) = \sum_{i=1}^{\infty} \theta_i \phi_i(x)$ (with the convention $\theta = (\theta_1, \ldots, \theta_k, 0, 0, \ldots)$ for $\theta \in \Theta(k) = \mathbb{R}^k$) and we assume that the true function f_{θ_0} belongs to a Sobolev-type smoothness class $S^\beta(L_0)$, defined as

$$S^\beta(L_0) = \left\{ f_\theta : \sum_{i=1}^{\infty} \theta_i^2 i^{2\beta} \le L_0 \right\}, \quad \text{for some } \beta, L_0 > 0, \tag{20}$$

with typically unknown regularity parameter $\beta > 0$. Slightly abusing our notation, we also write $\theta_0 \in S^\beta(L_0)$ if $f_{\theta_0} \in S^\beta(L_0)$. Note that depending on the basis functions ϕ_j this may or may not refer to the classical Sobolev balls; note also that the Fourier basis satisfies the assumptions below. The minimax estimation rate, which typically coincides with the minimax size of confidence sets (see, for instance, [39]) over $S^\beta(L_0)$ with respect to the ℓ_2-norm is $n^{-\beta/(1+2\beta)}$.

Let $\bar{\phi}_i = (\phi_i(x_1), \ldots, \phi_i(x_n))^T$ and for any $k \le n$ we introduce the notation $\Phi_k = (\bar{\phi}_1, \bar{\phi}_2, \ldots, \bar{\phi}_k) \in \mathbb{R}^{n \times k}$. Next, let $d_n(\theta, \theta')^2 = \frac{1}{n} \sum_{i=1}^n (f_\theta(x_i) - f_{\theta'}(x_i))^2$ be the empirical L_2-norm between the functions $f_\theta, f_{\theta'} \in L_2$. Furthermore, let us introduce the notation $f_{\theta,n} = (f_\theta(x_1), \ldots, f_\theta(x_n))$ and denote by $\theta^o_{[k]}$ the empirical L_2-norm projection of $f_{0,n} = (f_{\theta_0}(x_1), \ldots, f_{\theta_0}(x_n))^T$ to the space $\{\Phi_k \theta : \theta \in \mathbb{R}^k\}$ or in other words the d_n-projection of θ_0 onto $\mathbb{R}^k$. Then defining $b(k)$ in terms of the pseudometric $d_n(\cdot, \cdot)$ leads to $b(k) = d_n(\theta_0, \theta^o_{[k]})^2$ the approximation error of the true function with the k-dimensional projection. Assume furthermore that there exists a constant $C_0 \ge 1$ and a sequence K_n going to infinity such that

$$C_0^{-1} I_{K_n} \le \frac{\Phi_{K_n}^T \Phi_{K_n}}{n} \le C_0 I_{K_n}, \tag{21}$$

where I_k denotes the identity matrix in $\mathbb{R}_k$.

REMARK 5. The above assumptions on the choice of the basis functions $\phi_j(x) \in L_2[0,1]$ and the design points $x_1, x_2, \ldots, x_n$ are very mild and standard. There are many suitable choice of bases satisfying these properties. Orthonormal bases in $\mathbb{R}^n$, such as the discrete wavelet basis relative to the design points satisfy (21) with $K_n = n$; some orthonormal bases in L_2 will satisfy (21) for some finite value K_n. In the case of the Fourier basis for instance, (21) is valid as soon as $K_n = o(n)$.

REMARK 6. Let us consider a probability measure ν on $[0,1]$ and take an orthonormal $L_2(\nu)$ bases ϕ_j. Then in view of Rudelson's inequality, [43],

$$E_\nu \left\| \frac{\Phi_k^T \Phi_k}{n} - I_k \right\|_2 \leq M \sqrt{\frac{k \log n}{n}} \tag{22}$$

for all $k \leq k_0 n/\log n$ and some k_0 small enough. Hence following from Lemma A.6 in the Supplementary Material [42], if $K_n \log K_n = o(n)$ assertion (21) is verified with ν-probability going to 1.

Due to the condition (21), we have to slightly modify the polished tail condition by assuming that the approximation error using the largest model $\Theta(K_n)$ is not too large, that is, we take

$$\Theta_{0,n} = \Theta_{0,n}(R_0, k_0, \tau) \cap \{\theta_0 : b(K_n) \leq \delta K_n (\log n)/n\},$$

for some $\delta < 1 \wedge C_0$ and consider $\theta_0 \in \Theta_{0,n}$.

REMARK 7. To understand better the meaning of the restriction $\theta_0 \in \Theta_{0,n}$, assume that $\sum_{j=1}^\infty |\theta_{0,j}| < +\infty$ and $\max_j \|\phi_j\|_\infty < \infty$ so that $\|f_{\theta_0}\|_\infty < \infty$. If (21) is true for all $1 \leq k \leq Cn$, $C > 0$, then $\|\Delta_k\|_\infty = o(1)$ as k goes to infinity, with $\Delta_k = f_{\theta_0} - \sum_{j=1}^k \theta_{0,j}\phi_j$. Since $b(k) \leq \|\Delta_k\|_\infty^2$, then there exists K_n, with $Cn \geq K_n \geq 1$, such that $b(K_n) \leq \delta K_n (\log n)/n$ for all $n \geq 2$ and $\delta > 0$. Hence for all $L > 0$, $\{\theta_0 : \|\theta_0\|_1 \leq L\} \cap \Theta_{0,n} \subset \Theta_{0,n}$, when n is large enough, following from the inequality $\|f_\theta\|_\infty \leq \|\theta\|_1 \max_j \|\phi_j\|_\infty$. However, if (21) is only true for $K_n = o(n)$, then $\Theta_{0,n}$ will typically be more constraint. For instance, for $\theta_0 \in S^\beta(L_0)$, $\beta > 1/2$, we can bound $b(K_n) \leq \|\Delta_{K_n}\|_\infty^2 \lesssim K_n^{-2(\beta-1/2)}$ (using the Cauchy–Schwarz inequality) so that $b(K_n) \leq \delta K_n (\log n)/n$ if $K_n \gtrsim (n/\log n)^{1/(2\beta)}$. In case $K_n \asymp n/\log n$, $\beta > 1/2$ is enough. The upper bound $K_n^{-2(\beta-1/2)}$ is independent of the design and the chosen basis and can be improved in particular cases.

For instance in the random design case with distribution ν, bounded orthonormal basis $\max_j \|\phi_j\|_\infty < +\infty$ and writing $\theta_{0,[k]} = (\theta_{0,1}, \ldots, \theta_{0,k}) \in \mathbb{R}^k$ one has

$$\begin{aligned} \nu\big(d_n^2(\theta_0, \theta_{0,[K_n]}) > C\|\Delta_{K_n}\|_2^2\big) &= \nu\left(\sum_{i=1}^n \left(\sum_{j=K_n+1}^\infty \theta_{0,j}\phi_j(x_i)\right)^2 > nC\|\Delta_{K_n}\|_2^2\right) \\ &\leq \frac{E_\nu((\sum_{j=K_n+1}^\infty \theta_{0,j}\phi_j(X))^2)}{C\|\Delta_{K_n}\|_2^2} \leq \frac{1}{C}. \end{aligned}$$

Therefore, $b(K_n) \leq d_n^2(\theta_0, \theta_{0,[K_n]}) \leq C\|\Delta_{K_n}\|_2^2 \lesssim K_n^{-2\beta}$ with large probability.

REMARK 8. In the fixed design regression model with $K_n \geq n^{\frac{1}{2(\beta_0 - 1/2)}}$ (where $\beta_0 > 0$ is the smallest regularity level we are adapting to), the set $\Theta_{0,n}$ contains the set in $\{\theta_0 : b(K_n) \leq \delta K_n (\log n)/n\}$ satisfying the L_2 polished tail condition of [49], that is, if

$$\|\theta_{0,[R_0 k]} - \theta_0\|_2^2 \leq \tau_1 \|\theta_{0,[k]} - \theta_0\|_2^2, \quad \tau_1 < 1/(5C_0^2)$$

for all $k \geq k_0$, then $\theta_0 \in \Theta_{0,n}$. In the random design regression model (with arbitrary sequence K_n tending to infinity), the above inclusion holds with ν-probability arbitrary close to one. Therefore, the discussion in [49] on the L_2 polished tail condition, in terms of the force of the restriction induced by this condition applies here, namely that the condition is nonrestrictive from a statistical complexity, topological and Bayesian perspective.

