Bernoulli **26**(1), 2020, 127–158
https://doi.org/10.3150/19-BEJ1119

Spike and slab empirical Bayes sparse credible sets

ISMAËL CASTILLO[1] and BOTOND SZABÓ[2]

[1]*Sorbonne Université, Laboratoire Probabilités, Statistique et Modélisation; 4, place Jussieu, 75005 Paris, France. E-mail: ismael.castillo@upmc.fr*
[2]*Leiden University, The Netherlands. E-mail: b.t.szabo@math.leidenuniv.nl*

In the sparse normal means model, coverage of adaptive Bayesian posterior credible sets associated to spike and slab prior distributions is considered. The key sparsity hyperparameter is calibrated via marginal maximum likelihood empirical Bayes. First, adaptive posterior contraction rates are derived with respect to d_q-type-distances for $q \leq 2$. Next, under a type of so-called excessive-bias conditions, credible sets are constructed that have coverage of the true parameter at prescribed $1-\alpha$ confidence level and at the same time are of optimal diameter. We also prove that the previous conditions cannot be significantly weakened from the minimax perspective.

Keywords: convergence rates of posterior distributions; credible sets; empirical Bayes; spike and slab prior distributions

1. Introduction

1.1. Setting

In the sparse normal means model, one observes a sequence $X = (X_1, \ldots, X_n)$

$$X_i = \theta_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

with $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ and $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\mathcal{N}(0,1)$. Given θ, the distribution of X is a product of Gaussians and is denoted by P_θ. Further, one assumes that the 'true' vector θ_0 belongs to

$$\ell_0[s] = \big\{\theta \in \mathbb{R}^n, \big|\{i : \theta_i \neq 0\}\big| \leq s\big\},$$

the set of vectors that have at most s nonzero coordinates, where s is a sequence such that $s/n = o(1)$ and $s \to \infty$ as $n \to \infty$. A natural problem is that of reconstructing θ with respect to the ℓ^q-type-metric for $0 < q \leq 2$ (it is a true metric only for $q \leq 1$) defined by

$$d_q\big(\theta, \theta'\big) = \sum_{i=1}^{n} \big|\theta_i - \theta_i'\big|^q.$$

A benchmark is given by the minimax rate for this loss over the class of sparse vectors $\ell_0[s]$. The minimax rate over $\ell_0[s]$ for the loss d_q is of the order, as $n \to \infty$, see [13],

$$r_q := r_{n,q} = s\big[\log(n/s)\big]^{q/2}.$$

The sparse sequence model has become very central in statistics as one of the simplest and natural models to describe sparsity, in a similar way as the Gaussian white noise model in the setting of nonparametrics. Many authors have contributed to its study both from Bayesian and non-Bayesian perspectives, in particular in terms of convergence rates. Some seminal contributions include [1,6,16]. Methods using an empirical Bayes approach to study aspects of the posterior distribution include works by George and Foster [14], Johnstone and Silverman [19] (whose approach we describe in more detail in Sections 1.3–1.4) and Jiang and Zhang [17]. Works studying the full posterior have started more recently, and we include a brief overview below. Here our interest is in a popular class of Bayesian procedures associated to spike and slab prior distributions. We undertake a so-called frequentist analysis of the posterior distribution. That is, we first construct a prior distribution on the unknown sparse θ and then use the Bayesian framework to produce a posterior distribution, which is then studied under the frequentist assumption that the data has actually been generated from a 'true' unknown sparse parameter θ_0.

Our interest is in precise understanding of how posterior distributions for spike and slab priors work for *inference* in terms of convergence and confidence sets. Such priors play a central role in statistics, in sparse and non-sparse settings (such as nonparametric function estimation, see, for example, [21]), and also as tools for lower bounds. In sparsity contexts, especially for $\ell_0[s]$ classes, they are one of the most natural choice of priors. Despite recent advances, there are many open questions regarding mathematical properties of such fundamental priors for inference. A brief overview of the literature on sparse priors is given below. We note also that the present work is a natural continuation of [9], where rates of convergence in the case $q = 2$ were investigated. Here we handle the fundamentally different issue of building confidence regions, as well as posterior convergence rates, with respect to ℓ_q-type-metrics for all q in $(0, 2]$.

The construction of confidence sets is of key importance in statistics, but is a delicate issue. For convenience let us formally denote by $\{\Theta_\beta : \beta \in B\}$ a collection of models indexed by some parameter $\beta \in B$ (e.g. sparsity, regularity, dimension, etc.). In practice it is typically unknown which model Θ_β the true θ belongs to, hence one wants to develop adaptive methods not relying on the knowledge of β. Constructing adaptive confidence sets in high-dimensional and nonparametric problems is very challenging, in fact impossible in general, see, for instance, [15,22,25] in context of nonparametric models and [24] in (sparse) high dimensional problems. Therefore, it is sometimes necessary to introduce further assumptions on the models Θ_β, $\beta \in B$ to derive positive results, see [24] for more detailed description of the problem in the high-dimensional setting as well as Sections 2.4–2.6 below.

In various fields of applications, for their flexibility and practical convenience, Bayesian credible sets are routinely used as a measure of uncertainty. However, it is not immediately clear what the frequentist interpretation of these sets is, that is, whether such sets can be used as confidence sets or whether by doing so one provides a misleading haphazard uncertainty statement. The asymptotic properties of Bayesian credible sets have been investigated only in recent years, see, for instance, [10,28,30] and references therein. In the context of sparse high dimensional

problems, there are only a few results available. For the sparse normal means model, the frequentist coverage properties of a sparsity prior with empirically chosen Gaussian slabs [4] and of the horseshoe prior [32] were investigated, while in the more general linear regression model credible sets for the modified Gaussian slab prior are studied in [3], all for the quadratic risk. In the present paper, the focus is on the standard and popular spike and slab prior, which requires a substantially different analysis compared to the previous examples, as explained in more details below.

1.2. Spike and slab prior and associated posterior distribution

The spike and slab prior with sparsity parameter α is the prior Π_α on θ given by

$$\theta \sim \bigotimes_{i=1}^{n} \big((1-\alpha)\delta_0 + \alpha G(\cdot)\big) =: \Pi_\alpha, \tag{2}$$

where δ_0 denotes the Dirac mass at 0 and G is a given probability measure of density γ. It is often assumed that γ is a symmetric unimodal density on $\mathbb{R}$. We will make specific choices in the sequel. The posterior distribution under (1)–(2) is

$$\Pi_\alpha[\cdot \mid X] \sim \bigotimes_{i=1}^{n} \big(\big(1-a(X_i)\big)\delta_0 + a(X_i)\gamma_{X_i}(\cdot)\big), \tag{3}$$

where we have set, denoting ϕ the standard normal density and $g(x) = \phi * \gamma(x) = \int \phi(x-u)\,dG(u)$ the convolution of ϕ and G,

$$\begin{aligned} g(X_i) &= (\phi * \gamma)(X_i), \\ \gamma_{X_i}(\cdot) &= \frac{\phi(X_i - \cdot)\gamma(\cdot)}{g(X_i)}, \\ a(X_i) &= a_\alpha(X_i) = \frac{\alpha g(X_i)}{(1-\alpha)\phi(X_i) + \alpha g(X_i)}. \end{aligned}$$

If the choice of α is clear from the context, we denote $a(x)$ instead of $a(\alpha, x)$ for simplicity.

Introducing the posterior median threshold

For any symmetric γ density, the vector $\hat{\theta}_\alpha$ of medians of the coordinates of the posterior (3) (whose ith coordinate by (3) only depends on X_i) has been studied in [19]. The following property is used repeatedly in what follows, see Lemma 2 in [19]: the posterior coordinate-wise median has a thresholding property: there exists $t(\alpha) > 0$ such that $\hat{\theta}_\alpha(X)_i = 0$ if and only if $|X_i| \le t(\alpha)$.

1.3. Empirical Bayes estimation of α via marginal likelihood

In a seminal paper, Johnstone and Silverman [19] considered estimation of θ using spike and slab priors combined with a very simple empirical Bayes method for choosing α that we also follow here and describe next. The marginal likelihood in α is the density of $X \mid \alpha$ at the observation points in the Bayesian model. A simple calculation reveals that its logarithm equals

$$\ell(\alpha) = \ell_n(\alpha; X) = \sum_{i=1}^{n} \log\big((1-\alpha)\phi(X_i) + \alpha g(X_i)\big).$$

The corresponding score function equals $S(\alpha) := \ell'(\alpha) = \sum_{i=1}^{n} B(X_i, \alpha)$, where

$$B(x) = \frac{g}{\phi}(x) - 1; \qquad B(x, \alpha) = \frac{B(x)}{1 + \alpha B(x)}. \tag{4}$$

Then [19] define $\hat{\alpha}$ as the maximiser, henceforth abbreviated as MMLE, of the log-likelihood

$$\hat{\alpha} = \operatorname*{argmax}_{\alpha \in \mathcal{A}_n} \ell_n(\alpha; X), \tag{5}$$

where $\mathcal{A}_n = [\alpha_n, 1]$, and α_n is defined by, with $t(\alpha)$ the posterior median threshold as above,

$$t(\alpha_n) = \sqrt{2 \log n}.$$

1.4. Motivating risk results

Let as above $\hat{\theta}_\alpha$ denote the posterior coordinate-wise median associated to the posterior (3) with fixed hyper-parameter α and let $\hat{\theta} = \hat{\theta}_{\hat{\alpha}}$, with $\hat{\alpha}$ as in (5). Suppose that $s = o(n)$ as $n \to \infty$ and that for some constant $\kappa_1 > 0$,

$$\kappa_1 \log^2 n \le s. \tag{6}$$

Fact 1 (Direct consequence of [19], Theorem 1). Let γ be the Laplace or the Cauchy density. Suppose (6) holds. For any $0 < q \le 2$, there exists a constant $C = C(q, \gamma) > 0$ such that

$$\sup_{\theta_0 \in \ell_0[s]} E_{\theta_0} d_q(\hat{\theta}, \theta_0) \le C r_q,$$

thereby proving minimaxity (up to a constant multiplier) of the estimator $\hat{\theta} = \hat{\theta}_{\hat{\alpha}}$ over $\ell_0[s]$. The estimator is adaptive, as the knowledge of s is not required in its construction. Condition (6) is quite mild. In case it is not satisfied, the upper bound on the rate above is $Cr_q + \log^3 n$ instead of Cr_q, which means there may be a slight logarithmic penalty to use $\hat{\theta}$ in the extremely sparse situation where $s \ll \log^2 n$. Theorem 2 in [19] shows that the estimate $\hat{\alpha}$ can in fact be modified so that the minimax risk result holds even if the lower bound in (6) is not satisfied. For simplicity in the present paper, we work under (6) but presumably modifying the estimator as in [19] leads

to minimax optimality also in the extremely sparse range in the context of Theorem 1. In this respect, we note that [9], Theorem 5, shows that this is indeed the case when $q = 2$ and a Cauchy prior is used.

Consider the plug-in empirical Bayes posterior, for $\hat{\alpha}$ the MMLE as defined above,

$$\Pi_{\hat{\alpha}}[\cdot \mid X] \sim \bigotimes_{i=1}^{n} \big((1 - a_{\hat{\alpha}}(X_i))\delta_0 + a_{\hat{\alpha}}(X_i)\gamma_{X_i}(\cdot) \big).$$

Fact 2 ([9], Theorems 1 and 3). Let $\hat{\alpha}$ be the MMLE given by (5). Let γ be the Cauchy density. Under (6) there exists $C > 0$ such that, for n large enough,

$$\sup_{\theta_0 \in \ell_0[s]} E_{\theta_0} \int d_2(\theta, \theta_0)\, d\Pi_{\hat{\alpha}}(\theta \mid X) \leq Cs \log(n/s).$$

For γ the Laplace density, the result does not hold: there exists $\theta_0 \in \ell_0[s]$ and $c > 0$ such that

$$E_{\theta_0} \int d_2(\theta, \theta_0)\, d\Pi_{\hat{\alpha}}(\theta \mid X) \geq cM_n s \log(n/s),$$

where $M_n = \exp\{\sqrt{\log(n/s)}\}$ goes to infinity with n/s. This shows that if tails of the slab prior are not heavy enough, the corresponding posterior does not reach the optimal minimax rate over sparse classes. In particular, typical credible sets such as balls arising from this posterior will not have optimal diameter. These observations naturally lead to wonder if confidence sets in the squared euclidean norm $d_2 = \|\cdot\|^2$ could be obtained using a Cauchy slab, for which the optimal posterior contraction rate is guaranteed, and how the previous facts evolve if d_q-type-metrics for $q < 2$ are considered.

1.5. Brief overview of results on sparse priors

Many popular sparse priors can be classified into two categories: first, priors that put some coefficients to the exact zero value, such as spike and slab priors and second, priors that instead draw coefficients using absolutely continuous distributions, and thus do not generate exact zero values. In the first category, one can generalise the spike and slab prior scheme (2) by first selecting a random subset S of indexes within $\{1, \ldots, n\}$ and then given S setting $\theta_i = 0$ for $i \notin S$ and drawing θ_i for $i \in S$ from some absolutely continuous prior distribution. This scheme has been considered for example, in [12], where the case of an induced prior π_n on the number of non-zero coefficients of the form $\pi_n(k) \propto \exp[-c_1 k \log(c_2 n/k)]$, called complexity prior, is studied and the slab distribution has tails at least as heavy as Laplace. Belitser and coauthors [3,4], and Martin and Walker [23], consider the case of Gaussian slabs that are recentered at the observation points. Other proposals for slab distributions include non-local priors as in [18].

In the second category, one can replace the Dirac mass at zero of the spike by a density approaching it, as in the spike and slab LASSO introduced by Ročková and George, see [26,27]. One can also directly define a certain continuous density with a lot of mass at zero and heavy

tails, as does the horseshoe prior introduced in [8] and further studied in [31–33], see also [34] for other families of mixture priors. Other approaches to continuous shrinkage priors include the Dirichlet–Laplace priors of [5].

