The Annals of Statistics
2020, Vol. 48, No. 6, 3113–3137
https://doi.org/10.1214/19-AOS1845

EMPIRICAL BAYES ORACLE UNCERTAINTY QUANTIFICATION FOR REGRESSION

BY EDUARD BELITSER[1] AND SUBHASHIS GHOSAL[2]

[1]*Department of Mathematics, Vrije Universiteit Amsterdam, e.n.belitser@vu.nl*
[2]*Department of Statistics, North Carolina State University, sghosal@stat.ncsu.edu*

We propose an empirical Bayes method for high-dimensional linear regression models. Following an oracle approach that quantifies the error locally for each possible value of the parameter, we show that an empirical Bayes posterior contracts at the optimal rate at all parameters and leads to uniform size-optimal credible balls with guaranteed coverage under an "excessive bias restriction" condition. This condition gives rise to a new slicing of the entire space that is suitable for ensuring uniformity in uncertainty quantification. The obtained results immediately lead to optimal contraction and coverage properties for many conceivable classes simultaneously. The results are also extended to high-dimensional additive nonparametric regression models.

1. Introduction. A linear regression model with a large number of predictors is commonly adopted in modern statistics. Numerous methods exploiting a possible low-dimensional sparse structure have been proposed and their optimality properties for estimation and variable selection have been studied, the most notable of which is the Lasso [Tibshirani [32]]. Various oracle inequalities established for Lasso-type procedures imply optimal estimation accuracy adapted to the sparsity level and accuracy of model selection. However, confidence regions are rarely studied in the context of sparse regression. The issue of coverage with an adaptive optimal size is far more delicate than estimation accuracy. For instance, in the well-studied, infinitely many normal means model $X_i \overset{\text{ind}}{\sim} \mathrm{N}(\theta_i, n^{-1})$, $i = 1, 2, \ldots$, $\sum_{i=1}^{\infty} \theta_i^2 < \infty$, it is impossible for any confidence set to have coverage at all parameter values maintaining the optimal diameter adaptively over different classes [Li [20], Baraud [1], Cai and Low [8]]. This happens since under some true parameter values in each class, any optimal smoothing procedure can be tricked to believe that the true parameter is smoother than actually is, and hence it underestimates the bias, leading to the lack of adequate coverage. Only sufficiently conservative procedures can achieve coverage for such parameter values. Forcing coverage all over the parameter space makes the expected diameter of the confidence set to be at least of the order $n^{-1/4}$ for all values of the true parameter, and will be as bad as a constant for some parameter values. A sensible compromise is to remove a set of "deceptive parameters" responsible for the poor precision and obtain coverage of the confidence region with optimal size adaptively over the remaining part of the parameter space. As long as the set of deceptive parameters is small in an appropriate sense, a more precise procedure giving coverage at all nondeceptive parameters is preferable. Some conditions on the true parameter with increasing generality ensuring coverage of some confidence set in the context of infinitely many normal means are given by self-similarity [Picard and Tribouley [25]], polished tail [Szabó et al. [31]] and excessive bias restriction [Belitser [2]]. To the best of our knowledge, the only articles dealing with the construction of confidence regions in high-dimensional linear regression models are Nickl and van de Geer [24] and Cai and

Received October 2017; revised March 2019.
MSC2020 subject classifications. Primary 62G15, 62C05, 62H35; secondary 62G99.
Key words and phrases. Credible ball, coverage, empirical Bayes, excessive bias restriction, oracle rate.

Guo [7]. The former assumed all predictors are generated independently from the standard normal distribution and studied the size of a confidence set for the vector of linear regression coefficients under two different regimes of sparsity at a time. The latter treated only linear functionals of the parameter.

Bayesian procedures have a natural method for quantifying uncertainty through credible regions, which are typically easy to obtain using modern computing techniques. However, in general, a Bayesian credible set need not have frequentist coverage at the desired level, even approximately, for infinite-dimensional models; see Cox [13] and Freedman [16]. The problem arises essentially because in nonparametric problems, under optimal smoothing, the order of the bias matches the order of the posterior variation, and hence the coverage may be arbitrarily low. In recent years, the problem of constructing Bayesian credible regions with guaranteed coverage for nonparametric problems received considerable attention. The infinitely many normal means model, which is equivalent with the white noise model, is the most well-studied model in the nonparametric setting ostensibly because of the availability of explicit expressions. For the white noise model, Knapik et al. [19] obtained sufficient coverage by undersmoothing, while Castillo and Nickl [9] constructed credible sets with frequentist coverage through Bernstein–von Mises theorems in negative Sobolev spaces using wavelet-based priors. Ray [28] pursued the approach of Castillo and Nickl [9] in the adaptive setting and obtained coverage of credible balls for self-similar sequences with adaptive size. An approach based on inflating a credible region by a sufficiently large constant was adopted in Szabo et al. [31], Belitser [2] and Belitser and Nurushev [4]. In the latter two papers, an oracle approach was used to quantify accuracy locally for each possible value of the true parameter. The main advantage of quantifying error locally for each parameter value is that different scales may be used to quantify regularity. For example, a Bayes procedure is shown to attain automatically the best rate for the size of the credible region over different levels of regularity of the true parameter and simultaneously over many possible regularity scales (such as ellipsoids or hypercubes). Outside the normal mean models, only a few results on coverage of nonparametric credible sets are available. Castillo and Nickl [10] extended their earlier work to density estimation problems. Yoo and Ghosal [35] used inflated credible regions to obtain coverage for multivariate nonparametric regression for pointwise, L_2- and L_∞-credible sets using priors based on B-spline basis expansion. Sniekers and van der Vaart [29] used scaled Brownian motion to construct credible sets with coverage properties for nonparametric regression while Sniekers and van der Vaart [30] constructed inflated adaptive L_2-credible regions using a rescaled Gaussian process prior and showed asymptotic coverage for an analog of a polished tail condition. Coverage properties of posterior credible sets in a high-dimensional linear regression model have not been studied thus far.