The proof of the above remark is given in Section B.5 in the Supplementary Material [42].

Then we define the prior distribution on the regression function f by endowing the sequence of coefficients θ with the standard sieve prior, that is,

$$\theta = (\theta_1, \dots, \theta_k)|k \sim \prod_{i=1}^{k} g(\theta_i),$$

$$k \sim \text{Geom}(p) \text{ or } \text{Pois}(\lambda),$$

where Geom(p) and Pois(λ) denote the geometric and Poisson distributions, respectively, with parameters $p \in (0, 1)$ and $\lambda > 0$, and $g(\cdot)$ satisfies the standard assumption

$$G_1 e^{-G_2|x|^q} \le g(x) \le G_3 e^{-G_4|x|^q}, \tag{23}$$

for some positive constants G_1, G_2, G_3, G_4 and q. Alternatively, we can also estimate k by the MMLE (16) and plug in the estimator $\hat{k}_n$ into the posterior. These type of priors were considered for instance in [1] and [41], where it was shown that the corresponding hierarchical and empirical Bayes posterior distributions achieve (up to a $\log n$ factor) adaptive contraction rate around the true function f_0. The frequentist behavior of the Bayesian credible sets in the context of the regression model was investigated only in a few papers [45, 46, 55] for specific conjugate priors allowing direct computations, which cannot be applied in the present setting due to the lack of explicit expression for the posterior. Here, we consider both the inflated hierarchical Bayes credible set

$$\widehat{C}(L\sqrt{\log n}, \alpha) = \{\theta : d_n(\theta, \hat{\theta}) \le L\sqrt{\log n} r_\alpha\},$$

with $\pi(\theta : d_n(\theta, \hat{\theta}) \le r_\alpha | \mathbf{Y}) \ge 1 - \alpha$ and $\hat{\theta}$ satisfying Assumption A0 and the inflated MMLE empirical Bayes credible set defined along the same lines. By applying Theorems 1 and 2 together with Corollary 1, we can verify that both credible sets have good frequentist coverage and (almost) rate adaptive size under the general polished tail assumption.

PROPOSITION 1. *Consider the fixed design regression model* (19) *with* $f_{\theta_0} \in S^\beta(L_0)$ *for some* $\beta \ge \beta_0 > 1/2$ *and assume that condition* (21) *is satisfied with* $K_n > n^{\frac{\beta_0}{(1+2\beta_0)(\beta_0 - 1/2)}}$. *Denote both the inflated hierarchical Bayes and empirical Bayes credible sets, centered around any estimator* $\hat{\theta}$ *satisfying Assumption* A0 *by* $\widehat{C}_n(L\sqrt{\log n}, \alpha)$. *Then* $\widehat{C}_n(L\sqrt{\log n}, \alpha)$ *has* (*up to a* $\log n$ *factor*) *rate adaptive size and frequentist coverage tending to one under the general polished tail assumption* (7), *that is, for every* $\varepsilon, R_0, k_0, \tau > 0$ *there exist a large enough* $L, C > 0$ *such that*

$$\liminf_n \inf_{\theta_0 \in \Theta_{0,n}(R_0, k_0, \tau) \cap S^{\beta_0}(L_0)} P_{\theta_0}^{(n)}\big(\theta_0 \in \widehat{C}_n(L\sqrt{\log n}, \alpha)\big) \ge 1 - \varepsilon,$$

$$\liminf_n \inf_{\beta \ge \beta_0} \inf_{\theta_0 \in \mathcal{S}^\beta(L_0)} P_{\theta_0}^{(n)}\left(\text{diam}\big(\widehat{C}_n(1, \alpha), d_n\big) \le C\left(\frac{n}{\log n}\right)^{-\frac{\beta}{1+2\beta}}\right) \ge 1 - \varepsilon.$$

The proof of the proposition is given in Section A.1 of the Supplementary Material [42].

REMARK 9. Assumption A0 on the estimator is very mild, for instance, a typical draw from the posterior distribution satisfies it; see the comment above Lemma 3. Furthermore, standard estimators like the posterior mean also satisfies this assumption; see, for instance, [1]. We also note that similar results hold for the random design regression as well.

REMARK 10. Recall that by assumption (21) the empirical L_2 pseudometric d_n and the ℓ_2-norm are equivalent over $\Theta(k), k \leq K_n$. Furthermore, note that the inequality $\|\theta_0 - \theta_{0,[K_n]}\|_2^2 \lesssim K_n^{-2\beta} \leq 1/n$ holds for $K_n > n^{\frac{\beta_0}{(1+2\beta_0)(\beta_0-1/2)}}$, $\beta \geq \beta_0$. Hence we get that the same contraction rate and coverage statements as in Proposition 1 hold for the metric ℓ_2 as well.

The $\sqrt{\log n}$ blow up factor in the credible set is rather inconvenient and makes the procedure less appealing. The question naturally arises whether this blow up factor is just an artefact of the proof and can be removed or whether it is necessary to reach the desired frequentist coverage. We show below that without inflating the credible sets (centered at the posterior mean) with a multiple of $\sqrt{\log n}$ one would get coverage tending to zero for a large class of parameters satisfying the polished tail condition, justifying the presence of the inflating factor.

In view of [49], let us consider the class of self-similar functions

$$\mathcal{H}_s^\beta(L) = \{ f_\theta : L^{-1} i^{-\beta-1/2} \leq |\theta_i| \leq L i^{-\beta-1/2}, i = 1, 2, \ldots \},$$

where it was also shown that the present set is not substantially smaller than the entire hyperrectangle (the set without the lower bound assumption on $|\theta_i|$) from a topological and statistical complexity point of view. Note also that $\mathcal{H}_s^\beta(L) \subset S^{\beta+\varepsilon}(C)$, for arbitrary $\varepsilon > 0$ and some sufficiently large constant $C > 0$.

PROPOSITION 2. *Consider the fixed design regression model* (19) *with* $f_{\theta_0} \in \mathcal{H}_s^\beta(L)$ *for some* $\beta \geq \beta_0 > 1/2$ *and orthogonal basis* $\Phi_{K_n}^T \Phi_{K_n} = n I_{K_n}$ (*where* $K_n > n^{\frac{\beta_0}{(1+2\beta_0)(\beta_0-1/2)}}$). *Furthermore, take the prior* $g(\theta)$ *to be either the normal* $N(\mu, \sigma)$ *or Laplace* $\mathrm{Lap}(\mu, b)$ *distribution. Then the empirical Bayes credible set centered around the posterior mean* $\hat{\theta}$ *and inflated with a factor* $m_n \log^{1/2} n$, *for arbitrary* $m_n = o(1)$, *has frequentist coverage tending to zero, that is, for every* $\alpha > 0$,

$$\limsup_n \sup_{f_{\theta_0} \in \mathcal{H}_s^\beta(L)} P_{\theta_0}^{(n)}\big(\theta_0 \in \{\theta : d_n(\theta, \hat{\theta}) \leq m_n \sqrt{\log n} r_\alpha(\hat{k}_n)\}\big) = 0$$

The proof of the proposition is given in Section A.2 of the Supplementary Material [42].