Most previously cited works are concerned with posterior convergence rates, with the exception of [4]–[3] (that considers also oracle results) and [32], that derive properties of credible sets, all with respect to the squared ℓ^2-loss. The prior and confidence sets introduced in [4] are quite different from those considered here, in that, for instance, the radius of credible set we consider is determined directly from the posterior distribution, and the priors and confidence sets in [4] require some post-processing (e.g., specific separate estimation of the radius and recentering of the posterior selected components). As noted above, the horseshoe prior belongs to a different category of priors, not setting any coefficient to 0, and further, it is not clear if its Cauchy tails would be sufficiently heavy to handle d_q-losses for small q, at least via a MMLE–empirical Bayes choice of its tuning parameter τ. An overview of current research can be found in the discussion paper [32].

1.6. Outline and notation

Outline and summary of main results

Section 2 contains our main results. First, adaptive convergence rates in d_q-type-distances, $0 < q \leq 2$, are derived for the full empirical Bayes posterior $\Pi_{\hat{\alpha}}[\cdot \mid X]$, for a well-chosen slab distribution. Second, frequentist coverage results are obtained for credible balls centered at the posterior median estimator $\hat{\theta}$ and whose radius is a constant M times the posterior expected radius $\int d_q(\theta, \hat{\theta})\, d\Pi(\theta \mid X)$, both for deterministic and data-driven choice of the hyper-parameter α. In the later case, we prove that under an excessive-bias condition, the credible sets have optimal diameter and frequentist coverage, already for fixed large enough M (so without the need of a 'blow-up' $M = M_n \to \infty$). Focusing on the case $q = 2$, we then discuss the obtained excessive-bias condition and show that such a condition cannot be weakened from the minimax perspective. Section 3 briefly discusses the main findings of the paper. Proofs are organised as follows: Section 4 regroups some useful preliminary bounds, Section 5 is devoted to proofs for credible sets. A separate supplementary material [11] gathers proofs of technical lemmas, as well as the proof of the rate Theorem 1.

Notation. For two sequences a_n, b_n let us write $a_n \lesssim b_n$ if there exists a universal constant $C > 0$ such that $a_n \leq C b_n$, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously. We write $a_n \sim b_n$ for $b_n \neq 0$ if $a_n / b_n = 1 + o(1)$. We denote throughout by c and C universal constants whose value may change from line to line. Also, the d_q-diameter of a set $\mathcal{C}$ is written $\operatorname{diam}_q(\mathcal{C})$, that is,

$$\operatorname{diam}_q(\mathcal{C}) = \sup_{\theta, \theta' \in \mathcal{C}} d_q\big(\theta, \theta'\big).$$

For convenience in the case $q = 2$, we denote by $\operatorname{diam}(\mathcal{C})$ the d_2-diameter of the set $\mathcal{C}$. Finally, when $\mathcal{D}$ is a finite set, $|\mathcal{D}|$ denotes the cardinality of $\mathcal{D}$.

2. Main results

2.1. Slab prior distributions

For a fixed $\delta \in (0, 2)$, consider the unimodal symmetric density $\gamma = \gamma_\delta$ on $\mathbb{R}$ given by

$$\gamma(x) = \frac{c_\gamma}{(1+|x|)^{1+\delta}} = c_\gamma \Delta\big(1+|x|\big), \quad \text{for } \Delta(u) = u^{-1-\delta}, \tag{7}$$

where $c_\gamma = c_\gamma(\delta)$ is the normalising constant making γ a density. The purpose of this density is to have sufficiently heavy tails, possibly heavier than Cauchy. To fix ideas, we take the specific form (7), but as is apparent from the proofs, similar results continue to hold for densities with same tails: for instance, the Cauchy density could be used instead of γ_2. The possibility of having a broad range of heavy tails is essential to achieve optimal rates in terms of d_q for the considered empirical Bayes procedure. If $\delta \geq 1$, the function $u \to (1+u^2)\gamma(u)$ is bounded and the density γ falls in the framework of [19]. If $\delta \in (0, 1)$, we show below how this changes estimates quantitatively. In all cases, γ still satisfies

$$\sup_{u>0} \left| \frac{d}{du} \log \gamma(u) \right| =: \Lambda < \infty.$$

Recall $g = \phi * \gamma$ is the convolution of the heavy-tailed γ given by (7) and the noise density ϕ. Basic properties of g are gathered in Lemma 2, while Lemma 4 provides bounds on corresponding moments of the score function.

2.2. Adaptive risk bounds for integrated posterior

Theorem 1. *Fix $\delta \in (0, 2)$. Let $\gamma = \gamma_\delta$ be the density defined by (7) and let $\hat{\alpha}$ be the corresponding MMLE given by (5) and suppose (6) holds. Then there exists a universal constant $C > 0$, that in particular is independent of δ, q, such that, for large enough n, for any $q \in (2\delta, 2]$,*

$$\sup_{\theta_0 \in \ell_0[s]} E_{\theta_0} \int d_q(\theta, \theta_0)\, d\Pi_{\hat{\alpha}}(\theta \mid X) \leq Cs \log^{q/2}(n/s).$$

Theorem 1 shows that the posterior qth moment converges at the minimax rate for d_q-type-distances over $\ell_0[s]$. By contrast, note that the results in [12] for d_q covered complexity priors on the dimension, but not spike and slab priors (which induce a binomial prior on the dimension), and were results on the posterior convergence as a probability measure and as such did not imply convergence at minimax rate of for example, the posterior mean. The proof of Theorem 1 is given in the supplement, Section A.3.

Let us now briefly comment on the behaviour of some point estimators and on simulations from the empirical Bayes posterior. Under the conditions of Theorem 1, the posterior mean is rate-minimax for any $1 \leq q \leq 2$. This follows from Theorem 1 using the convexity of $\theta \to d_q(\theta, \theta_0)$ if $1 \leq q \leq 2$ and Jensen's inequality. More details on the posterior mean, in particular

its suboptimality when $q < 1$, can be found in the supplement, Section A.8. Concerning the posterior median, one can check that it is rate-minimax for any $0 < q \leq 2$, see Section A.8 in the supplement. We also note in passing that simulation from the considered empirical Bayes posterior distribution is fast: for Cauchy-type slab tails, one can directly use the EbayesThresh package of Johnstone and Silverman, see [20]. To compute $\hat{\alpha}$ corresponding to the precise slab form γ in (7), one can compute approximations of $g(x) = (\gamma * \phi)(x)$ by a numerical integration method and next insert this in the EbayesThresh subroutine computing $\hat{\alpha}$.

Remark 1. One may consider a density γ 'on the boundary' by setting, say,

$$\Delta(u) = u^{-1} \log^{-2} u.$$

For this choice of γ, it can be checked that the risk bound of Theorem 1 holds uniformly for $q \in (0, 2]$. However, this prior density has some somewhat undesired properties for confidence sets: it can be checked that the variance term of the empirical Bayes posterior is, for $q = 2$, of the order $s\tau(\alpha)^2/\log \tau(\alpha)$, which turns out to be sub-optimally small and a blow-up factor of order at least $\log\log(n/s)$ would be needed to guarantee coverage of the corresponding credible set.

2.3. Credible sets for fixed α

For $q \in (0, 2]$, and as before $\hat{\theta}_\alpha$ the posterior coordinate-wise median for fixed α, we set

$$\mathcal{C}_{q,\alpha} = \big\{\theta \in \mathbb{R}^n, d_q(\theta, \hat{\theta}_\alpha) \leq M v_{q,\alpha}(X)\big\}, \tag{8}$$

where M is a constant to be chosen below, and where we denote

$$v_{q,\alpha}(X) = \int d_q(\theta, \hat{\theta}_\alpha)\, d\Pi_\alpha(\theta \mid X).$$

Note that by Markov's inequality, for $M \geq 1/\beta$ it holds

$$\Pi_\alpha[\mathcal{C}_{q,\alpha} \mid X] \geq 1 - \beta,$$

so that $\mathcal{C}_{q,\alpha}$ is a $1-\beta$ credible set (actually it is sufficient to take $M = 1 + \varepsilon$, for arbitrary $\varepsilon > 0$, to achieve $1-\beta$ posterior coverage asymptotically, see Remark 2 below). The proposition below reveals the frequentist properties of the so-constructed credible sets for a fixed value of the tuning parameter α. Taking $\alpha \lesssim s \log^{\delta/2}(n/s)/n$, the *size* of the credible set is (nearly) optimal, reaching the (nearly) minimax rate $s \log^{q/2}(n)$, and by taking $\alpha \asymp s \log^{\delta/2}(n/s)/n$ the exact minimax rate (up to a constant) $s \log^{q/2}(n/s)$ is achieved. On the other hand, the frequentist *coverage* properties of $\mathcal{C}_{q,\alpha}$ behave in an opposite way with respect to α. Indeed, one can find elements of the class $\ell_0[s]$ for which too small choice of the hyperparameter α, i.e. $\alpha = o(s \log^{\delta/2}(n/s)/n)$, results in misleading uncertainty statements. At the same time sufficiently large values of α (i.e., $\alpha \gtrsim s \log^{\delta}(n/s)/n$) provide high frequentist coverage. Let us introduce the set

$$\tilde{\Theta}_{s,\alpha} = \big\{\theta \in \ell_0[s] : \big|\big\{i : t(\alpha)/8 \leq |\theta_{0,i}| \leq t(\alpha)/4\big\}\big| = s\big\}. \tag{9}$$

Note that this sets contains non-zero signals with large enough (but not too large) values.

Proposition 1. *Let $\delta \in (0,2)$ be arbitrary and Π_α be the spike and slab prior with $\gamma = \gamma_\delta$ the density defined by (7). Then for any $q > \delta$ and $s\log^{\delta/2}(n/s)/n \lesssim \alpha \le \alpha_1$ for sufficiently small constant α_1, the Bayes credible set (8) has, with respect to d_q, frequentist coverage tending to one for some sufficiently large choice of M*

$$\inf_{\theta_0 \in \ell_0[s]} P_{\theta_0}(\theta_0 \in \mathcal{C}_{q,\alpha}) \to 1.$$

However, for $\alpha = o(s\log^{\delta/2}(n/s)/n)$ the credible set has frequentist coverage tending to zero for true signals θ_0 in the set $\tilde{\Theta}_{s,\alpha}$ defined in (9), for arbitrary choice of $M > 0$, i.e.

$$\sup_{\theta_0 \in \tilde{\Theta}_{s,\alpha}} P_{\theta_0}(\theta_0 \in \mathcal{C}_{q,\alpha}) \to 0.$$

The next proposition shows that the region of α's where the diameter of the fixed α-credible set is optimal is in a sense 'reversed'. Both results are proved in Section 5.1.

Proposition 2. *Let $\delta \in (0,2)$ be arbitrary and Π_α be the spike and slab prior with $\gamma = \gamma_\delta$ the density defined by (7). Then for any $\delta < q \le 2$ and for any $s\log^{\delta/2}(n/s)/n \ll \alpha \le \alpha_1$ for sufficiently small constant α_1, the Bayes credible set (8) has, with respect to d_q, suboptimal diameter*

$$\inf_{\theta_0 \in \ell_0[s]} E_{\theta_0}\big[\operatorname{diam}_q(\mathcal{C}_{q,\alpha})\big] \gg s\log^{q/2}(n/s).$$

However, if $(s/n)^{c_1} \lesssim \alpha \lesssim (s/n)\log^{\delta/2}(n/s)$ for some $c_1 \ge 1$, the credible set has optimal diameter

$$\sup_{\theta_0 \in \ell_0[s]} E_{\theta_0}\big[\operatorname{diam}_q(\mathcal{C}_{q,\alpha})\big] \lesssim s\log^{q/2}(n/s).$$

Remark 2. One can consider other types of credible sets as well, for instance balls centered around the posterior coordinate-wise median, that is,

$$\tilde{C}_{q,\alpha} = \big\{\theta \in \mathbb{R}^n, d_q(\theta, \hat{\theta}_\alpha) \le r_\beta\big\}, \quad \text{with } r_\beta \text{ taken as } \Pi_\alpha(\tilde{C}_{q,\alpha}|X) = 1-\beta \tag{10}$$

(if the equation has no solution, one takes the smallest r_β such that $\Pi_\alpha(\tilde{C}_{q,\alpha}|X) \ge 1-\beta$).

One can show that these two types of credible sets are the same up to a $(1+o(1))$ blow-up factor for every fixed $0 < \beta < 1$, since

$$r_\beta = \big(1+o(1)\big)v_{q,\alpha}(X), \tag{11}$$

for all $\alpha \in (M_n(\log_2 n)^{\delta/2}/n, \alpha_1)$, with $\alpha_1 > 0$ a small enough constant and $M_n \to \infty$ arbitrarily slowly. The proof of this statement is given in Section A.7 of the supplement.

Therefore, by inflating the credible set (10) by a sufficiently large constant factor L, it has frequentist coverage tending to one for $s\log^{\delta/2}(n/s)/n \lesssim \alpha \le \alpha_1$, that is, for $\tilde{C}_{q,\alpha}(L) = \{\theta \in$

$\mathbb{R}^n, d_q(\theta, \hat{\theta}_\alpha) \leq Lr_\beta\}$, we have

$$\inf_{\theta_0 \in \ell_0[s]} P_{\theta_0}\big(\theta_0 \in \tilde{C}_{q,\alpha}(L)\big) \to 1.$$

However, for $\alpha = o(s \log^{\delta/2}(n/s)/n)$ the credible set has frequentist coverage tending to zero for true signals θ_0 in the set $\tilde{\Theta}_{s,\alpha}$.

2.4. Adaptive credible sets for $q = 2$

In this section, we investigate the adaptive version of the credible set $\mathcal{C}_\alpha$ introduced in (8), in the case $q = 2$. Define the random set, for $M \geq 1$ to be chosen,

$$\mathcal{C}_{\hat{\alpha}} = \mathcal{C}_{2,\hat{\alpha}} = \big\{\theta \in \mathbb{R}^n, \|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 \leq M v_{\hat{\alpha}}(X)\big\}, \tag{12}$$

where $\|\cdot\|^2 = d_2$ is the square of the standard euclidian norm and

$$v_{\hat{\alpha}}(X) = \int \|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 \, d\Pi_{\hat{\alpha}}(\theta \mid X).$$

By Markov's inequality the set $\mathcal{C}_{\hat{\alpha}}$ has at least $1 - \beta$ posterior coverage for $M \geq 1/\beta$. Also, it is a direct consequence of Theorem 1, Markov's inequality and the rate optimality of the posterior median estimator that the size of this sets adapts to the minimax rate: for every $\epsilon > 0$, there exists $M_\epsilon > 0$ such that for any $\theta_0 \in \ell_0[s]$,

$$P_{\theta_0}\big(v_{\hat{\alpha}}(X) \geq M_\epsilon s \log(n/s)\big) \leq \epsilon.$$

So the credible set has an optimal diameter uniformly. However, from similar arguments as in [24], this means that the present credible set cannot have honest coverage for every sparse θ_0, since the construction of adaptive and honest confidence sets for the quadratic risk is impossible in the sparse normal means model, see the Supplement [11] for a precise statement and proof.