It is common or folklore knowledge that empirical Bayes procedures underestimate uncertainty compared to hierarchical Bayes procedures. However, we prove below that under (somewhat) more restrictive conditions on θ_0 the same blow up factor is required for the hierarchical Bayes credible ball centered at the posterior mean. More precisely, let $\ell : \mathbb{N} \to \mathbb{R}_+$ be a slowly varying function going to 0 at infinity and set

$$\mathcal{H}_{ss}^\beta(L, \ell) = \{ f_\theta \in \mathcal{H}_s^\beta(L); \exists r_\infty \in [1/L, L]; |\theta_i^2 i^{2\beta+1} - r_\infty^2| \leq \ell(i), i \geq 1 \}.$$

PROPOSITION 3. *Consider the fixed design regression model* (19) *with* $f_{\theta_0} \in \mathcal{H}_s^\beta(L)$ *for some* $\beta > 1$ *and orthogonal basis* $\Phi_{K_n}^T \Phi_{K_n} = n I_{K_n}$ (*where* $K_n = n/\log n$). *Furthermore, assume that the log-prior* $\log g(\theta)$ *is continuously differentiable on* $\mathbb{R}$. *Then the hierarchical Bayes credible set centered around the posterior mean* $\hat{\theta}$ *and inflated with a factor* $m_n \delta_n^{-1/2}$, $\delta_n = 1/\log n + \ell(k_n)$ *for arbitrary* $m_n = o(1)$, *has frequentist coverage tending to zero, that is, for every* $\alpha > 0$,

$$\limsup_n \sup_{f_{\theta_0} \in \mathcal{H}_{ss}^\beta(L,\ell)} P_{\theta_0}^{(n)}\big(\theta_0 \in \{\theta : d_n(\theta, \hat{\theta}) \leq m_n \delta_n^{-1/2} r_\alpha\}\big) = 0.$$

In particular, if $\ell(i) \lesssim 1/\log(i)$, *then* $\delta_n^{-1/2} \asymp \sqrt{\log n}$.

The proof of the proposition is given in Section A.3 of the Supplementary Material [42].

REMARK 11. From the proofs of Propositions 2 and 3, it appears that the necessity of the logarithmic blow up follows from (1) the suboptimal behavior of the posterior mean and (2) the concentration of the posterior (or the empirical Bayes) distribution on values of k that are too small. We note, however, that other (typical) summary statistics of the posterior distribution have similar behavior as the bulk of the posterior is located at a suboptimal place. One can of course choose other centering points, not related to the posterior, to avoid the $\sqrt{\log n}$ blow up factor. For instance, one can consider Lepski's estimator in the sequence model (see Lemma B.1 in the Supplementary Material) but this brings us outside of the Bayesian framework and we are hesitant recommending such a solution. In this case, one does not need a logarithmic inflation factor, because it is already present in the radius of the credible ball.

3.2. *Application to density estimation using histogram priors.* In this section, we consider the density estimation model, that is, we assume to observe $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ i.i.d. samples from a true density function p_0 and our goal is to recover this density. We assume that p_0 is continuous, bounded from below by c_0 and from above by C_0. Furthermore, we assume that it belongs to a Hölder smoothness class $\mathcal{H}^\beta(L_0)$ for some $\beta \in (0, 1]$.

We investigate the Bayesian approach using histogram prior distributions; see, for instance, [13, 41, 44]. In other words, let $\Theta(k)$ denote the collection of k-bins random histogram where the bins are regular: $[(j-1)/k, j/k)$, $j = 1, \ldots, k$,

$$p_\theta(x) = k \sum_{j=1}^{k} \theta_j \mathbf{1}_{I_j}(x), \quad \theta_j \geq 0, \sum_{j=1}^{k} \theta_j = 1. \tag{24}$$

We therefore identify $\Theta(k)$ with the k-dimensional simplex $\mathcal{S}_k = \{x \in [0,1]^k; \sum_{i=1}^k x_i = 1\}$. First, we endow the hyperparameter k with either a Poisson Pois(λ) or a geometric Geom(p) hyperprior with $\lambda > 0$ and $0 < p < 1$. Given k consider a Dirichlet prior $\mathcal{D}(\alpha_{1,k}, \ldots, \alpha_{k,k})$ on $(\theta_1, \ldots, \theta_k)$, that is, the hierarchical prior π on the densities takes the form

$$\theta = (\theta_1, \ldots, \theta_k)|k \sim \mathcal{D}(\alpha_{1,k}, \ldots, \alpha_{k,k}),$$
$$c_1 k^{-a} \leq \alpha_{j,k} \leq C_1, k \sim \text{Geom}(p) \text{ or } \text{Pois}(\lambda),$$

for some $a \geq 0$ and $c_1, C_1 > 0$. Alternatively, we apply the MMLE $\hat{k}_n$ for the hyperparameter k and then consider the Dirichlet prior $\mathcal{D}(\alpha_{1,\hat{k}_n}, \ldots, \alpha_{\hat{k}_n,\hat{k}_n})$ on $(\theta_1, \ldots, \theta_{\hat{k}_n})$.

Then we consider the inflated hierarchical Bayes credible set

$$\widehat{C}(L\sqrt{\log n}, \alpha) = \{p_\theta : h(p_\theta, p_{\hat{\theta}}) \leq L\sqrt{\log n} r_\alpha\},$$

with $h(\cdot,\cdot)$ the Hellinger distance, $\hat{\theta}$ satisfying assumption (11) with $d(\theta, \theta') = h(p_\theta, p_{\theta'})$ and the radius r_α satisfies $\pi(\theta : h(p_\theta, p_{\hat{\theta}}) \leq r_\alpha|\mathbf{Y}) \geq 1 - \alpha$. Note that since the Hellinger metric is bounded and convex, and the posterior distribution contracts around the truth with the optimal rate $\varepsilon_n(k_n)$ the posterior mean satisfies condition (11); see page 507 of [19]. The inflated empirical Bayes credible set $\widehat{C}_{\hat{k}_n}(L\sqrt{\log n}, \alpha)$ is defined along the same lines. Applying again Theorems 1 and 2 together with Corollary 1, we can verify that both credible sets have high frequentist coverage and (almost) rate adaptive size under the general polished tail assumption.

PROPOSITION 4. *Consider the density estimation model with histogram priors* (24) *and assume that* $p_0 \in \mathcal{H}^\beta(L_0)$ *for some* $\beta \in [\beta_0, 1]$, $\beta_0 > 1/2$, *and it is bounded away from*

zero and infinity. Then both the inflated hierarchical Bayes and empirical Bayes credible sets with centering point $p_{\hat{\theta}}$ satisfying A0 *have (up to a* $\log n$ *factor) rate adaptive size and frequentist coverage tending to one under the polished tail assumption* (7)*, that is, for every* $\varepsilon, R_0, k_0, \tau > 0$ *there exist* $L, C > 0$ *such that*

$$\liminf_n \inf_{p_0 \in \Theta_{0,n}(R_0,k_0,\tau)\cap \mathcal{H}^{\beta_0}(L_0)} P_{p_0}^{(n)}\big(p_0 \in \widehat{C}_n(L\sqrt{\log n}, \alpha)\big) \geq 1-\varepsilon,$$

$$\liminf_n \inf_{\beta\in[\beta_0,1]} \inf_{p_0\in\mathcal{H}^{\beta}(L_0)} P_{p_0}^{(n)}\left(\operatorname{diam}\big(\widehat{C}_n(1,\alpha), h\big) \leq C\left(\frac{n}{\log n}\right)^{-\frac{\beta}{1+2\beta}}\right) \geq 1-\varepsilon,$$

where $\widehat{C}_n(L\sqrt{\log n}, \alpha)$ *could either denote the hierarchical or the empirical Bayes credible sets inflated by an* $L\sqrt{\log n}$ *multiplier.*

The proposition is verified in Section A.4 of the Supplementary Material [42].

REMARK 12. Using Lemma 3 of [42], we have $h(p_0, p_\theta) \asymp \|p_0 - p_\theta\|_2$ in a neighborhood of p_0 if k is not too large, so that the polished tail condition in the Hellinger distance is equivalent to the polished tail condition in the L_2-norm (associated to different constants). To understand the latter note that if $p_{0,[k]}$ is the L_2 projection of p_0 and $b_2(k)$ is the L_2 bias, then for any positive integer R_0, $b_2(k) = b_2(2R_0k) + \|p_{0,[k]} - p_{0,[2R_0k]}\|_2^2$ so that the L_2 polished tail condition is equivalent to

$$\|p_{0,[k]} - p_{0,[2R_0k]}\|_2^2 \geq (1-\tau)\|p_0 - p_{0,[k]}\|_2^2,$$

which has a similar flavor to the polished tail condition of [49].

3.3. *Application to density estimation with exponential families of prior.* In this subsection, we consider again the density estimation problem on $[0, 1]$, that is, we assume that we observe independent and identically distributed draws $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ from a distribution with density function f_0 (with respect to the Lebesgue measure). Then we assume that the true density can be written as an infinite dimensional exponential distribution

$$\begin{aligned} f_0(x) &= \exp\left(\sum_{j=1}^{\infty} \theta_{0,j}\phi_j(x) - c(\theta_0)\right), \quad x \in [0,1], \\ \text{with } e^{c(\theta_0)} &= \int_0^1 \exp\left(\sum_{j=1}^{\infty} \theta_{0,j}\phi_j(x)\right) dx, \end{aligned} \tag{25}$$

for some $\theta_0 = (\theta_{0,1}, \theta_{0,2}, \ldots) \in \ell_2$. For any $\theta \in \ell_2$, we define $f_\theta = \exp(\sum \theta_j\phi_j - c(\theta))$, and hence $f_0 = f_{\theta_0}$. This model is also known as the log-linear model. Furthermore, we also assume that $\|\log f_0\|_\infty < +\infty$, that $\phi_j(x)$, $j = 1, 2, \ldots$ forms an orthonormal basis (together with $\phi_0(x) \equiv 1$ and, therefore, satisfies $\int_0^1 \phi_j(x)\,dx = 0$ for all $j \geq 1$) and that $\theta_0 \in \mathcal{S}^\beta(L_0)$ for some $\beta, L_0 > 0$ as in (20).

Then we define the prior distribution on the densities with hyperparameter k by endowing the sequence θ in the log-linear model with the standard sieve prior, that is,

$$\theta = (\theta_1, \ldots, \theta_k)|k \sim \prod_{i=1}^{k} g(\theta_i),$$

$$k \sim \text{Geom}(p) \text{ or } \text{Pois}(\lambda),$$

for some fixed $p \in (0,1)$ or $\lambda > 0$ and $g(\cdot)$ satisfying (23). Alternatively, one can estimate k from the data by the MMLE and plug in the estimator $\hat{k}_n$ into the posterior distribution. Similar to Section 3.1, $\Theta(k) = \mathbb{R}^k$.

This type of priors was considered, for instance, in [1, 37, 38, 41, 53, 54], where (nearly) adaptive posterior contraction rates were derived. However, the reliability of Bayesian uncertainty quantification in this model has not yet been investigated in the literature.

By using the corresponding posterior distribution, we construct the inflated hierarchical credible set as

$$\widehat{C}(L\sqrt{\log n}, \alpha) = \{f_\theta : h(f_\theta, f_{\hat{\theta}}) \le L\sqrt{\log n} r_\alpha\},$$

where $h(\cdot,\cdot)$ denotes the Hellinger distance, the radius r_α satisfies $\pi(\theta : h(f_\theta, f_{\hat{\theta}}) \le r_\alpha | \mathbf{Y}) \ge 1-\alpha$ and $\hat{\theta}$ is an arbitrary estimator satisfying A0 with $d(\theta,\theta') = h(f_\theta, f_{\theta'})$. We note that similar to the histogram example above the posterior mean satisfies condition (11) hence can be used as a centering point of the credible set. The construction of the inflated empirical Bayes credible set $\widehat{C}_{\hat{k}_n}(L\sqrt{\log n}, \alpha)$ goes similarly. Using again Theorems 1 and 2 together with Corollary 1, we can verify that the preceding credible sets have high frequentist coverage and (nearly) rate adaptive size under the general polished tail assumption.

PROPOSITION 5. *Consider the log-linear model* (25). *Then both the inflated hierarchical and empirical Bayes credible sets have* (*up to a* $\log n$ *factor*) *rate adaptive size and frequentist coverage tending to one under the general polished tail assumption* (7), *that is*, *for every* $\beta_0 > 1/2$ *and* $\varepsilon, R_0, k_0, \tau > 0$ *there exist* $L, C > 0$ *such that*

$$\liminf_n \inf_{\theta_0 \in \Theta_{0,n}(R_0,k_0,\tau) \cap \mathcal{S}^{\beta_0}(L_0)} P_{\theta_0}^{(n)}\big(f_{\theta_0} \in \widehat{C}_n(L\sqrt{\log n}, \alpha)\big) \ge 1-\varepsilon,$$

$$\liminf_n \inf_{\beta \ge \beta_0} \inf_{\theta_0 \in \mathcal{S}^\beta(L_0)} P_{\theta_0}^{(n)}\left(\operatorname{diam}\big(\widehat{C}_n(1,\alpha), h\big) \le C\left(\frac{n}{\log n}\right)^{-\frac{\beta}{1+2\beta}}\right) \ge 1-\varepsilon,$$

where $\widehat{C}_n(L\sqrt{\log n}, \alpha)$ *denotes either the inflated hierarchical or empirical Bayes credible set with a blow up factor* $L\sqrt{\log n}$.

The proof of the proposition is given in Section A.5 of the Supplementary Material [42].

REMARK 13. In view of Lemma A.5 in the Supplementary Material, we note that the rate and coverage statements of Proposition 4 also hold for the ℓ_2-metric.

REMARK 14. Again, similar to before, if $f_{\theta_0} \in \mathcal{S}^{\beta_0}(L)$ with $\beta_0 > 1/2$ and if $k \le \bar{K}_n$ then for all $k_0 \le k \le k_n$ we have $\|\theta_0 - \theta_{0,[k]}\|_2 \le Lk^{-\beta_0}$, where $\theta_{0,[k]} = (\theta_{0,1}, \ldots, \theta_{0,k}, 0, 0, \ldots)$, and if $k_0 \ge (L/\varepsilon)^{1/\beta_0}$ with $\varepsilon > 0$ arbitrarily small, using Lemma 5 in the Supplementary Material [42],

$$b(k) \asymp \|\theta_0 - \theta_{0,[k]}\|_2^2 \quad \text{for } k \ge k_0.$$

Therefore, the parameters θ satisfying the L_2 polished tail condition of [49] (see also (8)) is a subset of $\Theta_{0,n}(R_0, k_0, \tau)$.

3.4. *Application to nonparametric classification.* In this section, we apply our general theorem to the nonparametric classification (or also known as binary regression) model. We assume to observe the sequence $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n) \in \{0,1\}^n$ satisfying

$$P(Y_i = 1 | x_i) = q_0(x_i) \quad \text{for some } q_0 : [0,1] \mapsto (0,1), \tag{26}$$

with $x_i \in [0,1]$, $i = 1, \dots, n$ fixed design points. We also take $\mu(x) = e^x/(1+e^x)$ to be the logistic link function.

We assume that under the true distribution associated to q_0, $f_0 = \mu^{-1}(q_0) \in \mathcal{S}^\beta(L_0)$, with unknown smoothness parameter $\beta > 0$. Minimax estimation rates with respect to the L_2-norm, that is, $n^{-\beta/(1+2\beta)}$ for $S^\beta(L_0)$, and an adaptive estimator achieving these rates in the random design case was derived, for instance, in [56, 57]. Similar to the density estimation and classification models, it was also shown that with respect to the L_2-norm it is impossible to construct adaptive confidence sets over a full scale of regularity classes $\bigcup_{\beta \geq \beta_0} S^\beta(L_0)$, for any $\beta_0 > 0$; see [31].

In the Bayesian approach, one endows the nonparametric function f with a prior distribution resulting in a prior on the binary regression function q. The theoretical properties of the Bayesian approach in the present model were investigated, for instance, in [20] with linear function f, in [53] with Gaussian process priors on the nonparametric function f and in [26] in context of classification of the nodes of large graphs. In the preceding papers, adaptive posterior contraction rates were derived. However, the coverage properties of Bayesian credible sets remained unknown. Due to the lack of an explicit formula for the posterior distribution direct computations are not feasible to quantify the reliability of Bayesian uncertainty quantification. Therefore, we tackle this until now unanswered question by applying our general, abstract theorem.