To achieve good frequentist coverage one has to introduce certain extra assumptions on the parameter set $\ell_0[s]$. We consider the excessive-bias restriction investigated in the context of the sparse normal means model in [4,31], that is, we say that $\theta_0 \in \ell_0[s]$ satisfies the *excessive-bias restriction* for constants $A > 1$ and $C_2, D_2 > 0$, if there exists an integer $s \geq \ell \geq \log^2 n$, with

$$\sum_{i:|\theta_{0,i}| < A\sqrt{2\log(n/\ell)}} \theta_{0,i}^2 \leq D_2 \ell \log(n/\ell), \qquad \big|\big\{i : |\theta_{0,i}| \geq A\sqrt{2\log(n/\ell)}\big\}\big| \geq \frac{\ell}{C_2}. \tag{13}$$

We denote the set of all such vectors θ_0 by $\Theta_0^2[s] = \Theta_0^2[s; A, C_2, D_2]$, and let $\tilde{s} = \tilde{s}(\theta_0)$ be $|\{i : |\theta_{0,i}| \geq A\sqrt{2\log(n/\ell)}\}|$, for the smallest possible ℓ such that (13) is satisfied. We note that the assumption $\ell \geq \log^2 n$ can be relaxed to $\ell \geq 1$ by considering a modified MLE estimator, as discussed below assumption (6). However for the sake of simplicity and better readability we work under the assumption $\ell \geq \log^2 n$. The necessity of condition (13) is investigated in Section 2.6.

By definition $\tilde{s} \le s$ and possibly, if θ_0 has many small coefficients, one can have $\tilde{s} = o(s)$. We call the quantity $\tilde{s}$ the *effective sparsity* of $\theta_0 \in \ell_0[s]$. It shows the number of large enough signal components which can be distinguished from the noise. The rest of the signals are too small to be detectable, but at the same time their energy (the sum of their squares) is not too large so the bias of standard estimators (which will shrink or truncate the observations corresponding to small signals) won't be dominant. Our goal is to adapt to the present effective sparsity value and at the same time have appropriate frequentist coverage for the credible sets.

Theorem 2. *Let $\delta \in (0,2)$ be arbitrary and let Π_α be the spike and slab prior with $\gamma = \gamma_\delta$ the density defined by (7). Let $\hat{\alpha}$ be the corresponding MMLE given by (5) and suppose that the excessive-bias condition (13) hold. Then the MMLE empirical Bayes credible set (12) has, with respect to the d_2-type distance ($q = 2$), adaptive size to the effective sparsity $\tilde{s}$ and frequentist coverage tending to one, that is, for any $s = o(n)$*

$$\inf_{\theta_0 \in \Theta_0^2[s; A, C_2, D_2]} P_{\theta_0}\big(\operatorname{diam}(\mathcal{C}_{\hat{\alpha}}) \le C\tilde{s}\log(n/\tilde{s})\big) \to 1, \tag{14}$$

$$\inf_{\theta_0 \in \Theta_0^2[s; A, C_2, D_2]} P_{\theta_0}(\theta_0 \in \mathcal{C}_{\hat{\alpha}}) \to 1, \tag{15}$$

for sufficiently large constants $C, M > 0$ (the latter in (12)), depending on A, C_2 and D_2 in the excessive bias condition.

This result is a particular case of Theorem 3 below, whose proof is given in Section 5.2. Similar to the risk results presented in Section 1.4, choosing a Laplace slab would lead to suboptimal diameter of the confidence sets. A heavy tail slab is crucial when following an empirical Bayes method for estimating α in the spike and slab prior.

2.5. Adaptive credible sets: Extension to the case $q < 2$

Let $q \in (0,2]$ and let us start by defining an analogous condition to (13) in order to control the bias of the posterior. We say that $\theta_0 \in \ell_0[s]$ satisfies the *d_q-excessive-bias restriction*, in short EB(q), for constants $A > 1$ and C_q, $D_q > 0$, if there exists an integer $s \ge \ell \ge \log^2 n$ with

$$\sum_{i:|\theta_{0,i}| < A\sqrt{2\log(n/\ell)}} |\theta_{0,i}|^q \le D_q \ell \log^{q/2}(n/\ell), \qquad \big|\big\{i : |\theta_{0,i}| \ge A\sqrt{2\log(n/\ell)}\big\}\big| \ge \frac{\ell}{C_q}. \tag{16}$$

The set of all such vectors θ_0 is denoted $\Theta_0^q[s] = \Theta_0^q[s; A, C_q, D_q]$. For any $q' \in (q, 2]$,

$$\Theta_0^q[s; A, C_q, D_q] \subset \Theta_0^{q'}\big[s; A, C_q, D_q(\sqrt{2}A)^{q'-q}\big]. \tag{17}$$

This means that up to a change in the constants, the EB(q) condition becomes stronger when q decreases. Let $\tilde{s}_q = \tilde{s}_q(\theta_0)$ be $|\{i : |\theta_{0,i}| \ge A\sqrt{2\log(n/\ell)}\}|$, for the smallest possible ℓ such that (16) is satisfied.

Next, we define the random set, for any $q \in (0, 2]$ and $M \geq 1$ to be chosen,

$$\mathcal{C}_{q,\hat{\alpha}} = \big\{\theta \in \mathbb{R}^n, d_q(\theta, \hat{\theta}_{\hat{\alpha}}) \leq M v_{q,\hat{\alpha}}(X)\big\}, \tag{18}$$

where $v_{q,\hat{\alpha}}(X) = \int d_q(\theta, \hat{\theta}_{\hat{\alpha}})\, d\Pi_{\hat{\alpha}}(\theta \mid X)$. By Markov's inequality, this set has at least $1 - \beta$ posterior coverage for $M \geq 1/\beta$. Once again, from Theorem 1 and the optimality of the posterior median estimator in d_q, the size of these sets adapts to the minimax rate: for every $\epsilon > 0$, there exists $M_\epsilon > 0$ such that for any $\theta_0 \in \ell_0[s]$,

$$P_{\theta_0}\big(v_{q,\hat{\alpha}}(X) \geq M_\epsilon s \log^{q/2}(n/s)\big) \leq \epsilon.$$

Theorem 3. *Fix $\delta \in (0, 2)$ and let $q \in (2\delta, 2]$. Let $\gamma = \gamma_\delta$ be the density defined by (7). Let $\hat{\alpha}$ be the corresponding MMLE given by (5) and suppose (16) holds. Then the MMLE empirical Bayes credible set (18) for sufficiently large M ($M > 3c_0(2^q D_q C_q + 1)(2^{q-1} \vee 1)2^6(q - \delta)/\delta$ is sufficiently large, where c_0 is given in Lemma 8), has adaptive size to the effective sparsity $\tilde{s}_q$ and frequentist coverage tending to one, that is, for any $s = o(n)$*

$$\inf_{\theta_0 \in \Theta_0^q[s; A, C_q, D_q]} P_{\theta_0}\big(\operatorname{diam}_q(\mathcal{C}_{q,\hat{\alpha}}) \leq C\tilde{s}_q \log^{q/2}(n/\tilde{s}_q)\big) \to 1, \tag{19}$$

$$\inf_{\theta_0 \in \Theta_0^q[s; A, C_q, D_q]} P_{\theta_0}(\theta_0 \in \mathcal{C}_{q,\hat{\alpha}}) \to 1, \tag{20}$$

for some sufficiently large constant $C > 0$ (depending on A, C_q and D_q)

The proof of this result is given in Section 5.2.

Remark 3. The results of Theorem 3 are uniform over $q \in (2\delta, 2)$ provided M is larger than $\sup_{q \in (2\delta, 2)} 3c_0(2^q D_q C_q + 1)(2^{q-1} \vee 1)2^6(q - \delta)/\delta$ as seen in the proof of Theorem 3.

Remark 4. In Lemma 8, we shall see that, under condition EB(q), the parameters $\tilde{s}$ and $\tilde{s}_q$ are equivalent up to a constant multiplier, hence the result above also holds with the effective sparsity $\tilde{s}$ corresponding to the parametrisation $\Theta_0^2(s; A, C_q, D_q(\sqrt{2}A)^{2-q})$.

Remark 5. We note that the same results as in Theorem 3 hold for the inflated version of the credible set define in (10) as well. The proof is deferred to Section A.7 of the supplementary material.

2.6. Excessive bias conditions: Comparison and minimax necessity

In this section for simplicity, we restrict the discussion to the case $q = 2$. Let us briefly summarise the results obtained in [24], where the authors work in the random design sparse regression model. If s is of smaller order than $\sqrt{n}$, the authors in [24] show, see their Theorem 4 part (A), that construction of adaptive and honest confidence sets is impossible (strictly speaking this result is for the regression model, and for completeness we derive a sequence model version of it in

the Supplement [11]). They also show, see their Theorem 4 part (B), that if one cuts out part of the parameter set, thus obtaining a certain *slicing* formulated in terms of a certain separation (or 'testing') condition, adaptive confidence sets do again exist. If one knows beforehand that one deals with a moderately sparse vector, for which s is of larger order than $\sqrt{n}$, then construction of adaptive confidence sets is possible as well, but requires a different procedure than in the highly sparse case under the testing condition, see, for example, Theorem 1 in [24].

First, we compare the excessive-bias condition with the testing condition introduced in [24] adapted to the sparse sequence model (of course they work on a somewhat different model, but on the same parameter space $\ell_0[s]$). The testing condition was originally given for two sparseness classes $\ell_0[s_1]$ and $\ell_0[s_2]$ for some $s_1 \leq s_2 \wedge n^{1/2}$ and it was shown in Theorems 3 and 4 of [24] that constructing adaptive and honest confidence sets is possible when restricting true signals to the set

$$\ell_0[s_1] \cup \mathcal{T}[s_1, s_2; c] \quad \text{with } \mathcal{T}[s_1, s_2; c] := \big\{\theta \in \ell_0[s_2] : \big\|\theta - \ell_0[s_1]\big\|_*^2 \geq c\big[n^{1/2} \wedge (s_2 \log n)\big]\big\},$$

for some large enough constant $c > 0$, where in the setting of [24] the loss is $\|\cdot\|_*^2 = n \times d_2$, while here we take $\|\cdot\|_*^2 = \|\cdot\|^2 = d_2$.

This condition can be extended to cover every sparsity class up to a certain level s (possibly $s = n$) for instance by introducing the dyadic partition $s_i = 2^i$, $i = 1, 2, \ldots, \lfloor \log_2 s \rfloor$ and formulating the testing condition on every consecutive sparsity class on this grid. A similar type of dyadic partitioning was introduced in [7] in the nonparametric regression and density estimation for Hölder smoothness classes. Set, for given $c > 0$ and $0 \leq s \leq n$,

$$\mathcal{T}_d[s; c] := \bigcup_{i=1}^{\lfloor \log_2 s \rfloor - 1} \mathcal{T}[s_i, s_{i+1}; c].$$

Then one can construct adaptive and honest confidence sets on the set $\mathcal{T}_d[s; c]$ provided c is large enough, see, for instance, the closely related result in context of the nonparametric regression model in [7]. If a vector $\theta \in \mathbb{R}^n$ belongs to $\mathcal{T}_d[s; c]$ for some s, c, we say that it satisfies the *testing condition*. The next lemma, whose proof can be found in Section 6, shows that for well-chosen constants $A, C_2, D_2 > 0$ the excessive-bias condition is a weaker condition than the testing condition for sparsities $s = o(\sqrt{n})$ (up to a log factor).

Lemma 1. *Let $c > 0$ and $s \leq s_{\max} = \lfloor \sqrt{n}/(c \log n) \rfloor$. For $1 \leq s_{\max} \leq n/e$, we have*

$$\mathcal{T}_d[s; c] \subset \Theta_0^2[s; \sqrt{c/2}, 1, c].$$

Further if $c \log(n/s_{max}) > 1$, we have the strict inclusion $\mathcal{T}_d[s; c] \subsetneq \Theta_0^2[s; \sqrt{c/2}, 1, c]$.

Next, we show that the slicing of the parameter space induced by the excessive-bias condition in some sense cannot be weakened to construct adaptive confidence sets even between two sparsity classes. To do so, we proceed in a similar way as for the testing condition in [24], and consider three different types of weakening of the excessive-bias assumption. Let m_n denote a sequence tending to zero arbitrary slowly. First, we relax the upper bound on the energy of the

small signal component from $Cq\log(n/\ell)$ to $m_n^{-1}q\log(n/\ell)$, second we relax the lower bound on the number of signal components above the detection boundary from ℓ/C_2 to $m_n\ell$ and third we relax the detection threshold $A\sqrt{2\log(n/\ell)}$ to $m_n\sqrt{2\log(n/\ell)}$. More formally, the three different relaxations are

$$\sum_{i:|\theta_{0,i}|<A\sqrt{2\log(n/\ell)}} \theta_{0,i}^2 \le m_n^{-1}\ell\log(n/\ell), \qquad \left|\left\{i : |\theta_{0,i}| \ge A\sqrt{2\log(n/\ell)}\right\}\right| \ge \ell/C_2. \tag{21}$$

$$\sum_{i:|\theta_{0,i}|<A\sqrt{2\log(n/\ell)}} \theta_{0,i}^2 \le D_2\ell\log(n/\ell), \qquad \left|\left\{i : |\theta_{0,i}| \ge A\sqrt{2\log(n/\ell)}\right\}\right| \ge m_n\ell, \tag{22}$$

$$\sum_{i:|\theta_{0,i}|<m_n\sqrt{2\log(n/\ell)}} \theta_{0,i}^2 \le D_2\ell\log(n/\ell), \qquad \left|\left\{i : |\theta_{0,i}| \ge m_n\sqrt{2\log(n/\ell)}\right\}\right| \ge \ell/C_2. \tag{23}$$

Theorem 4 below, whose proof is given in Section 6, shows that under neither of these relaxations is it possible to construct adaptive confidence sets.

Theorem 4. *Take any $L > 0$, $s_2 = n^{1/2-\varepsilon}$, for some $\varepsilon > 0$, and $s_1 = m_n s_2$, for some $m_n = o(1)$. Then under neither of the weaker excessive-bias condition (21) or (22) or (23) (each of them denoted by Θ_0 for simplicity) with $A > 1$ and $C_2, D_2 > 0$, exists a confidence set $\mathcal{C}_n(X)$ satisfying simultaneously for $i = 1, 2$ that*

$$\underline{\lim}_n \inf_{\theta_0 \in \Theta_0 \cap \ell_0[s_i]} P_{\theta_0}\big(\theta_0 \in \mathcal{C}_n(X)\big) \ge 1 - \beta, \tag{24}$$

$$\underline{\lim}_n \inf_{\theta_0 \in \Theta_0 \cap \ell_0[s_i]} P_{\theta_0}\big(\operatorname{diam}\big(\mathcal{C}_n(X)\big) \le Ls_i\log(n/s_i)\big) \ge 1 - \beta', \tag{25}$$

for some $\beta, \beta' \in (0, 1/3)$.

The above result shows that, if one slices the parameter space according to an excessive-bias condition, the slicing cannot be refined by making constants arbitrarily smaller: one cannot construct adaptive confidence sets for the resulting larger parameter set. In [24], a similar result is shown for the testing condition above. The slicing induced by the excessive-bias condition is a bit more general as indicated by Lemma 1: it has more flexibility given that it depends also on more parameters. While the impossibility of the weakening (23) can proved appealing to as similar proof as for the testing condition in [24], the other two weakenings correspond to slicing the space in different directions and require a completely new proof.

Theorem 4 can be interpreted as showing the optimality of the slicing within the excessive-bias scale, in the highly sparse regime $s = o(\sqrt{n})$, where adaptive confidence sets do not exist without further assumptions on the space. It could be also interesting to consider the dense regime where one knows beforehand that s is of larger order than $\sqrt{n}$, although it is a qualitatively different question which is not considered here from the optimality perspective. In that case, a different empirical Bayes choice of the sparsity parameter could presumably be used, in a similar spirit as [29] in context of the Gaussian white noise model, enabling the construction of adaptive confidence sets from corresponding posteriors, but this is beyond the scope of the present paper.

3. Discussion

In the paper, we show that the empirical Bayes posterior distribution corresponding to the spike and slab prior, with heavy enough slab tails, results in optimal recovery and reliable uncertainty quantification in ℓ_q-type-norm, $q \in (0, 2]$, under the excessive-bias assumption. We have further shown that the excessive-bias assumption is optimal in a minimax sense for $s = o(\sqrt{n})$ and ℓ_2-norm. A natural extension of the derived results could be to consider hierarchical Bayes methods. Relatively similar results are expected, but computationally simulating from the posterior can be more involved.

We note that the derived contraction and coverage results heavily depend on the choice of the slab prior. The empirical Bayes procedure with Laplace slabs results in sub-optimal contraction rate and therefore too conservative credible sets in ℓ_2-norm, see [9]. Therefore, to achieve optimal recovery of the truth one has to use slab priors with polynomial tails. Considering ℓ_q-type-metrics, $q \in (0, 2]$, one has to carefully choose the order of the polynomial, for instance, for $q \in (0, 1)$ sub-Cauchy tails have to be applied. In view of Propositions 1 and 2, one can see that the optimal choice of mixing hyper parameter α in terms of rates and coverage is $\alpha \asymp (s/n)\log^{\delta/2}(n/s)$, for $\delta < q$. Also note that we have looked at a special excessive-bias type slicing with specific effective sparsity definition; but the 'effective sparsity' in particular could presumably be more general, thus leading to an even more general slicing (but presumably more difficult to study).

4. Preliminaries to the proofs

In Section 4.1, we introduce quantities used repeatedly in the proofs, and state their properties. Notably, the function B appearing in the score function, see (4), is shown to be increasing on $\mathbb{R}^+$, and bounds for its moments are given. In Section 4.2, risk bounds for fixed α are derived, that will be useful both for the rate and confidence sets results. The proofs of these results is given in the Supplement [11], Section A.2.

4.1. General properties and useful thresholds

Lemma 2. *For γ defined by (7), $\delta \in (0, 2)$ and $g = \phi * \gamma$, as $x \to \infty$,*

$$g(x) \asymp \gamma(x),$$

$$g(x)^{-1} \int_x^\infty g(u)\,du \asymp x/\delta.$$

Also, g/ϕ is strictly increasing from $(g/\phi)(0) < 1$ to $+\infty$ as $x \to \infty$.

Threshold $\zeta(\alpha)$. The monotonicity property in Lemma 2 enables one to define a pseudo-threshold from the function $B = (g/\phi) - 1$ as

$$\zeta(\alpha) = B^{-1}\big(\alpha^{-1}\big).$$

Using the bounds from [19] Sections 5.3 and 5.4, noting that these do not use any moment bound on γ (so their bounds also hold even if $\delta < 1$ in the prior density (7)), one can link thresholds $t(\alpha)$ (the threshold of the posterior median for given α, see Section 1.3) and $\zeta(\alpha)$ as follows: $t(\alpha)^2 < \zeta(\alpha)^2$, and $\phi(t(\alpha)) < C\phi(\zeta(\alpha))$, where C is independent of δ, and $B(\zeta(\alpha)) \le 2 + B(t(\alpha))$, arguing as in the proof of Lemma 3 of [19].

Threshold $\tau(\alpha)$. As g/ϕ is continuous, one can define $\tau(\alpha)$ as the solution in x of

$$\Omega(x,\alpha) := \frac{a(x)}{1-a(x)} = \frac{\alpha}{1-\alpha}\frac{g}{\phi}(x) = 1.$$

Equivalently, $a(\tau(\alpha)) = 1/2$. Define α_0 as $\tau(\alpha_0) = 1$ and set

$$\tilde{\tau}(\alpha) = \tau(\alpha \wedge \alpha_0).$$

This is the definition from [19], but note that for α small enough, $\tilde{\tau}(\alpha) = \tau(\alpha)$. Also, it follows from the definition of $\tau(\alpha)$ that $B(\tau(\alpha)) = B(\zeta(\alpha)) - 2 \le B(t(\alpha))$, by the inequality mentioned above, so that $\tau(\alpha) \le t(\alpha)$, as B is increasing. This and the previous inequalities relating the thresholds $\zeta(\alpha)$, $t(\alpha)$, $\tau(\alpha)$ will be freely used in the sequel.

Lemma 3. *For γ defined by (7), $\delta \in (0,2)$, there exists $C_1 > 0$, $C_2 \in \mathbb{R}$ and $a_1 > 0$ such that for any $\alpha \le a_1$,*

$$2\log(1/\alpha) + C_1 \le \zeta(\alpha)^2 \le 2\log(1/\alpha) + (1+\delta)\log\log(1/\alpha) + C_2.$$

The same bounds hold, with possibly different constants C_1, C_2, for $\tau(\alpha)^2$ and $t(\alpha)^2$.

Moments of the score function. Recall that $B(x,\alpha) = B(x)/(1+\alpha B(x))$ and set

$$\tilde{m}(\alpha) = -E_0 B(X,\alpha), \qquad m_1(\mu,\alpha) = E_\mu B(X,\alpha), \qquad m_2(\mu,\alpha) = E_\mu B(X,\alpha)^2.$$

Lemma 4. *The function $\alpha \to \tilde{m}(\alpha)$ is nonnegative and increasing in α. For every fixed $\alpha \in (0,1)$, the function $\mu \mapsto m_1(\mu,\alpha)$ is symmetric and monotone increasing for $\mu \ge 0$. For every fixed $\mu > 0$, the map $\alpha \to m_1(\mu,\alpha)$ is decreasing. As $\alpha \to 0$,*

$$\tilde{m}(\alpha) \asymp \int_\zeta^\infty g(u)\,du \asymp \zeta g(\zeta) \asymp \zeta^{-\delta}/\delta.$$

We have $m_1(\mu,\alpha) \le (\alpha \wedge c)^{-1}$ and $m_2(\mu,\alpha) \le (\alpha \wedge c)^{-2}$ for all μ, and

$$m_1(\mu,\alpha) \le \begin{cases} -\tilde{m}(\alpha) + C\zeta(\alpha)^{2-\delta}\mu^2, & \text{for } |\mu| < 1/\zeta(\alpha), \\ C\dfrac{\phi(\zeta/2)}{\alpha}, & \text{for } |\mu| < \zeta(\alpha)/2, \end{cases}$$

$$m_2(\mu,\alpha) \le \begin{cases} \dfrac{C\delta}{\zeta(\alpha)^2}\dfrac{\tilde{m}(\alpha)}{\alpha}, & \text{for } |\mu| < 1/\zeta(\alpha), \\ \dfrac{C}{\zeta}\dfrac{\phi(\zeta/2)}{\alpha^2}, & \text{for } |\mu| < \zeta(\alpha)/2, \end{cases}$$

for universal constants $c, C > 0$. Finally, as $\alpha \to 0$, one has $m_1(\zeta,\alpha) \sim \frac{1}{2}\alpha^{-1}$, as $\alpha \to 0$.

Bounds on $a(x)$. By definition of $a(x)$, for any real x and $\alpha \in [0,1]$,

$$\alpha \frac{g}{g \vee \phi}(x) \le a(x) \le 1 \wedge \frac{\alpha}{1-\alpha}\frac{g}{\phi}(x). \tag{26}$$

The following bound in terms of $\tau(\alpha)$, see [19] p. 1623, that again extends to the case of a density (7) with $\delta < 1$, is useful for large x: for $\alpha \le \alpha_0$, so that $\tilde{\tau}(\alpha) = \tau(\alpha)$,

$$1 - a(x) \le \mathbb{1}_{|x| \le \tau(\alpha)} + e^{-\frac{1}{2}(|x|-\tau(\alpha))^2}\mathbb{1}_{|x|>\tau(\alpha)}. \tag{27}$$

4.2. Risk bounds for fixed α

Let us consider the posterior d_q-loss for a fixed value of the tuning parameter α, that is

$$\int d_q(\theta,\theta_0)\, d\Pi_\alpha(\theta \mid X) = \sum_{i=1}^n \int |\theta_i - \theta_{0,i}|^q \, d\Pi_\alpha(\theta_i \mid X).$$

To study $\int |\theta_i - \theta_{0,i}|^q \, d\Pi_\alpha(\theta_i \mid X)$, let $\pi_\alpha(\cdot \mid x)$ be the posterior given α evaluated at $X = x$,

$$r_q(\alpha,\mu,x) = \int |u-\mu|^q \, d\pi_\alpha(u \mid x) = \big(1-a(x)\big)|\mu|^q + a(x)\int |u-\mu|^q \gamma_x(u)\, du.$$

Lemma 5. *Denoting $\gamma_x(u) = \gamma(u)\phi(x-u)/g(x)$, for any $q \in (0,2]$ and $\mu, x \in \mathbb{R}$,*

$$\int |u-\mu|^q \gamma_x(u)\, du \le C\big[|x-\mu|^q + 1\big]. \tag{28}$$

For α small enough, for any $q > 2\delta$ and any $\mu, x \in \mathbb{R}$, the following risk bounds hold

$$E_0 r_q(\alpha,0,x) \lesssim \alpha\tau(\alpha)^{q-\delta}/(q-\delta) + \tau(\alpha)^{q-1}\phi\big(\tau(\alpha)\big) + \phi\big(\tau(\alpha)\big)/\tau(\alpha),$$
$$E_\mu r_q(\alpha,\mu,x) \lesssim \big(1+\tau(\alpha)^q\big).$$

5. Proofs for credible sets

The credible set $\mathcal{C}_{q,\alpha}$ in (8) is centered around the posterior median $\hat{\theta}_\alpha$. We shall use below a few basic properties of this estimator, which were established in [19] (they extend without difficulty to the case of heavier tails than Cauchy, as no moments conditions on γ are needed for their proof): the fact that $\hat{\theta}_\alpha$ is a shrinkage rule with the bounded shrinkage property, from Lemma 2 of [19], and the bounds of the risk $E_{\theta_0} d_q(\hat{\theta}_\alpha, \theta_0)$ for fixed α in Lemmas 5 and 6 of [19], recalled in the Supplement [11], see (A.6)–(A.7) there.