In our analysis, we consider again the popular sieve prior. For given k, we introduce the parametrization

$$f_\theta(x_i) = \sum_{j=1}^{k} \theta_j \phi_j(x_i) = \Phi_k(x_i)\theta,$$

with $\theta = (\theta_1, \dots, \theta_k)^T \in \Theta(k) = \mathbb{R}^k$ and $\Phi_k(x_i) = (\phi_1(x_i), \phi_2(x_i), \dots, \phi_k(x_i))$, as in Section 3.1, satisfying assumption (21). We work with the average (empirical) Hellinger pseudometric

$$h_n^2(q_1, q_2) = \frac{1}{n}\sum_{i=1}^{n} h_b^2\big(q_1(x_i), q_2(x_i)\big), \quad \text{where}$$

$$h_b\big(q_1(x_i), q_2(x_i)\big) = \big(\sqrt{q_1(x_i)} - \sqrt{q_2(x_i)}\big)^2 + \big(\sqrt{1-q_1(x_i)} - \sqrt{1-q_2(x_i)}\big)^2.$$

REMARK 15. Since assumption (21) is in a general and weak form, similar to the nonparametric regression example we have to slightly strengthen our polished tail assumption. To see this, first note that $h_n^2(q_1, q_2) \leq d_n^2(f_1, f_2)$ with $f_j(x) = \mu^{-1}(q_j(x))$, $j = 1, 2$. Similar to before, to understand the coverage properties of the credible balls, we need to study the bias function $b(k)$ with respect to the pseudometric h_n. Assume $\theta_0 \in \mathcal{S}^\beta(L_0)$ for $\beta \geq \beta_0 > 1/2$ and $L_0 > 0$. Denote by $\tilde{b}(\cdot)$ the bias function associated to $d_n(f_{\theta_0}, f_\theta)$ and studied in Section 3.1. Assume that K_n satisfies $\tilde{b}(K_n) \leq \delta K_n(\log n)/n$ for some small enough δ. Then since $b(K_n) \leq \tilde{b}(K_n)$ we get $b(K_n) \leq \delta K_n(\log n)/n$. The discussion on the feasibility of the constraint $\tilde{b}(K_n) \leq \delta K_n(\log n)/n$ is similar to that of Section 3.1. As in the case of the regression model, using (A.27) of the Supplementary Material [42], if $f_{\theta_0} \in \mathcal{S}^{\beta_0}(L)$ with $\beta_0 > 1/2$, $d_n(\theta, \theta_0) \asymp h_n(\theta, \theta_0)$ locally. Using the same arguments as in Section 3.1, if θ_0 satisfies the L_2 polished tail condition of [49], then it satisfies the general polished tail condition.

In this example, we consider the prior

$$\theta = (\theta_1, \ldots, \theta_k)|k \sim \prod_{i=1}^{k} g(\theta_i),$$

$$k \sim \mathrm{Geom}(p) \text{or} \mathrm{Pois}(\lambda),$$

with $g(\cdot)$ satisfying (23), and $p \in (0, 1)$ or $\lambda > 0$, resulting in the two level hierarchical prior $\pi(\cdot)$. Alternatively, we estimate k using the MMLE and plug it in into the posterior for θ given k. Then we consider credible balls in terms of $q(x) = \mu(f(x))$, and the empirical Hellinger pseudometric $h_n(\cdot, \cdot)$.

The inflated hierarchical Bayes credible balls are defined as

$$\widehat{C}(L\sqrt{\log n}, \alpha) = \{q_\theta(\cdot) : h_n(q_\theta, \hat{q}_n) \le L\sqrt{\log n} r_\alpha\},$$

with radius r_α given by $\pi(\theta : h_n(q_\theta, \hat{q}_n) \le r_\alpha|\mathbf{Y}) \ge 1 - \alpha$ and taking the posterior mean $\hat{q}_n = E_{\pi(\cdot|\mathbf{Y})}(q_\theta)$ as the centering point. Note that by convexity and boundedness of $q \to h_n^2(q, q_0)$, the posterior mean $\hat{q}_n$ satisfies condition (11). Alternatively, we can use any centering point satisfying condition (11). The inflated empirical Bayes credible ball $\widehat{C}_{\hat{k}_n}(L\sqrt{\log n}, \alpha)$ is defined similarly.

By applying our main Theorems 1 and 2 and Corollary 1, we show that under the general polished tail assumption (7) both of the inflated credible sets have (nearly) optimal frequentist behavior.

PROPOSITION 6. *Consider the classification model given in* (26) *with* $q_{\theta_0} = \mu(f_{\theta_0})$ *satisfying* $\theta_0 \in S^\beta(L_0)$, $\beta \ge \beta_0 > 1/2$ *and* $K_n \gg n^{\frac{1}{2(\beta_0 - 1/2)}}$. *Then the inflated credible set* $\widehat{C}_n(L\sqrt{\log n}, \alpha)$*—denoting either* $\widehat{C}(L\sqrt{\log n}, \alpha)$ *in the hierarchical approach or* $\widehat{C}_{\hat{k}_n}(L\sqrt{\log n}, \alpha)$ *in the empirical approach—have* (*up to a* $\log n$ *factor*) *rate adaptive size and frequentist coverage arbitrary close to one under the general polished tail assumption, that is, for every* $\varepsilon, R_0, k_0, \tau > 0$, *there exist constants* $L, C > 0$ *such that*

$$\liminf_n \inf_{\theta_0 \in \Theta_{0,n}(R_0, k_0, \tau) \cap \mathcal{S}^{\beta_0}(L_0)} P_{\theta_0}^{(n)}\big(q_{\theta_0} \in \widehat{C}_n(L\sqrt{\log n}, \alpha)\big) \ge 1 - \varepsilon,$$

$$\liminf_n \inf_{\beta \ge \beta_0} \inf_{\theta_0 \in \mathcal{S}^\beta(L_0)} P_{\theta_0}^{(n)}\left(\mathrm{diam}\big(\widehat{C}_n(1, \alpha), h_n\big) \le C\left(\frac{n}{\log n}\right)^{-\frac{\beta}{1+2\beta}}\right) \ge 1 - \varepsilon.$$

The proof of the proposition is deferred to Section A.6 of the Supplementary Material [42].

REMARK 16. We note that the empirical Hellinger pseudometric h_n is locally equivalent to the empirical L_2 pseudometric $d_n(f_{\theta_1}, f_{\theta_2})$ and the ℓ_2-norm $\|\theta_1 - \theta_2\|_2$; see assertions (A.27) and (A.28) in the Supplementary Material [42]. Therefore, the same coverage and contraction rate results can be shown for these pseudometrics as well.

3.5. *Parametric models and the BIC formula.* Interestingly, the lower bound obtained in Propositions 2 and 3 also holds for simple regular parametric models. Consider the same structure as model (3) with $k \le K < +\infty$, and $\Theta(k - 1) \subset \Theta(k)$. In this case, it is common knowledge that under usual regularity conditions, for a fixed $\theta_0 \in \Theta(k_0)$ with $k_0 \le K$, the posterior distribution on k converges to k_0 and that the Bernstein–von Mises theorem holds, that is, the posterior distribution of $\sqrt{n}(\theta - \hat{\theta}_{k_0})$ is asymptotically Gaussian with mean 0 and variance $I_{k_0}(\theta_0)$ where $\hat{\theta}_{k_0}$ is the maximum likelihood estimator within model $\Theta(k_0)$ and $I_{k_0}(\theta_0)$ is the Fisher information matrix (per observation, in model $\Theta(k_0)$) computed

at θ_0. However, this result does not hold uniformly, due to the $\log n$ penalization induced by integrating over the parameters. We will show below that the posteriors are not so well behaved under parameter values which are on the boundaries of the sets $\Theta(k)$, in the sense that they are close to $\Theta(k-1)$ without belonging to it.