5.1. Proof of Propositions 1 and 2

For any given θ_1, θ_2 in the credible set $\mathcal{C}_{q,\alpha}$ from (8), by Lemma 7,

$$d_q(\theta_1,\theta_2) \le \big(2^{q-1}\vee 1\big)\big(d_q(\theta_1,\hat\theta_\alpha)+d_q(\theta_2,\hat\theta_\alpha)\big) \le 2\big(2^{q-1}\vee 1\big)M v_{q,\alpha}(X).$$

Using the risk bounds from Lemma 5 with $\mu=\theta_{0,i}$ for each index i between 1 and n leads to, distinguishing between the signal case ($\theta_{0,i}=0$) and non-zero signal case ($\theta_{0,i}\neq 0$),

$$E_{\theta_0}\int d_q(\theta,\theta_0)\,d\Pi_\alpha(\theta\mid X) \le C\big[(n-s)\alpha\tau(\alpha)^{q-\delta}+s\big(1+\tau(\alpha)^q\big)\big].$$

Also, combining Lemmas 5 and 6 in [19] with fixed non-random threshold equal to $t(\alpha)$, one gets

$$E_{\theta_0} d_q(\hat\theta_\alpha,\theta_0) \le C\big[(n-s)t(\alpha)^{q-1}\phi\big(t(\alpha)\big)+s\big(1+t(\alpha)^q\big)\big].$$

From the last displays one deduces that, for any θ_1,θ_2 in $\mathcal{C}_{q,\alpha}$,

$$E_{\theta_0} d_q(\theta_1,\theta_2) \le CM\big[(n-s)\big\{\alpha\tau(\alpha)^{q-\delta}+t(\alpha)^{q-1}\phi\big(t(\alpha)\big)\big\}+s\big(1+\tau(\alpha)^q+t(\alpha)^q\big)\big].$$

Let us recall the inequality $\alpha^{-1}=B(\zeta(\alpha))\le 2+B(t(\alpha))$ stated in Section 4.1 and $B(\cdot)=(g/\phi)(\cdot)-1$ by (4). As $t(\alpha)\to+\infty$ when $\alpha\to 0$, we have that for small α it holds $1+(g/\phi)(t(\alpha))\lesssim (g/\phi)(t(\alpha))$. Lemma 2 gives that $g(x)$ has tails $x^{-1-\delta}$ as $x\to\infty$, which implies, using again that $t(\alpha)$ goes to infinity as α goes to 0, that for α small enough,

$$\phi\big(t(\alpha)\big)\lesssim \alpha g\big(t(\alpha)\big)\lesssim \alpha t(\alpha)^{-1-\delta}. \tag{29}$$

Using Lemma 3, the above inequality on the diameter becomes the following, hereby proving the second part of Proposition 2: for any θ_1,θ_2 in $\mathcal{C}_{q,\alpha}$ and small enough α,

$$E_{\theta_0} d_q(\theta_1,\theta_2) \le CM\big[n\alpha\log^{(q-\delta)/2}(1/\alpha)+s\log^{q/2}(1/\alpha)\big].$$

Confidence. Next, one considers the coverage probability

$$P_{\theta_0}[\theta_0\in\mathcal{C}_{q,\alpha}] = P_{\theta_0}\big[d_q(\theta_0,\hat\theta_\alpha)\le M v_{q,\alpha}(X)\big].$$

Let $\mu_n=\mu_n(X)=d_q(\theta_0,\hat\theta_\alpha)$ and $S_0:=\{i:\theta_{0,i}\neq 0\}$, then

$$\mu_n=\sum_{i\in S_0}|\theta_{0,i}-\hat\theta_{\alpha,i}|^q+\sum_{i\notin S_0}|\hat\theta_{\alpha,i}|^q\mathbb{1}_{|\varepsilon_i|>t(\alpha)} =: \mu_1+\mu_2.$$

With $v_{q,\alpha}=v_{q,\alpha}(X)$, $\omega_q(x)=\int|u|^q\gamma_x(u)\,du$, and $\kappa_{\alpha,q,i}(x)=\int|u-\hat\theta_{\alpha,i}|^q\gamma_x(u)\,du$,

$$\begin{aligned} v_{q,\alpha} &= \sum_{i=1}^n\int|\theta_i-\hat\theta_{\alpha,i}|^q\,d\Pi(\theta_i\mid X_i)\\ &= \sum_{i:|X_i|\le t(\alpha)}\int|\theta_i|^q\,d\Pi(\theta_i\mid X_i)+\sum_{i:|X_i|>t(\alpha)}\int|\theta_i-\hat\theta_{\alpha,i}|^q\,d\Pi(\theta_i\mid X_i)\end{aligned}$$

$$= \sum_{i\in S_0:|X_i|\le t(\alpha)} a(X_i)\omega_q(X_i) + \sum_{i\in S_0:|X_i|>t(\alpha)} \big[a(X_i)\kappa_{\alpha,q,i}(X_i) + \big(1-a(X_i)\big)|\hat\theta_{\alpha,i}|^q\big]$$
$$+ \sum_{i\notin S_0:|\varepsilon_i|\le t(\alpha)} a(\varepsilon_i)\omega_q(\varepsilon_i) + \sum_{i\notin S_0:|\varepsilon_i|>t(\alpha)} \big[a(\varepsilon_i)\kappa_{\alpha,q,i}(\varepsilon_i) + \big(1-a(\varepsilon_i)\big)|\hat\theta_{\alpha,i}|^q\big]$$
$$=: v_1 + v_2 + v_3 + v_4. \tag{30}$$

Step 1, lower bound on the variance. The variance at fixed α is bounded from below by, using first Lemma 10 to bound ω_q from below, and next using $(1-\alpha)\phi(\varepsilon_i) + \alpha g(\varepsilon_i) \le 2(1-\alpha)\phi(\varepsilon_i)$ when $|\varepsilon_i| \le \tau(\alpha)$ to bound $a(\varepsilon_i)$ from below,

$$v_3 \ge \sum_{i\notin S_0, C_0\le|\varepsilon_i|\le\tau(\alpha)} a(\varepsilon_i)\omega_q(\varepsilon_i) \ge c\alpha \sum_{i\notin S_0} \frac{g}{\phi}(\varepsilon_i)\big(|\varepsilon_i|^q+1\big)\mathbb{1}_{C_0\le|\varepsilon_i|\le\tau(\alpha)}, \tag{31}$$

for any $C_0 > 0$ and $\alpha < 1/2$ (say). One deduces that, in view of the assumption $q > \delta$, for small enough choice of $C_0 > 0$

$$E\bigg[\alpha\sum_{i\notin S_0}\frac{g}{\phi}(\varepsilon_i)\big(|\varepsilon_i|^q+1\big)\mathbb{1}_{C_0\le|\varepsilon_i|\le\tau(\alpha)}\bigg] \ge \frac{(n-s)\alpha}{4}\int_{C_0}^{\tau(\alpha)} g(x)x^q\,dx$$
$$\ge \frac{1}{2^3(q-\delta)} n\alpha\tau(\alpha)^{q-\delta}.$$

Furthermore in view of the inequality $(1+|u|^q)^2 \le Cu^{2q}$ and Lemma 2, we get that

$$\mathrm{Var}\bigg[\alpha\sum_{i\notin S_0}\frac{g}{\phi}(\varepsilon_i)\big(|\varepsilon_i|^q+1\big)\mathbb{1}_{C_0\le|\varepsilon_i|\le\tau(\alpha)}\bigg] \le n\alpha^2\int_{C_0}^{\tau(\alpha)}\frac{g^2}{\phi^2}(u)u^{2q}\phi(u)\,du$$
$$\lesssim n\alpha^2\int_{C_0}^{\tau(\alpha)} u^{-2-2\delta+2q}\phi(u)^{-1}\,du.$$

An integration by parts shows that, setting $d = -2-2\delta+2q$, the right hand side of the preceding display is further bounded from above by a multiple of $n\alpha^2\tau(\alpha)^{d-1}e^{\tau(\alpha)^2/2}$. Using that $\tau(\alpha) \le t(\alpha)$ and $t(\alpha_n)^2 = 2\log n$ (so that $e^{\tau(\alpha)^2/2} \le n$ for $\alpha > \alpha_n$), one gets that the variance term is of smaller order than the square of the expectation, so

$$v_{q,\alpha} \ge \frac{1}{2^4(q-\delta)} n\alpha\tau(\alpha)^{q-\delta}, \tag{32}$$

with high probability. The diameter lower bound in Proposition 2 follows, as $Mv_{q,\alpha} \le \mathrm{diam}_q(\mathcal{C}_{q,\alpha})$, where we have used that $\delta < q$, the lower bound on $\tau(\alpha)$ from Lemma 3, the inequality $n\alpha\tau(\alpha)^{q-\delta} \gtrsim n\alpha(2\log(1/\alpha)+C)^{(q-\delta)/2}$, and the latter is increasing in α.

Step 2, upper bound on the bias. As a first step, we give upper bounds for the terms μ_1 and μ_2. The posterior median is, as stated in [19] Lemma 2, a shrinkage rule with the bounded shrinkage

property: there exists $b > 0$ such that for all $x \geq 0$ and α,

$$\big(x - t(\alpha) - b\big) \vee 0 \leq \hat{\theta}_\alpha(x) \leq x. \tag{33}$$

This implies, as $\hat{\theta}_\alpha(-x) = -\hat{\theta}_\alpha(x)$, that by using Lemma 7

$$\mu_1 \leq \big(2^{q-1} \vee 1\big)\bigg(\sum_{i\in S_0} |\varepsilon_i|^q + \sum_{i\in S_0} \big(t(\alpha)+b\big)^q\bigg) \lesssim \sum_{i\in S_0} |\varepsilon_i|^q + s\big(t(\alpha)+b\big)^q. \tag{34}$$

Let us now use a standard chi-square bound: if Z_i are $\mathcal{N}(0,1)$ i.i.d., for any integer $s \geq 1$ and $t > 0$, one has $P[\sum_{i=1}^s (Z_i^2 - 1) \geq t] \leq \exp\{-t^2/[4(s+t)]\}$. For $t = s(\log(n/s))^{1/2}$, the bound is $\exp\{-s(\log(n/s))^{1/2}\} \leq \exp\{-c\log^{1/2} n\} = o(1)$, so

$$P\bigg[\sum_{i\in S_0} \big(\varepsilon_i^2 - 1\big) > s\sqrt{\log(n/s)}\bigg] = o(1).$$

Also note that by Hölder's inequality, $\sum_{i\in S_0} |\varepsilon_i|^q \leq (\sum_{i\in S_0} \varepsilon_i^2)^{q/2} s^{1-q/2}$. Hence,

$$\sum_{i\in S_0} |\varepsilon_i|^q \lesssim \big(s + s\sqrt{\log(n/s)}\big)^{q/2} s^{1-q/2} \lesssim s\log^{q/4}(n/s), \tag{35}$$

with probability tending to one. For μ_2, using again that the posterior median is a shrinkage rule, $\mu_2 \leq \mu_3 := \sum_{i=1}^n |\varepsilon_i|^q \mathbb{1}_{|\varepsilon_i| > t(\alpha)}$. Then in view of Lemma 12 (with $t_1 = 0$, and $t_2 = t(\alpha) \geq 1$) we have with probability tending to one that

$$\mu_3 \lesssim t(\alpha)^{q-1}\big[ne^{-t(\alpha)^2/2}\big] + M_n t(\alpha)^{(q+1)/2}\big[ne^{-t(\alpha)^2/2}\big]^{1/2}.$$

For the first term to dominate this expression it is enough, provided $M_n \to \infty$ slow enough, that

$$t(\alpha)^2 \leq 2\log n - (3 - q + c)\log\log n = 2\log\big(n/\big(\log^{(3-q+c)/2} n\big)\big), \tag{36}$$

for some $c > 0$, which follows from the assumption $\alpha \gg s/n \gtrsim (\log n)^2/n$ and Lemma 3. Therefore with probability tending to one

$$\mu_1 + \mu_2 \lesssim s\log^{\frac{q}{4}}(n/s) + st(\alpha)^q + t(\alpha)^{q-1}\big[ne^{-t(\alpha)^2/2}\big].$$

Now recall the bound $v_{q,\alpha} \gtrsim n\alpha\tau(\alpha)^{q-\delta}$. We conclude the proof of the first part of Proposition 1 (the positive coverage result) by noting that for $\alpha \geq s\log^{\delta/2}(n/s)/n$, in view of (29) we have $\mu_1 + \mu_2 \lesssim s\log^{q/2}(n/s) + n\alpha t(\alpha)^{q-\delta-2} \lesssim v_{q,\alpha}$. Indeed, for the second term one has $q - \delta - 2 < 0$, so $t(\alpha)^{q-\delta-2} \leq \tau(\alpha)^{q-\delta-2}$ using $t(\alpha) \geq \tau(\alpha)$. For the first term, using the lower bound in Lemma 3, $\tau(\alpha)^2 \gtrsim \log(1/\alpha)$ and next that $\alpha \to \alpha\log(1/\alpha)^{(q-\delta)/2}$ is increasing, so that, using $\alpha \geq s\log^{\delta/2}(n/s)/n$ again, one gets

$$v_{q,\alpha} \gtrsim n\big\{s\log^{\delta/2}(n/s)/n\big\}\tau\big(s\log^{\delta/2}(n/s)/n\big)^{q-\delta} \gtrsim s\log^{q/2}(n/s).$$

Hence for large enough choice of M in (12), one gets frequentist coverage tending to one.

To obtain the second part of Proposition 1 (the non-coverage result), we will use that by assumption $\alpha = o(s\log^{\delta/2}(n/s)/n)$ and show below that for some $C > 0$,

$$\inf_{\theta_0 \in \tilde{\Theta}_{s,\alpha}} P_{\theta_0}\big(\mu_n \geq Cs\log^{q/2}(n/s)\big) \to 1, \tag{37}$$

$$v_{q,\alpha} = o\big(s\log^{q/2}(n/s)\big). \tag{38}$$

Then the result follows by combining the above statements.

Step 3, lower bound on the bias. To show (37) note that for $\theta_0 \in \tilde{\Theta}_{s,\alpha}$, and as $\tau(\alpha) \leq t(\alpha)$,

$$\begin{aligned}
\mu_n &\geq \sum_{i\in S_0} |\hat{\theta}_{\alpha,i} - \theta_{0,i}|^q \\
&\geq \sum_{\tau(\alpha)/8 \leq |\theta_{0,i}| \leq \tau(\alpha)/4} |\theta_{0,i}|^q \mathbb{1}_{\varepsilon_i \in (-(3/4)\tau(\alpha),(3/4)\tau(\alpha))} \\
&\geq 2^{-3q}\tau(\alpha)^q \sum_{i\in S_0} \mathbb{1}_{\varepsilon_i \in (-(3/4)\tau(\alpha),(3/4)\tau(\alpha))}.
\end{aligned}$$

Then as $P(\varepsilon_i \in (-(3/4)\tau(\alpha), (3/4)\tau(\alpha))) \geq 3/4$ if $\alpha \leq \alpha_1$ for some sufficiently small $\alpha_1 > 0$, we get by Hoeffding's inequality that

$$P\left(\sum_{i\in S_0} \mathbb{1}_{\varepsilon_i \in (-(3/4)\tau(\alpha),(3/4)\tau(\alpha))} \leq s/2\right) \lesssim e^{-s/2} = o(1).$$

This implies that $\mu_n \geq 2^{-3q-1}s\tau(\alpha)^q$ with high probability. This is at least of the order $Cs\log^{q/2}(n/s)$ using the assumption on α.

Step 4, upper bound on the variance. Next, we deal with (38) by giving upper bounds for v_1, v_2, v_3 and v_4 in (30) for $\theta_0 \in \tilde{\Theta}_{s,\alpha}$, separately. Let us start with v_1. Recalling that $\tau(\alpha)$ and $t(\alpha)$ differ slightly, let us split the sum defining v_1 over indexes $i \in S_0$ with $|X_i| \leq \tau(\alpha)$ and $\tau(\alpha) < |X_i| \leq t(\alpha)$, respectively. In view of (26) and (28) (with $\mu = 0$), one gets

$$\begin{aligned}
v_1 &\lesssim \alpha \sum_{i\in S_0 : |X_i| \leq \tau(\alpha)} \frac{g}{\phi}(X_i)\big(1 + |X_i|^q\big) + \sum_{i\in S_0 : \tau(\alpha) \leq |X_i| \leq t(\alpha)} \big(1 + |X_i|^q\big) \\
&\lesssim \alpha s e^{\tau(\alpha)^2/4}\tau(\alpha)^{q-\delta} + M_n \alpha s^{1/2} e^{3\tau(\alpha)^2/8}\tau(\alpha)^{q-\delta-1/2} \\
&\quad + s\tau(\alpha)^{q-1}e^{-3^2\tau(\alpha)^2/2^5} + M_n s^{1/2}\tau(\alpha)^{(q+1)/2}e^{-3^2\tau(\alpha)^2/2^6} \\
&= O(s),
\end{aligned}$$

where the second inequality follows from Lemma 11 (with $t = \tau(\alpha)/4$) and Lemma 12 (with $t_2 = \tau(\alpha)$, $t_1 = \tau(\alpha)/4$). Then in view of $|\hat{\theta}_{\alpha,i}|^q \leq |X_i|^q$ and $\kappa_{\alpha,q,i} \lesssim |\hat{\theta}_{\alpha,i}|^q + |X_i|^q + 1 \lesssim$

$|X_i|^q + 1$, see equation (28), we get that

$$\begin{aligned} v_2 &= \sum_{i\in S_0, |X_i|>t(\alpha)} a(X_i)\kappa_{\alpha,q,i} + \big(1-a(X_i)\big)|\hat\theta_{\alpha,i}|^q \\ &\lesssim \sum_{i\in S_0, |X_i|>t(\alpha)} \big(1+|X_i|^q\big) \\ &\lesssim st(\alpha)^{q-1}e^{-3^2t(\alpha)^2/2^5} + M_n s^{1/2}t(\alpha)^{(q+1)/2}e^{-3^2t(\alpha)^2/2^6} = O(s), \end{aligned}$$

where the last line follows from Lemma 12 (with $t_2 = t(\alpha)$, $t_1 = t(\alpha)/4$) and from $e^{-c_1 t(\alpha)^2}t(\alpha)^{c_2} = O(1)$ for $c_1, c_2 > 0$ and $\alpha \lesssim \alpha_1$. Next in view of Lemma 11 (with $t = 0$) and Lemma 12 (with $t_1 = 0$ and $t_2 = \tau(\alpha)$)

$$\begin{aligned} v_3 &\lesssim \alpha \sum_{i\notin S_0:|\varepsilon_i|\le\tau(\alpha)} \frac{g}{\phi}(\varepsilon_i)\big(1+|\varepsilon_i|^q\big) + \sum_{i\notin S_0:\tau(\alpha)\le|\varepsilon_i|\le t(\alpha)} \big(1+|\varepsilon_i|^q\big) \\ &\lesssim \alpha n\tau(\alpha)^{q-\delta} + n\tau(\alpha)^{q-1}e^{-\tau(\alpha)^2/2} + M_n\sqrt{n}\tau(\alpha)^{(q+1)/2}e^{-\tau(\alpha)^2/4} = o\big(s\log^{q/2}(n/s)\big), \end{aligned}$$

where the last line follows from the definition of $\tau(\alpha)$ (implying $e^{-\tau(\alpha)^2/2} \lesssim \alpha g(\tau(\alpha)) \lesssim \alpha\tau(\alpha)^{-1-\delta}$) as well as $n\alpha = o(s\log^{\delta/2}(n/s))$ as assumed. To conclude the proof of Proposition 1, it is enough to note that, using similar arguments,

$$\begin{aligned} v_4 &\lesssim \sum_{i\notin S_0, |\varepsilon_i|>t(\alpha)} \big(1+|\varepsilon_i|^q\big) \\ &\lesssim nt(\alpha)^{q-1}e^{-\frac{t(\alpha)^2}{2}} + M_n n^{\frac12}t(\alpha)^{\frac{q+1}{2}}e^{-t(\alpha)^2/4} \\ &= o\left(s\log^{\frac q2}\left(\frac ns\right)\right). \end{aligned}$$

5.2. Proof of Theorem 3

Step 0, Concentration of $\hat\alpha$ under the excessive-bias condition. A key component is to describe the behaviour of the MMLE $\hat\alpha$ over the set $\Theta_0^q[s; A, C_q, D_q]$. In view of (17) (see also Lemma 8, below) let us consider the larger set $\Theta_0^2[s; A, C_q, D_q(\sqrt{2}A)^{2-q}]$ and denote by $\tilde s = \tilde s(\theta_0)$ the effective sparsity of $\theta_0 \in \Theta_0^2[s; A, C_q, D_q(\sqrt{2}A)^{2-q}]$, i.e.

$$\tilde s = \tilde s(\theta_0) = \big|\big\{i : |\theta_{0,i}| \ge A\sqrt{2\log(n/\ell)}\big\}\big|, \tag{39}$$

where $\ell \ge (\log_2 n)^2$ is the smallest integer satisfying (13) with parameters $A = A$, $C_2 = C_q$ and $D_2 = D_q(\sqrt{2}A)^{2-q}$.

Next, we introduce weights $\tilde{\alpha}_i$, for $i = 1, 2$, as the solution of the respective equations

$$d_i \tilde{\alpha}_i \tilde{m}(\tilde{\alpha}_i) = \tilde{s}/n, \tag{40}$$

where the constants $0 < d_2 < d_1$ are specified later. We note that in view of the fact that $\alpha \to \alpha\tilde{m}(\alpha)$ is increasing, which follows from Lemma 4,

$$\tilde{\alpha}_2/\tilde{\alpha}_1 \le d_1/d_2, \qquad \tilde{\alpha}_2 = o(1). \tag{41}$$

We now show that by appropriate choice of constants $d_1 > d_2$ the corresponding $\tilde{\alpha}_1 < \tilde{\alpha}_2$ are upper and lower bounds, respectively, for $\hat{\alpha}$ under the excessive-bias condition.

Lemma 6. *Under the conditions of Theorem 3, for $\tilde{\alpha}_1$, $\tilde{\alpha}_2$ as in (40), we have*

$$\inf_{\theta_0 \in \Theta_0^q[s; A, C_q, D_q]} P_{\theta_0}(\tilde{\alpha}_1 \le \hat{\alpha} \le \tilde{\alpha}_2) \to 1. \tag{42}$$

The proof of the lemma is deferred to Section A.4 of the Supplement [11].

Step 1, lower bound on the variance. In view of (30)–(31) and Lemma 6, with P_{θ_0}-probability tending to one

$$\begin{aligned} v_{q,\hat{\alpha}} &\ge v_3(\hat{\alpha}) \\ &\ge c\hat{\alpha} \sum_{i \notin S_0} \frac{g}{\phi}(\varepsilon_i)\big(|\varepsilon_i|^q + 1\big)\mathbb{1}_{C_0 \le |\varepsilon_i| \le \tau(\hat{\alpha})} \\ &\ge c\tilde{\alpha}_1 \sum_{i \notin S_0} \frac{g}{\phi}(\varepsilon_i)\big(|\varepsilon_i|^q + 1\big)\mathbb{1}_{C_0 \le |\varepsilon_i| \le \tau(\tilde{\alpha}_2)}. \end{aligned}$$

Then following from the computations above assertion (32), Lemma 4 and the inequality $\tau(\alpha) \le \zeta(\alpha)$ we have, for large enough n,

$$v_{q,\hat{\alpha}} \ge \frac{1}{2^4(q-\delta)} \tilde{\alpha}_1 n \tau(\tilde{\alpha}_2)^{q-\delta} \ge \frac{1}{2^4(q-\delta)} \frac{\tilde{s}\tau(\tilde{\alpha}_2)^{q-\delta}}{d_1 \zeta(\tilde{\alpha}_1)^{-\delta}/\delta} \ge \frac{\delta}{2^4 d_1 (q-\delta)} \tilde{s} \log^{q/2}(n/\tilde{s}).$$

Step 2, upper bound on the bias. Let split the bias term

$$\mu_n(\hat{\alpha}) = \sum_{i=1}^n |\theta_{0,i} - \hat{\theta}_{\hat{\alpha},i}|^q$$

along the index sets $Q_1 = \{i : |\theta_{0,i}| \le 1/t(\tilde{\alpha}_2)\}$, $Q_2 = \{i : 1/t(\tilde{\alpha}_2) \le |\theta_{0,i}| \le t(\tilde{\alpha}_2)/2\}$ and $Q_3 = \{i : |\theta_{0,i}| \ge t(\tilde{\alpha}_2)/2\}$, i.e.

$$\mu_n(\hat{\alpha}) = \sum_{i \in Q_1 \cup Q_2} |\theta_{0,i} - \hat{\theta}_{\hat{\alpha},i}|^q + \sum_{i \in Q_3} |\theta_{0,i} - \hat{\theta}_{\hat{\alpha},i}|^q =: \mu_1(\hat{\alpha}) + \mu_2(\hat{\alpha}).$$

Using that $|\hat{\theta}_{\hat{\alpha},i}|^q \leq \mathbb{1}_{|X_i|>t(\hat{\alpha})}|X_i|^q$, Lemma 7, the monotone decreasing property of the functions $t \mapsto \mathbb{1}_{|X_i|>t}$ and Lemma 6 we have with probability tending to one that

$$\begin{aligned}\mu_1(\hat{\alpha}) &\leq \big(2^{q-1}\vee 1\big)\bigg(\sum_{i\in Q_1\cup Q_2}|\theta_{0,i}|^q + \sum_{i\in Q_1}|X_i|^q\mathbb{1}_{|X_i|>t(\hat{\alpha})} + \sum_{i\in Q_2}|X_i|^q\mathbb{1}_{|X_i|>t(\hat{\alpha})}\bigg)\\ &\leq \big(2^{q-1}\vee 1\big)\bigg(\sum_{i\in Q_1\cup Q_2}|\theta_{0,i}|^q + \sum_{i\in Q_1}|X_i|^q\mathbb{1}_{|X_i|>t(\tilde{\alpha}_2)} + \sum_{i\in Q_2}|X_i|^q\mathbb{1}_{|X_i|>t(\tilde{\alpha}_2)}\bigg). \quad (43)\end{aligned}$$

The first term is smaller than $C_q D_q c_0 \tilde{s}\log^{q/2}(n/\tilde{s})$ following from Lemma 8 and $t(\tilde{\alpha}_2)/2 \sim \{0.5\log(n/\tilde{s})\}^{1/2} < A'\{2\log(n/\tilde{s})\}^{1/2}$, for any $A' > 0$ and large enough n. Furthermore, by applying Lemma 12 (with $t_1 = 1/t(\tilde{\alpha}_2)$, $t_2 = t(\tilde{\alpha}_2)$) and the inequality $|Q_1| < n$ the second term on the right hand side of the preceding display is bounded from above with P_{θ_0}-probability tending to one by a multiple of, for arbitrary $M_n \to \infty$,

$$nt(\tilde{\alpha}_2)^{q-1}e^{-t(\tilde{\alpha}_2)^2/2} + M_n t(\tilde{\alpha}_2)^{(q+1)/2}\big(ne^{-t(\tilde{\alpha}_2)^2/2}\big)^{1/2}.$$

Then by noting that in view of Lemma 9, we have $nt(\tilde{\alpha}_2)e^{-t(\tilde{\alpha}_2)^2/2} \gtrsim \tilde{s} \geq (\log n)^2$,

$$t(\tilde{\alpha}_2)^{(q+1)/2}\big(ne^{-t(\tilde{\alpha}_2)^2/2}\big)^{1/2} = o\big(nt(\tilde{\alpha}_2)^{q-1}e^{-t(\tilde{\alpha}_2)^2/2}\big). \quad (44)$$

Finally, in view of Lemma 12 (with $t_1 = t(\tilde{\alpha}_2)/2$ and $t_2 = t(\tilde{\alpha}_2)$) the third term in the right-hand side of (43) is bounded with probability tending to one by a multiple of

$$|Q_2|t(\tilde{\alpha}_2)^{q-1}e^{-t(\tilde{\alpha}_2)^2/8} + M_n t(\tilde{\alpha}_2)^{(q+1)/2}\big(|Q_2|e^{-t(\tilde{\alpha}_2)^2/8}\big)^2.$$

Then note that in view of Lemma 13 (with $t = 1/t(\tilde{\alpha}_2)$) the cardinality of the set Q_2 is bounded from above by a multiple of $\tilde{s}t(\tilde{\alpha}_2)^q\log^{q/2}(n/\tilde{s})$, hence by using that the function $t \mapsto t^{c_1}e^{-c_2t^2}$ tends to zero as t goes to infinity for arbitrary $c_1 \in \mathbb{R}$ and $c_2 > 0$, the preceding display is of smaller order than $\tilde{s}$. By putting together the obtained bounds, one concludes that by choosing M_n tending to infinite sufficiently slowly, for example, $M_n = o(\log^{q/4}(n/\tilde{s}))$, one gets

$$\mu_1(\hat{\alpha}) \leq \big(2^{q-1}\vee 1\big)\big(Cnt(\tilde{\alpha}_2)^{q-1}e^{-t(\tilde{\alpha}_2)^2/2} + D_q C_q c_0\tilde{s}\log^{q/2}(n/\tilde{s})\big).$$

It remained to deal with $\mu_2(\hat{\alpha})$. In view of assertion (34)

$$\mu_2(\hat{\alpha}) \leq \big(2^{q-1}\vee 1\big)\bigg(\sum_{i\in Q_3}|\varepsilon_i|^q + \sum_{i\in Q_3}\big(t(\hat{\alpha})+b\big)^q\bigg). \quad (45)$$

Lemma 13 with $t = t(\tilde{\alpha}_2)/2$ implies $|Q_3| \leq c_0(2^q D_q C_q + 1)\tilde{s}$, so the second term is bounded with probability tending to one using Lemma 6 and then Lemma 9 by

$$\sum_{i\in Q_3}\big(t(\hat{\alpha})+b\big)^q \leq |Q_3|\big(t(\tilde{\alpha}_1)+b\big)^q \leq 3c_0\big(2^q D_q C_q + 1\big)\tilde{s}\log^{q/2}(n/\tilde{s}),$$

for n large enough. Next, we deal with the first term on the right hand side of (45). In view of (35) we have with probability tending to one that

$$\sum_{i\in Q_3} |\varepsilon_i|^q \lesssim |Q_3| \log^{q/4}\big(n/|Q_3|\big) \le c_0\big(2^q D_q C_q + 1\big)\tilde{s}\log^{q/4}(n/\tilde{s}).$$

Now combining (29) and Lemmas 4 and 9, one sees that

$$nt(\tilde{\alpha}_2)^{q-1}e^{-t(\tilde{\alpha}_2)^2/2} \lesssim n\tilde{\alpha}_2 t(\tilde{\alpha}_2)^{q-2-\delta} \lesssim \zeta(\tilde{\alpha}_2)^{\delta}\tilde{s}t(\tilde{\alpha}_2)^{q-2-\delta} = o(\tilde{s}) \tag{46}$$

so putting the bounds on μ_1, μ_2 together, with probability tending to one,

$$\mu_n(\hat{\alpha}) \le \big(2^{q-1}\vee 1\big)3c_0\big(2^q D_q C_q + 1 + o(1)\big)\tilde{s}\log^{q/2}(n/\tilde{s}).$$

Therefore by choosing $M > 3c_0(2^q D_q C_q + 1)(2^{q-1}\vee 1)2^4 d_1(q-\delta)/\delta$, we get frequentist coverage tending to one. Note that the preceding results hold simultaneously for all $q \in (2\delta, 2)$ if $M > \sup_{q\in(2\delta,2)} 3c_0(2^q D_q C_q + 1)(2^{q-1}\vee 1)2^4 d_1(q-\delta)/\delta$.