More precisely, define the set $\Theta_n(\mathcal{K}(k_0), k_1, \delta, C, \tau)$ with $\tau, \delta \in (0,1)$, $C > 1$ and $k_1 < k_0$, as

$$\Theta_n\big(\mathcal{K}(k_0), k_1, \delta, C, \tau\big) = \Big\{\theta_0 \in \mathcal{K}(k_0); \inf_{\theta \in \Theta(k_1-1)} \|\theta_0 - \theta\|_2 \geq C\sqrt{(\log n)/n}, \\ \tau\delta\sqrt{(\log n)/n} \leq \inf_{\theta \in \Theta(k_1)} \|\theta_0 - \theta\|_2 \leq \delta\sqrt{(\log n)/n}\Big\},$$

where $\mathcal{K}(k_0)$ is a compact subset of $\Theta(k_0)$. In other words, these parameters are close to $\Theta(k_1)$ but do not belong to $\Theta(k_1)$. For instance, if $\Theta(k) = \mathbb{R}^k$ the following parameters belong to $\Theta_n(\mathcal{K}(k_0), k_1, \delta, C, \tau)$, with $\mathcal{K}(k_0) = \{\theta \in \mathbb{R}^{k_0}; |\theta_j| \leq C'\}$:

$$C' \geq |\theta_{0,j}| \geq c,\ j \leq k_1, \quad |\theta_{0,k_1+1}| \geq \tau\delta\sqrt{(\log n)/n}, \qquad \sum_{j=k_1+1}^{k_0} \theta_{0,j}^2 \leq \frac{\delta^2 \log n}{n}.$$

Note in particular that the parameters $\theta_0 \in \Theta_n(\mathcal{K}(k_0), k_1, \delta, C, \tau)$ vary with n. Then under usual regularity assumptions (see Section B.7 of the Supplementary Material) for any $\alpha \in (0,1)$, and any $L_n = o(\sqrt{\log n})$,

$$\sup_{\theta_0 \in \Theta_n(\mathcal{K}(k_0), k_1, \delta, C, \tau)} \big|E_{\theta_0}\pi(k = k_1|\mathbf{Y}) - 1\big| = o(1), \\ \sup_{\theta_0 \in \Theta_n(\mathcal{K}(k_0), k_1, \delta, C, \tau)} P_{\theta_0}\big(\theta_0 \in \hat{C}(L_n, \alpha)\big) = o(1), \tag{27}$$

where $\hat{C}(1,\alpha) = \{\|\theta - \hat{\theta}\|_2 \leq r_\alpha\}$ is the α credible ball centered at the posterior mean $\hat{\theta}$.

In other words, because of the $\log n$ penalization induced by the integration over the parameter spaces $\Theta(k)$, signals of order $\sqrt{(\log n)/n}$ may be estimated at 0 and the posterior concentrates on a smaller dimensional parameter set, thus underestimating the uncertainty.

4. Discussion. In this paper, we have provided some general tools to study the frequentist properties of inflated credible balls in infinite dimensional models based on sieve priors. We have also studied three types of models: regression, density estimation and classification. As we can see from our results, a key condition for the good behavior of these inflated balls is the fact that the posterior distribution concentrates on the values of k for which $b(k) \asymp k(\log n)/n$ and this is verified under the generalized polished tail condition, together with some other technical conditions. An intriguing feature of our result is the fact that we had to inflate the credible balls by a factor of order $\sqrt{\log n}$. In the case of the regression model, under both the empirical and hierarchical Bayes posteriors we have shown that this inflation is necessary, in order to obtain good frequentist coverage. The reason behind it is that the marginal maximum likelihood estimator $\hat{k}_n$ corresponds to a value k such that the bias $b(k) \asymp k(\log n)/n$, while the estimation error (and thus the radius r_α^2) is of order k/n. We believe that this (negative) result remains valid for the other models (density estimation and classification).

We believe that this is in fact an important takeaway message from our results, that is, the model selection priors induce a penalization of order $d_k \log n$, where d_k is the dimension of the parameter space reminiscent of the BIC formula, which in turn induce a loss in uniform coverage. This is even still true in simple regular parametric models, as discussed in Section 3.5.

From a practical perspective, these credible balls can be approximately visualized by plotting the curves under the posterior distribution which satisfy the constraint $d(\theta,\hat{\theta}) \leq L\sqrt{\log n} r_\alpha$, as was done for instance in [36] and in [49]. In general, visualization of confidence sets outside of the L_∞ or the pointwise case is challenging and we are not aware of any practical solution, for instance, for L_2- or Hellinger-confidence balls.

The paper focuses on priors based on the structure (3). This represents a general family of prior models but of course does not cover every possible prior. In particular, hierarchical priors based on a continuous hyperparameter, such as hierarchical Gaussian processes, are not tackled by the present approach. There is so far no general theory for such priors and the only existing results so far are based of particular models and particular priors for which explicit computations can be derived, as in [49].

5. Proof of Theorem 1. Theorem 1 is a simple consequence of the following lemma which allows to control the prior mass of neighborhoods of $\hat{\theta}$.

LEMMA 4. *Under the same assumptions as in Theorem 1 for every $\varepsilon > 0$, there exists a small enough $\delta_\varepsilon > 0$ such that for $\rho_n = \delta_\varepsilon/\sqrt{\log n}$,*

$$\sup_{\theta_0 \in \Theta_{0,n}} E_{\theta_0}^{(n)}\big(\pi\big(d(\theta,\hat{\theta}) \leq \rho_n \varepsilon_n(k_n)|\mathbf{Y}\big)\big) \leq \varepsilon.$$

The proof of Lemma 4 is presented in Section 5.1. We now give the proof of Theorem 1.

PROOF OF THEOREM 1. Let $L_n = L_{\varepsilon,\alpha}\sqrt{\log n}$ (for some $L_{\varepsilon,\alpha} > 0$ to be specified later) and $\varepsilon_n = \varepsilon_n(k_n)$. Then by assumption (11) and definition (5) we have for every $\varepsilon > 0$ that

$$\begin{aligned}
&P_{\theta_0}^{(n)}\big(\theta_0 \in \widehat{C}(L_n,\alpha)\big)\\
&\quad = P_{\theta_0}^{(n)}\big(d(\theta_0,\hat{\theta}) \leq L_n r_\alpha\big)\\
&\quad \geq P_{\theta_0}^{(n)}\big[\pi\big(d(\theta,\hat{\theta}) \leq d(\theta_0,\hat{\theta})/L_n|\mathbf{Y}\big) \leq 1-\alpha\big]\\
&\quad \geq P_{\theta_0}^{(n)}\big[\pi\big(\theta : d(\theta,\hat{\theta}) \leq M_\varepsilon \varepsilon_n/L_n|\mathbf{Y}\big) \leq 1-\alpha\big] - \varepsilon.
\end{aligned}$$

We show below that the first term on the right-hand side is bounded from below by $1-\varepsilon$. In view of Lemma 4, there exists $\delta_{\varepsilon,\alpha} > 0$ small enough such that

$$\sup_{\theta_0 \in \Theta_{0,n}} E_{\theta_0}^{(n)}\big(\pi\big(d(\theta,\hat{\theta}) \leq \delta_{\varepsilon,\alpha}\varepsilon_n/\sqrt{\log n}|\mathbf{Y}\big)\big) \leq \varepsilon(1-\alpha),$$

and, therefore, by taking $L_{\varepsilon,\alpha} = M_\varepsilon/\delta_{\varepsilon,\alpha}$ and applying Markov's inequality

$$P_{\theta_0}^{(n)}\left[\pi\left(d(\theta,\hat{\theta}) \leq \frac{M_\varepsilon \varepsilon_n}{L_n}|\mathbf{Y}\right) > 1-\alpha\right] \leq \frac{E_{\theta_0}^{(n)}(\pi(d(\theta,\hat{\theta}) \leq \frac{\delta_{\varepsilon,\alpha}\varepsilon_n}{\sqrt{\log n}}|\mathbf{Y}))}{1-\alpha} \leq \varepsilon,$$

completing the proof of our statement. □

5.1. *Proof of Lemma 4.* For notational convenience, let $\varepsilon_n = \varepsilon_n(k_n)$ and $\Theta_n = \bigcup_k \Theta_n(k)$. Then in view of Lemma 2, for large enough choice of $M > 0$,

$$\begin{aligned}
&E_{\theta_0}^{(n)}\pi\big(d(\theta,\hat{\theta}) \leq \rho_n\varepsilon_n|\mathbf{Y}\big)\\
&\quad \leq E_{\theta_0}^{(n)}\Bigg(\sum_{k\in\mathcal{K}_n(M)} \pi_{|k}\big(\{d(\theta,\hat{\theta}) \leq \rho_n\varepsilon_n\} \cap \Theta_n(k)|\mathbf{Y}\big)\pi_k(k|\mathbf{Y})\Bigg)\\
&\qquad + E_{\theta_0}^{(n)}\Bigg(\sum_{k\in\mathcal{K}_n(M)} \pi_{|k}\big(\Theta_n(k)^c|\mathbf{Y}\big)\pi_k(k|\mathbf{Y})\Bigg) + \varepsilon
\end{aligned}$$