To conclude the proof of Theorem 3, it is enough to obtain the diameter bound, which is done using similar arguments as for the upper bound of Proposition 2, once the MMLE α is controlled using Lemma 6 as above. The detailed argument can be found in Section A.5 of the Supplement [11]. This concludes the proof.

5.3. Technical lemmas for credible sets

Lemma 7. *For any reals x, y, and $q > 0$,*

$$|x+y|^q \le \big(2^{q-1}\vee 1\big)\big[|x|^q + |y|^q\big]. \tag{47}$$

Proof. First note that for $q < 1$ the map $x \to |x+y|^q - |x|^q - |y|^q$ is monotone. Furthermore, since the map $x \mapsto |x|^q$ is convex for $q \ge 1$ we have by Jensen's inequality that $|(x+y)/2|^q \le (|x|^q + |y|^q)/2$. □

The proofs of Lemmas 8 up to 13 can be found in Section A.6 of the Supplement [11].

Lemma 8. *Assume that θ_0 satisfies the excessive-bias condition* $\mathrm{EB}(q)$ *in* (16) *for some positive parameters A, C_q, D_q and therefore it belongs also to $\Theta_0^2[s; A, C_q, D_q(\sqrt{2}A)^{2-q}]$. Let us denote by $\tilde{s}_q$ and $\tilde{s}$ the corresponding effective sparsity parameters. Then there exists a large enough $c_0 \ge 1$ such that*

$$\tilde{s} \le \tilde{s}_q \le c_0\tilde{s}.$$

Furthermore, for every $1 < A' < A$, and large enough n,

$$\sum_{|\theta_{0,i}|\le A'\sqrt{2\log(n/\tilde{s})}} |\theta_{0,i}|^q \le C_q D_q c_0 \tilde{s}\log^{q/2}(n/\tilde{s}), \quad \tilde{s} \le \big|\big\{|\theta_{0,i}| \ge A'\sqrt{2\log(n/\tilde{s})}\big\}\big|.$$

Lemma 9 (Basic bounds on $\zeta(\alpha_1)$, $\tau(\alpha_1)$ and $t(\alpha_1)$ and tilde versions). *Let α_1 be defined by* (A.8) *for d a given constant, and let $\zeta_1 = \zeta(\alpha_1)$. Then for some constants C_1, C_2,*

$$2\log(n/s) + C_1 \le \zeta(\alpha_1)^2 \le 2\log(n/s) + \log\bigl(1 + \log(n/s)\bigr) + C_2.$$

The same bounds hold, with possibly different constants C_1 and C_2, for $\tau(\alpha_1)^2$ and $t(\alpha_1)^2$. Furthermore, the same result holds (with possibly different constants $C_1, C_2 > 0$) when α_1 is replaced by $\tilde{\alpha}_1$ or $\tilde{\alpha}_2$ (defined in (40)) *and s by $\tilde{s}$.*

Lemma 10. *There exists $c > 0$ such that for any $q > 0$ and $\mu \in \mathbb{R}$, and any $x \in \mathbb{R}$,*

$$\int |u - \mu|^q \gamma_x(u)\, du \ge c\bigl(1 + |x - \mu|^q\bigr).$$

Lemma 11. *Let $q \in (0, 4]$ and suppose $\delta < q \wedge 2$, with δ as in* (7). *There exists a constant $C > 0$ such that for a set $R_t \subseteq \{i : |\theta_{0,i}| \le t\}$ and $t \ge 0$ arbitrary, we have, with P_{θ_0}-probability tending to one*

$$\sum_{i \in R_t, |X_i| \le \tau(\alpha)} \frac{g}{\phi}(X_i)\bigl(|X_i|^q + 1\bigr) \le \begin{cases} C|R_t|\tau(\alpha)^{q-\delta} & \text{for } t \le 1/\tau(\alpha), \\ C|R_t|e^{\tau(\alpha)^2/4}\tau(\alpha)^{q-\delta}\left[1 + M_n \dfrac{e^{\tau(\alpha)^2/8}}{|R_t|^{1/2}\tau(\alpha)^{1/2}}\right] & \text{for } t \le \tau(\alpha)/4, \end{cases} \tag{48}$$

where M_n is an arbitrary sequence such that $M_n \to \infty$.

Lemma 12. *Let $q \in (0, 4]$. Let $t_1 \ge 0$ and $t_2 \ge \max\{2, 2t_1\}$. For $R_{t_1} \subseteq \{i : |\theta_{0,i}| \le t_1\}$, we have with P_{θ_0}-probability tending to one that*

$$\sum_{i \in R_{t_1}} \bigl(|X_i|^q + 1\bigr)\mathbb{1}_{|X_i| > t_2} \le \tilde{c}_0 |R_{t_1}| t_2^{q-1} e^{-(t_2 - t_1)^2/2} + M_n t_2^{(q+1)/2}\bigl(|R_{t_1}| e^{-(t_2-t_1)^2/2}\bigr)^{1/2},$$

for arbitrary $M_n \to \infty$ and $\tilde{c}_0 = (2^q \vee 2)[4((2-q) \vee 1) + 2]/\sqrt{2\pi}$.

Lemma 13. *Under the excessive-bias restriction* $\mathrm{EB}(q)$ (16) *the size of the set $R_t = \{i : t \le |\theta_{0,i}|\}$ is bounded from above by $c_0(C_q D_q t^{-q} \log^{q/2}(n/\tilde{s}) + 1)\tilde{s}$, with $\tilde{s}$ as in* (39).

6. Proofs for the excessive-bias assumption

Proof of Lemma 1. First, we show that for every $i = 1, \dots, \log_2 s - 1$ and $c > 0$

$$\mathcal{T}[s_i, s_{i+1}; c] \subset \Theta_0^2[s; \sqrt{c/2}, 1, c]. \tag{49}$$

Take any $\theta \in \mathcal{T}[s_i, s_{i+1}; c]$ and note that $\theta^2_{(s_i+1)} \geq c \log n$ (where $\theta_{(j)}$ denotes the jth decreasingly ordered value of the parameter of interest). Let us denote by I the largest index satisfying $\theta^2_{(I)} \geq c \log(n/s_i)$ and note that $I \in \{s_i + 1, \ldots, s_{i+1}\}$. Then one can also see that this I satisfies the condition (13), since $|\{j : \theta_j^2 \geq c \log(n/I)\}| \geq |\{j : \theta_j^2 \geq c \log(n/s_1)\}| = I$ and $\sum_{|\theta_j| \leq \sqrt{c \log(n/I)}} \theta_i^2 \leq (s_2 - I) c \log(n/I) \leq s_1 c \log(n/s_1)$, for $I \leq n/e$, using $s_2 \leq 2s_1$ and $I \geq s_1 + 1$. To show the strict inclusion, let θ_0 be defined as

$$\theta^2_{0,j} = \begin{cases} n, & \text{for } 1 \leq j \leq s_i, \\ 1, & \text{for } s_i + 1 \leq j \leq 2s_i, \\ 0, & \text{for } 2s_i + 1 \leq j, \end{cases}$$

for any $i = 1, \ldots, \log_2 s - 1$. Then $\|\theta_0 - \ell_0(s_i)\|_2^2 \leq \sum_{j=s_i+1}^n \theta^2_{0,j} = s_i < c 2 s_i \log n$ so $\theta_0 \notin \mathcal{T}[s_i, s_{i+1}; c]$ and therefore $\theta_0 \notin \mathcal{T}_d[s; c]$. Furthermore, by choosing $\ell = s_i$ we have that $|\{j : |\theta_{0,j}| \geq \log(n/\ell)\}| \geq \ell$ and $\sum_{|\theta_{0,j}| \leq \sqrt{c \log(n/\ell)}} \theta^2_{0,i} = \ell \leq c \ell \log(n/\ell)$, satisfying the excessive-bias restriction with parameters given in the lemma. □

Proof of Theorem 4. First, consider the result under condition (23). We show below that

$$\mathcal{T}[s_1, s_2; c] \subset \Theta_0^2[s_3; \sqrt{c/2}, 1, 1] \tag{50}$$

for every $c > 0$, and $s_1 < s_2 \leq s_3 = o(\sqrt{n}/\log n)$, satisfying $s_2 < c^{-1} s_1$. This inclusion is similar in spirit to (49), but allows that $s_2 \geq 2s_1$ (for $c < 1/2$), which was required in (49). It is in particular true for $c = m_n^2 = o(1)$ and s_1, s_2 as in the theorem, hence $\mathcal{T}[s_1, s_2; m_n^2] \subset \Theta_0^2[s; m_n/\sqrt{2}, 1, 1]$. Then the proof of the statement simply follows from of Theorem 4(A) of [24] (the authors consider the loss $\|\cdot\|_*^2 = n \times d_2$, but the same argument goes through with the loss $\|\cdot\|_*^2 = d_2$ in the sequence model), where it is shown that it is not possible to construct adaptive confidence sets over the classes $\{\theta \in \ell_0[s_2] : \|\theta - \ell_0[s_1]\|_2^2 \geq m_n^2 s_2 \log n\}$ and $\ell_0[s_1]$ (more precisely in the proof it is shown that it is not possible to construct a confidence set with size bounded by a multiple of $s_2 \log n$ over $\{\theta \in \ell_0[s_2] : \|\theta - \ell_0[s_1]\|_2^2 \geq m_n^2 s_2 \log n\}$ and by $s_1 \log n$ for $\theta = 0$, which completes the proof since the zero signal satisfies the excessive-bias assumption with $\tilde{s} = 0$).

We now prove assertion (50), along the lines of (49). We highlight here only the differences. Note that in view of the proof of Lemma 1, for every $\theta \in \mathcal{T}[s_1, s_2; c]$, we have $\sum_{|\theta_i| \leq \sqrt{c \log(n/I)}} \theta_i^2 \leq (s_2 - I) c \log(n/s_2) \leq I \log(n/I)$ (for $I \leq n/e$), where we use $c s_2 < s_1$ and $I \in \{s_1 + 1, \ldots, s_2\}$.

Next, we deal with the conditions (21) and (22). In these cases, one can not directly apply the results of [24], since elements of the set $\{\theta \in \ell_0[s_2] : \|\theta - \ell_0[s_1]\|_2^2 \geq m_n^2 s_2 \log n\}$ will not necessarily satisfy the excessive-bias assumptions (21) and (22) (consider for instance a signal with $n - s_2$ and s_2 coefficients of size zero and $m_n \sqrt{\log n}$, respectively). The proof of the nonexistence result combines then ideas from [2] and [24], and adapts them to the present setting.

We argue by contradiction. Let us assume that under (21) or (22) one can construct confidence sets satisfying assertions (24) and (25). We show below that this would imply the existence of a

test ψ such that

$$\overline{\lim_n}\Big(E_{\theta_0}\psi + \sup_{\theta\in B_0} E_\theta(1-\psi)\Big) \le \gamma' + 2\gamma, \tag{51}$$

where the signal $\theta_0 = (\theta_{0,1}, \dots, \theta_{0,n})$ and the set B_0 respectively are defined by

$$\theta_{0,i} = \begin{cases} A\sqrt{2\log(n/s_1)}, & \text{for } i \le s_1, \\ 0, & \text{else,} \end{cases}$$

$$B_0 = \big\{\theta \in \ell_0[s_2] : \theta_1 = \cdots = \theta_{s_1} = A\sqrt{2\log(n/s_1)}, \big|\{i : \theta_i^2 = c^2 2\log(n/s_2)\}\big| = s_2 - s_1\big\},$$

with $c^2 < 2\varepsilon/(1+2\varepsilon) < 1$, it that is, the parameters in B_0 have signal strength $A\sqrt{2\log(n/s_1)}$ in the first s_1 coordinates, amongst the rest of the coefficients $s_2 - s_1$ is of squared size $2c^2\log(n/s_2)$, while the rest of the coefficients are zero. Assuming further that $2c^2 < 1$, note that θ_0 and any $\theta \in B_0$ satisfy both of the conditions (21) (with $\ell = s_1$ and $C_2 = 1$) and (22) (with $\ell = s_2$ and $D_2 = 1$). We also show below that for any $\gamma_1 \in (0,1)$, $\varepsilon_1 \in (0, 1-\gamma_1)$,

$$\underline{\lim_n} \inf_{\phi_{\gamma_1}} \sup_{\theta\in B_0} E_\theta(1-\phi_{\gamma_1}) \ge \varepsilon_1, \tag{52}$$

where the infimum is taken over every test of level γ_1. This leads to a contradiction with (51) (noting that $\gamma, \gamma' < 1/3$).