for all $\theta_0 \in \Theta_{0,n}$. Let $\Omega_{n,0} = \{m_n(k_n) > e^{-(c_3+c_4+1)n\varepsilon_n^2}\}$, with $m_n(k) = \int_{\Theta(k)} e^{\ell_n(\theta)-\ell_n(\theta_0)} \times \pi_{|k}(d\theta)$. In view of Lemma 10 of [20] (with $n\varepsilon^2 = n\varepsilon_n^2(k_n) \le 2k_n \log n$, $k = r$ and $C = 1$), we get following A1 that $P_{\theta_0}^{(n)}(\Omega_{n,0}^c) \le (k_n \log n)^{-r/2} = o(1)$. Furthermore, note that for $k \in \mathcal{K}_n(M)$ we have by Assumption A2(i) that

$$E_{\theta_0}^{(n)}\big[\mathbf{1}_{\Omega_{n,0}}\pi_{|k}\big(\Theta_n(k)^c|\mathbf{Y}\big)\big] = \pi_{|k}\big(\Theta_n(k)^c\big)ee^{(c_3+c_4+1)n\varepsilon_n^2} \lesssim e^{-(c_2+1)n\varepsilon_n^2}.$$

By combining the above inequalities and in view of Lemma 1, we get that

$$E_{\theta_0}^{(n)}\left(\mathbf{1}_{\Omega_{n,0}} \sum_{k\in\mathcal{K}_n(M)} \pi_{|k}\big(\Theta_n(k)^c|\mathbf{Y}\big)\pi_k(k|\mathbf{Y})\right) \lesssim k_n e^{-n\varepsilon_n^2} = o(1).$$

Next, we show that with probability at least $1 - \tilde{C}\varepsilon$, for some universal $\tilde{C} > 0$, we have for every $k \in \mathcal{K}_n(M)$,

$$\pi_{n,k} := \pi_{|k}\big(\big\{d(\theta,\hat{\theta}) \le \rho_n\varepsilon_n\big\} \cap \Theta_n(k)|\mathbf{Y}\big) \le \varepsilon. \tag{28}$$

Then the statement of the lemma follows by noting that

$$E_{\theta_0}^{(n)}\pi\big(d(\theta,\hat{\theta}) \le \rho_n\varepsilon_n|\mathbf{Y}\big) \le \big(\tilde{C} + 1 + o(1)\big)\varepsilon + \sum_{k\in\mathcal{K}_n(M)} \varepsilon\pi_k(k|Y) \le (\tilde{C}+3)\varepsilon.$$

It remained to prove (28). As a first step, we introduce the notation, for $C, B > 0$,

$$\Omega_n(C) = \left\{\max_{k\in\mathcal{K}_n(M)} e^{Ck} \frac{\int_{\Theta(k)} e^{\ell_n(\theta)-\ell_n(\theta_{[k]}^o)}\pi_{|k}(d\theta)}{\pi_{|k}(d(\theta,\theta_{[k]}^o)^2 \le k/n)} \ge 1\right\}, \tag{29}$$

$$\Gamma_n(B) = \left\{\max_{k\in\mathcal{K}_n(M)} \sup_{\Theta_n(k)\cap B_k(\hat{\theta},\rho_n\varepsilon_n,d)} \big(\ell_n(\theta) - \ell_n\big(\theta_{[k]}^o\big) - Bk\big) < 0\right\}. \tag{30}$$

Using Assumption A0, we have with probability greater than $1-\varepsilon$, $d(\hat{\theta},\theta_0) \le M_\varepsilon\varepsilon_n$, therefore, as soon as $\rho_n \le 1$,

$$B_k(\hat{\theta},\rho_n\varepsilon_n,d) \subset B_k\big(\theta_0,(M_\varepsilon+1)\varepsilon_n,d\big).$$

Hence in view of Assumption A4(ii), there exists a large enough constant $B_\varepsilon > 0$ such that $\inf_{\theta_0\in\Theta_{0,n}} P_{\theta_0}^{(n)}(\Gamma_n(B_\varepsilon)) \ge 1-\varepsilon$. Also note that following from A4(i) and by using the standard technique for lower bound for the likelihood ratio (e.g., Lemma 10 of [20] with $1+C = c_7 + 1/\sqrt{\varepsilon}$ and $n\varepsilon^2 = k$) we have, for any $k \in \mathcal{K}_n(M)$, with $P_{\theta_0}^{(n)}$-probability bounded from below by $1-(\varepsilon/k)^{r/2} \ge 1-\varepsilon/k$ that

$$\begin{aligned}
&\int_{\Theta(k)} e^{\ell_n(\theta)-\ell_n(\theta_{[k_n]}^o)}\pi_{|k}(d\theta)\\
&\quad\ge e^{-(c_7+1/\sqrt{\varepsilon})k}\pi_{|k}\big(S_n(k,c_7,c_8,r)\big)\\
&\quad\ge e^{-(c_7+1/\sqrt{\varepsilon})k}\pi_{|k}\big(B_k\big(\theta_{[k]}^o,\sqrt{k/n},d\big)\big),
\end{aligned} \tag{31}$$

hence in view of Lemma 1, $P_{\theta_0}^{(n)}(\Omega_n^c(c_7+1/\sqrt{\varepsilon})) \le C\varepsilon$.

Then we have, on $\Omega_n(c_7+1/\sqrt{\varepsilon}) \cap \Gamma_n(B_\varepsilon)$, that for any $k \in \mathcal{K}_n(M)$,

$$\pi_{n,k} \le e^{(c_7+B_\varepsilon+1/\sqrt{\varepsilon})k} \frac{\pi_{|k}(\Theta_n(k) \cap \{d(\theta,\hat{\theta}) \le \rho_n\varepsilon_n\})}{\pi_{|k}(d(\theta,\theta_{[k]}^o) \le \sqrt{k/n})}.$$

We recall that Lemma 1 implies that $k_n \leq Ck$ for all $k \in \mathcal{K}_n(M)$ and by definition of $\varepsilon_n(k_n)$, $n\varepsilon_n^2 \leq 2k_n \log n$. Therefore, we have $\rho_n \varepsilon_n \leq \delta_\varepsilon \sqrt{2k_n/n} \leq (2C)^{1/2}\delta_\varepsilon \sqrt{k/n}$ for all $k \in \mathcal{K}_n(M)$. In view of Assumption A4(iii) (with $\delta_{n,k} = (2C)^{1/2}\delta_\varepsilon$),

$$\pi_{n,k} \lesssim e^{(c_7 + B_\varepsilon + 1/\sqrt{\varepsilon} + c_9 \log(\sqrt{2C}\delta_\varepsilon))k} \leq \varepsilon, \tag{32}$$

for small enough choice of $\delta_\varepsilon > 0$ (the choice $\log(\delta_\varepsilon) \leq -c_9^{-1}(1/\sqrt{\varepsilon} + c_7 + B_\varepsilon + \log \varepsilon^{-1}) - \log \sqrt{2C}$ is sufficiently small).

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. The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

This work was supported in part by the Chaire Havas.

The first author was supported in part by the ANR IPANEMA, the labex ECODEC.

The second author was supported by Netherlands Organization for Scientific Research NWO.

SUPPLEMENTARY MATERIAL

Supplement to "Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors" (DOI: 10.1214/19-AOS1881SUPP; .pdf). Supplementary information.

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 https://doi.org/10.1002/sjos.12002

[2] BELITSER, E. (2017). On coverage and local radial rates of credible sets. *Ann. Statist.* **45** 1124–1151. MR3662450 https://doi.org/10.1214/16-AOS1477

[3] BELITSER, E. and NURUSHEV, N. (2015). Needles and straw in a haystack: Empirical Bayes confidence for possibly sparse sequences. Preprint. Available at arXiv:1511.01803.

[4] BULL, A. D. (2012). Honest adaptive confidence bands and self-similar functions. *Electron. J. Stat.* **6** 1490–1516. MR2988456 https://doi.org/10.1214/12-EJS720

[5] BULL, A. D. and NICKL, R. (2013). Adaptive confidence sets in L^2. *Probab. Theory Related Fields* **156** 889–919. MR3078289 https://doi.org/10.1007/s00440-012-0446-z

[6] CAI, T. T. and LOW, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** 1805–1840. MR2102494 https://doi.org/10.1214/009053604000000049

[7] CARPENTIER, A. (2013). Honest and adaptive confidence sets in L_p. *Electron. J. Stat.* **7** 2875–2923. MR3148371 https://doi.org/10.1214/13-EJS867

[8] CARPENTIER, A. and NICKL, R. (2015). On signal detection and confidence sets for low rank inference problems. *Electron. J. Stat.* **9** 2675–2688. MR3432430 https://doi.org/10.1214/15-EJS1087

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

[10] CASTILLO, I. and MISMER, R. (2018). Empirical Bayes analysis of spike and slab posterior distributions. *Electron. J. Stat.* **12** 3953–4001. MR3885271 https://doi.org/10.1214/18-EJS1494

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

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

[13] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** 2353–2383. MR3405597 https://doi.org/10.1214/15-AOS1336

[14] CASTILLO, I. and SZABO, B. (2020). Spike and slab empirical Bayes sparse credible sets. *Bernoulli* **26** 127–158. https://doi.org/10.3150/19-BEJ1119

[15] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.* **42** 1787–1818. MR3262468 https://doi.org/10.1214/14-AOS1235

[16] COX, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923. MR1232525 https://doi.org/10.1214/aos/1176349157

[17] FREEDMAN, D. (1999). On the Bernstein–von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* **27** 1119–1140. MR1740119 https://doi.org/10.1214/aos/1017938917
[18] GAO, C. and ZHOU, H. H. (2016). Rate exact Bayesian adaptation with modified block priors. *Ann. Statist.* **44** 318–345. MR3449770 https://doi.org/10.1214/15-AOS1368
[19] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228
[20] 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 https://doi.org/10.1214/009053606000001172
[21] GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. MR2604707 https://doi.org/10.1214/09-AOS738
[22] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge Univ. Press, New York. MR3588285 https://doi.org/10.1017/CBO9781107337862
[23] HADJI, A. and SZABO, B. (2019). Can we trust bayesian uncertainty quantification from Gaussian process priors with squared exponential covariance kernel? Preprint. Available at arXiv:1904.01383.
[24] HOFFMANN, M. and NICKL, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409. MR2906872 https://doi.org/10.1214/11-AOS903
[25] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **43** 2259–2295. MR3396985 https://doi.org/10.1214/15-AOS1341
[26] KIRICHENKO, A. and VAN ZANTEN, H. (2017). Estimating a smooth function on a large graph by Bayesian Laplacian regularisation. *Electron. J. Stat.* **11** 891–915. MR3629018 https://doi.org/10.1214/17-EJS1253
[27] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39** 2626–2657. MR2906881 https://doi.org/10.1214/11-AOS920
[28] LEAHU, H. (2011). On the Bernstein–von Mises phenomenon in the Gaussian white noise model. *Electron. J. Stat.* **5** 373–404. MR2802048 https://doi.org/10.1214/11-EJS611
[29] LOW, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. MR1604412 https://doi.org/10.1214/aos/1030741084
[30] MCVINISH, R., ROUSSEAU, J. and MENGERSEN, K. (2009). Bayesian goodness of fit testing with mixtures of triangular distributions. *Scand. J. Stat.* **36** 337–354. MR2528988 https://doi.org/10.1111/j.1467-9469.2008.00620.x
[31] MUKHERJEE, R. and SEN, S. (2018). Optimal adaptive inference in random design binary regression. *Bernoulli* **24** 699–739. MR3706774 https://doi.org/10.3150/16-BEJ893
[32] NICKL, R. and SZABÓ, B. (2016). A sharp adaptive confidence ball for self-similar functions. *Stochastic Process. Appl.* **126** 3913–3934. MR3565485 https://doi.org/10.1016/j.spa.2016.04.017
[33] NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450 https://doi.org/10.1214/13-AOS1170
[34] PETRONE, S., ROUSSEAU, J. and SCRICCIOLO, C. (2014). Bayes and empirical Bayes: Do they merge? *Biometrika* **101** 285–302. MR3215348 https://doi.org/10.1093/biomet/ast067
[35] PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28** 298–335. MR1762913 https://doi.org/10.1214/aos/1016120374
[36] RAY, K. (2017). Adaptive Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **45** 2511–2536. MR3737900 https://doi.org/10.1214/16-AOS1533
[37] RIVOIRARD, V. and ROUSSEAU, J. (2012). Bernstein–von Mises theorem for linear functionals of the density. *Ann. Statist.* **40** 1489–1523. MR3015033 https://doi.org/10.1214/12-AOS1004
[38] RIVOIRARD, V. and ROUSSEAU, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Anal.* **7** 311–333. MR2934953 https://doi.org/10.1214/12-BA710
[39] ROBINS, J. and VAN DER VAART, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* **34** 229–253. MR2275241 https://doi.org/10.1214/009053605000000877
[40] ROUSSEAU, J. (2007). Approximating interval hypothesis: p-values and Bayes factors. In *Bayesian Statistics* 8 (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). *Oxford Sci. Publ.* 417–452. Oxford Univ. Press, Oxford. MR2433202
[41] ROUSSEAU, J. and SZABO, B. (2017). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *Ann. Statist.* **45** 833–865. MR3650402 https://doi.org/10.1214/16-AOS1469
[42] ROUSSEAU, J. and SZABO, B. (2020). Supplement to "Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors." https://doi.org/10.1214/19-AOS1881SUPP.
[43] RUDELSON, M. (1999). Random vectors in the isotropic position. *J. Funct. Anal.* **164** 60–72. MR1694526 https://doi.org/10.1006/jfan.1998.3384

[44] SCRICCIOLO, C. (2007). On rates of convergence for Bayesian density estimation. *Scand. J. Stat.* **34** 626–642. MR2368802 https://doi.org/10.1111/j.1467-9469.2006.00540.x

[45] SERRA, P. and KRIVOBOKOVA, T. (2017). Adaptive empirical Bayesian smoothing splines. *Bayesian Anal.* **12** 219–238. MR3597573 https://doi.org/10.1214/16-BA997

[46] 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 https://doi.org/10.1214/15-EJS1078

[47] SÖHL, J. and TRABS, M. (2016). Adaptive confidence bands for Markov chains and diffusions: Estimating the invariant measure and the drift. *ESAIM Probab. Stat.* **20** 432–462. MR3581829 https://doi.org/10.1051/ps/2016017

[48] SZABÓ, B. (2015). On Bayesian based adaptive confidence sets for linear functionals. In *Bayesian Statistics from Methods to Models and Applications*. *Springer Proc. Math. Stat.* **126** 91–105. Springer, Cham. MR3374423 https://doi.org/10.1007/978-3-319-16238-6_8

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

[50] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

[51] VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* **12** 1221–1274. MR3724985 https://doi.org/10.1214/17-BA1065

[52] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[53] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663 https://doi.org/10.1214/009053607000000613

[54] VERDINELLI, I. and WASSERMAN, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.* **26** 1215–1241. MR1647645 https://doi.org/10.1214/aos/1024691240

[55] WEIMIN YOO, W. and VAN DER VAART, A. W. (2017). The Bayes Lepski's method and credible bands through volume of tubular neighborhoods. ArXiv e-prints.

[56] YANG, Y. (1999). Minimax nonparametric classification. I. Rates of convergence. *IEEE Trans. Inform. Theory* **45** 2271–2284. MR1725115 https://doi.org/10.1109/18.796368

[57] YANG, Y. (1999). Minimax nonparametric classification. II. Model selection for adaptation. *IEEE Trans. Inform. Theory* **45** 2285–2292. MR1725116 https://doi.org/10.1109/18.796369

[58] YOO, W. W. and GHOSAL, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.* **44** 1069–1102. MR3485954 https://doi.org/10.1214/15-AOS1398