First we verify assertion (51), by constructing a test ψ satisfying

$$\overline{\lim_n} E_{\theta_0}\psi \le \gamma' + \gamma, \tag{53}$$

$$\overline{\lim_n} \sup_{\theta\in B_0} E_\theta(1-\psi) \le \gamma. \tag{54}$$

Let us consider the test $\psi = \mathbb{1}\{\mathcal{C}_n(X) \cap B_0 \ne \varnothing\}$, and using (24),

$$\overline{\lim_n} \sup_{\theta\in B_0} E_\theta(1-\psi) = \overline{\lim_n} \sup_{\theta\in B_0} P_\theta\big(\mathcal{C}_n(X)\cap B_0 = \varnothing\big) \le \overline{\lim_n} \sup_{\theta\in B_0} P_\theta\big(\theta \notin \mathcal{C}_n(X)\big) \le \gamma,$$

while one also has, using the diameter bound (25) and coverage (24),

$$\begin{aligned}\overline{\lim_n} E_{\theta_0}\psi &= \overline{\lim_n} P_{\theta_0}\big(\theta_0 \in \mathcal{C}_n(X), \mathcal{C}_n(X)\cap B_0 \ne \varnothing\big) + \overline{\lim_n} P_{\theta_0}\big(\theta_0 \notin \mathcal{C}_n(X), \mathcal{C}_n(X)\cap B_0 \ne \varnothing\big) \\ &\le \overline{\lim_n} P_{\theta_0}\big(\operatorname{diam}\big(\mathcal{C}_n(X)\big) \ge Ls_1\log(n/s_1)\big) + \overline{\lim_n} P_{\theta_0}\big(\theta_0 \notin \mathcal{C}_n(X)\big) \le \gamma' + \gamma,\end{aligned}$$

where in the first inequality we have used that

$$\inf_{\theta\in B_0} \|\theta_0 - \theta\|_2^2 \ge 2c^2(s_2 - s_1)\log(n/s_2) \gg Ls_1\log(n/s_1).$$

Hence, it remained to verify assertion (52). The minimax risk over B_0 in (52) is bounded from below by any Bayes risk for a prior distribution on B_0. Let us define a specific prior Π on the

set B_0 as follows. Let the first s_1 coordinates be fixed to the value $A\sqrt{2\log(n/s_1)}$, and next let S be sampled uniformly at random over subsets of cardinality $s_2 - s_1$ among the remaining coordinates $\{s_1 + 1, \dots, n\}$. Let $\{\epsilon_j\}$ be i.i.d. Rademacher and, given S, set

$$\theta_j = c\sqrt{2\log(n/s_2)}\epsilon_j \quad \text{if } j \in S,$$

and $\theta_k = 0$ otherwise. For convenience let us introduce the notation $\lambda = c\sqrt{2\log(n/s_2)}$. The corresponding marginal likelihood ratio $L_\Pi(Y) = \int (dP_\theta/dP_{\theta_0})(Y)\, d\Pi(\theta)$ is

$$L_\Pi(Y) = \frac{1}{\binom{n-s_1}{s_2-s_1}} \sum_{S \in \mathcal{S}(s_2-s_1, n-s_1)} E_{\epsilon|S}\left[\exp\left(-(s_2-s_1)\lambda^2/2 + \lambda \sum_{j\in S} \epsilon_j Y_j\right)\right],$$

where $\mathcal{S}(s_2-s_1, n-s_1)$ denotes the subsets of size s_2-s_1 of a set of size $n-s_1$ of $\{s_1+1, \dots, n\}$, and $E_{\epsilon|S}$ denotes the expected value corresponding to the i.i.d. Rademacher random variables $\epsilon = \{\epsilon_j : j \in S\}$. Let us introduce the notation $K(Y_{s_1+1}) = E_{\epsilon|S}[\exp(-\lambda^2/2 + \lambda\epsilon_j Y_j)]$, for some $j \in S$ and

$$a = E_{\theta_0}\big[K(Y_{s_1+1})^2\big], \qquad b = E_{\theta_0}\big[K(Y_{s_1+1})\big].$$

Note that a, b do not depend on θ_0, since $\theta_{0,s_1+1} = 0$. Then

$$\begin{aligned}
E_{\theta_0}\big[L_\Pi(Y)^2\big] &= \frac{1}{\binom{n-s_1}{s_2-s_1}^2} \sum_{S,S' \in \mathcal{S}(s_2-s_1, n-s_1)} E_{\theta_0}\Bigg[\prod_{j\in S} E_{\epsilon|S} \exp\big(-\lambda^2/2 + \lambda\epsilon_j Y_j\big) \\
&\qquad \times \prod_{j\in S'} E_{\epsilon|S'} \exp\big(-\lambda^2/2 + \lambda\epsilon_j Y_j\big)\Bigg] \\
&= \frac{1}{\binom{n-s_1}{s_2-s_1}^2} \sum_{S,S' \in \mathcal{S}(s_2-s_1, n-s_1)} E_{\theta_0}\Bigg[\prod_{j\in S\cap S'} E_{\epsilon|S\cap S'} \exp\big(-\lambda^2 + 2\lambda\epsilon_j Y_j\big) \\
&\qquad \times \prod_{j\in S\Delta S'} E_{\epsilon|S\Delta S'} \exp\big(-\lambda^2/2 + \lambda\epsilon_j Y_j\big)\Bigg] \\
&= \frac{1}{\binom{n-s_1}{s_2-s_1}^2} \sum_{S,S' \in \mathcal{S}(s_2-s_1, n-s_1)} a^{|S\cap S'|} b^{|S\Delta S'|} \\
&= b^{2(s_2-s_1)} \sum_{j=1}^{s_2-s_1} \big(a/b^2\big)^j p_{j,s_2-s_1,n-s_1},
\end{aligned}$$

where $A\Delta B = \{A\backslash B\} \cup \{B\backslash A\}$ and $p_{j,s_2-s_1,n-s_1} = \binom{n-s_1}{s_2-s_1}^{-2} |\{(S,S') \in \mathcal{S}(s_2-s_1, n-s_1)^2 : |S\cap S'| = j\}|$. One can notice that the random variable X satisfying $P(X = j) = p_{j,s_2-s_1,n-s_1}$ has a hypergeometric distribution with parameters $n-s_1$, s_2-s_1 and $(s_2-s_1)/(n-s_1)$ and the right-hand side of the preceding display can be written as $b^{2(s_2-s_1)} E[(a/b^2)^X]$. Then in view of

(27) of [2] (with $\cosh(\lambda^2)$ replaced by a/b^2 and k by $s_2 - s_1$), one obtains

$$E_{\theta_0}\big[L_\Pi(Y)^2\big] \le b^{2(s_2-s_1)} \exp\bigg((s_2 - s_1) \log\bigg[1 + \frac{s_2 - s_1}{n - s_1}\bigg(\frac{a}{b^2} - 1\bigg)\bigg]\bigg)$$
$$\le b^{2(s_2-s_1)} \exp\bigg(\frac{(s_2 - s_1)^2}{n - s_1}\bigg(\frac{a}{b^2} - 1\bigg)\bigg). \tag{55}$$

Note that in view of $E_{\theta_0}\cosh(\lambda Y_{s_1+1}) = e^{\lambda^2/2}$ for any $\lambda \in \mathbb{R}$ we have

$$b = e^{-\lambda^2/2} E_{\theta_0}\cosh(\lambda Y_{s_1+1}) = 1.$$

Furthermore, in view of $E_{\theta_0}\cosh^2(\lambda Y_{s_1+1}) = e^{\lambda^2}\cosh(\lambda^2)$

$$a = E_{\theta_0}\big(e^{-\lambda^2/2}\cosh(\lambda Y_{s_1+1})\big)^2 = \cosh\big(\lambda^2\big) = \big(1/2 + o(1)\big)(n/s_2)^{2c^2}.$$

By noting that $s_1 = o(s_2)$ and $s_2 \le n^{1/2-\varepsilon}$ and substituting the preceding two displays into (55) one gets, using $(s_2 - s_1)/(n - s_1) \le s_2/n$ and $1 + o(1) \le 2$ for large enough n,

$$E_{\theta_0}\big[L_\Pi(Y)^2\big] \le \exp\bigg(\frac{s_2^2}{n}\bigg(\frac{n^{2c^2}}{2s_2^{2c^2}} - 1\bigg)\big(1 + o(1)\big)\bigg)$$
$$\le 1 + s_2^{2-2c^2} n^{2c^2-1}$$
$$\le 1 + n^{c^2(1+2\varepsilon)-2\varepsilon} = 1 + o(1). \tag{56}$$

Finally, in view of (24) of [2] (and the display below (24) in [2]), we obtain that

$$\inf_{\phi_{\gamma_1}} \sup_{\theta \in B_0} E_\theta(1 - \phi_{\gamma_1}) \ge \inf_{\phi_{\gamma_1}} \int_{B_0} E_\theta(1 - \phi_{\gamma_1})\Pi(d\theta)$$
$$\ge 1 - \gamma_1 - \bigg\| \int_{B_0} P_\theta \Pi(d\theta) - P_{\theta_0} \bigg\|_{\mathrm{TV}} \Big/ 2$$
$$\ge 1 - \gamma_1 - \big(E_{\theta_0}\big[L_\Pi(Y)^2\big] - 1\big)^{1/2}/2$$
$$\ge 1 - \gamma_1 - o(1),$$

where the last display follows from assertion (56), concluding the proof of assertion (52) and therefore the proof of the theorem. □

Acknowledgements

The authors would like to thank Richard Nickl for insightful comments.

Work of Ismaël Castillo was partly supported by the grant ANR-17-CE40-0001-01 of the French National Research Agency ANR (project BASICS).

Botond Szabó was supported by the Netherlands Organization for Scientific Research. The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

Supplementary Material

Supplement to "Spike and slab empirical Bayes sparse credible sets" (DOI: 10.3150/19-BEJ1119SUPP; .pdf). A separate supplement collects the remaining proofs.

References

[1] Abramovich, F., Benjamini, Y., Donoho, D.L. and Johnstone, I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584–653. MR2281879 https://doi.org/10.1214/009053606000000074

[2] Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606. MR1935648

[3] Belitser, E. and Ghosal, S. (2019). Empirical Bayes oracle uncertainty quantification for regression. *Ann. Stasist.* To appear.

[4] Belitser, E. and Nurushev, N. (2019). Needles and straw in a haystack: Robust empirical Bayes confidence for possibly sparse sequences. *Bernoull*. To appear.

[5] Bhattacharya, A., Pati, D., Pillai, N.S. and Dunson, D.B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *J. Amer. Statist. Assoc.* **110** 1479–1490. MR3449048 https://doi.org/10.1080/01621459.2014.960967

[6] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* (*JEMS*) **3** 203–268. MR1848946 https://doi.org/10.1007/s100970100031

[7] 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

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

[9] 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

[10] 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

[11] Castillo, I. and Szabo, B. (2020). Supplement to "Spike and slab empirical Bayes sparse credible sets." https://doi.org/10.3150/19-BEJ1119SUPP.

[12] Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077 https://doi.org/10.1214/12-AOS1029

[13] Donoho, D.L., Johnstone, I.M., Hoch, J.C. and Stern, A.S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* **54** 41–81. With discussion and a reply by the authors. MR1157714

[14] George, E.I. and Foster, D.P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747. MR1813972 https://doi.org/10.1093/biomet/87.4.731

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

[16] Golubev, G.K. (2002). Reconstruction of sparse vectors in white Gaussian noise. *Problemy Peredachi Informatsii* **38** 75–91. MR2101314 https://doi.org/10.1023/A:1020098307781
[17] Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467 https://doi.org/10.1214/08-AOS638
[18] Johnson, V.E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107** 649–660. MR2980074 https://doi.org/10.1080/01621459.2012.682536
[19] Johnstone, I.M. and Silverman, B.W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135 https://doi.org/10.1214/009053604000000030
[20] Johnstone, I.M. and Silverman, B.W. (2005). EbayesThresh: R programs for Empirical Bayes thresholding. *J. Stat. Softw.* **12** .
[21] Johnstone, I.M. and Silverman, B.W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33** 1700–1752. MR2166560 https://doi.org/10.1214/009053605000000345
[22] Low, M.G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. MR1604412 https://doi.org/10.1214/aos/1030741084
[23] Martin, R. and Walker, S.G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Stat.* **8** 2188–2206. MR3273623 https://doi.org/10.1214/14-EJS949
[24] 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
[25] Robins, J. and van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* **34** 229–253. MR2275241 https://doi.org/10.1214/009053605000000877
[26] Ročková, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Ann. Statist.* **46** 401–437. MR3766957 https://doi.org/10.1214/17-AOS1554
[27] Ročková, V. and George, E.I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.* **113** 431–444. MR3803476 https://doi.org/10.1080/01621459.2016.1260469
[28] Rousseau, J. and Szabó, B. (2019). Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors in general settings. *Ann. Statist.* To appear.
[29] Szabó, B., van der Vaart, A. and van Zanten, H. (2015). Honest Bayesian confidence sets for the L^2-norm. *J. Statist. Plann. Inference* **166** 36–51. MR3390132 https://doi.org/10.1016/j.jspi.2014.06.005
[30] 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
[31] van der Pas, S., Szabó, B. and van der Vaart, A. (2017). Adaptive posterior contraction rates for the horseshoe. *Electron. J. Stat.* **11** 3196–3225. MR3705450 https://doi.org/10.1214/17-EJS1316
[32] 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
[33] van der Pas, S.L., Kleijn, B.J.K. and van der Vaart, A.W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electron. J. Stat.* **8** 2585–2618. MR3285877 https://doi.org/10.1214/14-EJS962
[34] van der Pas, S.L., Salomond, J.-B. and Schmidt-Hieber, J. (2016). Conditions for posterior contraction in the sparse normal means problem. *Electron. J. Stat.* **10** 976–1000. MR3486423 https://doi.org/10.1214/16-EJS1130

Received August 2018 and revised January 2019