In this paper, we adopt the oracle approach of Belitser [2] and Belitser and Nurushev [4] to quantify the estimation accuracy, coverage and precision of an empirical Bayesian procedure for linear regression with a large number of predictors. Suppose that we observe

$$Y = \mathrm{X}\theta + \sigma\varepsilon, \tag{1.1}$$

where $\mathrm{X} = (X_1, \dots, X_p)$ is a deterministic $n \times p$ design matrix, $\theta = (\theta_1, \dots, \theta_p) \in \mathbb{R}^p$ is an unknown regression parameter where possibly p can be much larger than n, and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ is the (column) vector of independent normally distributed random errors with zero mean and unit standard deviation. We do not impose any a priori restriction on the design matrix. If the predictors arise randomly from certain distributions, the results will have to be applied conditionally on their realizations. As in comparable works in the literature, the noise intensity $\sigma > 0$ will be considered known in our setting. For a predictor of the form $\hat{Y} = \mathrm{X}\hat{\theta}$ for some estimator $\hat{\theta}$ of θ, the prediction risk is defined by $R^2(I, \theta) = \mathrm{E}\|\mathrm{X}\hat{\theta}(I) - \mathrm{X}\theta\|^2$, where $\hat{\theta}(I)$ is $\hat{\theta}$ computed on the basis of the model indexed by $I \subseteq \{1, \dots, p\}$, that is,

assuming $\theta_{I^c} = (\theta_j : j \in I^c) = 0$, or equivalently, when the predictors $(X_j : j \in I^c)$ are made irrelevant. The prediction risk optimized over some natural family of predictors serves as a natural quantification of the estimation accuracy and the precision of a confidence region for $\mathrm{X}\theta$ at any given value of θ, and will be referred to as the (squared) *oracle rate*. For a good choice of the predictors family, the oracle rate is dominated by the scale of minimax rates over traditional sparsity classes. We shall consider an empirical Bayes approach for the problem by putting a conjugate normal prior on θ and selecting the mean of the prior distribution by the empirical Bayes approach. In the Supplementary Material [3], we describe a computing strategy based on the simulated annealing technique. We show that the empirical Bayes posterior concentrates around the true value nearly at the oracle rate, thus leading to adaptive minimax optimality of the empirical Bayes posterior mean as a point estimator over sparsity classes. We shall also construct a posterior credible ball centered at $\hat{Y}$ with size nearly equaling a multiple of the oracle rate provided that certain deceptive parameters are excluded from consideration. Clearly, it is desirable to exclude as little of the parameter space as possible. It will be seen that a condition, to be called the *excessive bias restriction*, arises naturally in controlling the bias in terms of the oracle rate. The condition is milder than other conditions such as self-similarity or polishedness of the tail used in the literature for analogous purposes. In effect, the excessive bias restriction condition gives rise to a new sparsity scale which slices the entire space and is very suitable for ensuring uniformity in uncertainty quantification. The structural parameter of the EBR-scale measures the extent of deceptiveness, which determines the needed amount of inflation of the confidence ball for high coverage.

The results obtained for linear regression models are further extended to additive nonparametric regression models where we show that the posterior distribution adapts to both sparsity of the model and the smoothness of the component functions and the resulting credible regions have coverage when a suitable excessive bias restriction condition holds.

The paper is organized as follows. The setup is formulated and the main results are presented in Section 2. Some extensions are given in Section 3, along with a thorough discussion of the key condition of excessive bias restriction. Extensions of the results to additive nonparametric regression models are given in Section 4. Proofs are presented in Section 5. Empirical Bayes interpretation of the posterior distribution, its computational scheme and some proofs are provided in the Supplementary Material [3].

2. Setup and main results. Consider the linear regression model (1.1) with $\varepsilon \sim \mathrm{N}_n(0, \mathrm{I})$, the n-variate normal distribution with mean vector 0 and dispersion matrix I, the identity matrix. We let X be fixed in our setup. Various objects defined in course may depend on X.

2.1. *Notation and conventions.* We use the following notation and conventions throughout the paper. We denote the probability distribution of Y from the model (1.1) by P_θ and the corresponding expectation is denoted by E_θ. Let $|S|$ denote the cardinality of a set S, $v_I = (v_i, i \in I) \in \mathbb{R}^{|I|}$ denote a subvector of $v \in \mathbb{R}^p$, and $\|v\|$ denote the Euclidean norm of v. We interpret vectors as column vectors and denote matrices by upright capital letters. For positive integers k, l, let I_k and $\mathrm{O}_{k,l}$ denote the $k \times k$ identity matrix and the $k \times l$ zero matrix, respectively, and let 0_k stand for the zero vector of length k, with the index(es) omitted if clear from the context. For two square matrices A and B, we say that $\mathrm{A} \le \mathrm{B}$ (resp., $\mathrm{A} < \mathrm{B}$) if $\mathrm{B} - \mathrm{A}$ is nonnegative definite (resp., positive definite). The symbol $\triangleq$ will refer to equality by definition. Let $\mathbb{1}$ stand for the indicator function.

Let $\mathcal{I} = \{\varnothing \ne I \subseteq \{1, \dots, p\} : \text{the columns } (X_i, i \in I) \text{ are linearly independent}\}$, X_I be the submatrix of X with columns $(X_i, i \in I)$. Clearly, $|\mathcal{I}| \le 2^p$ and $|I| \le \nu$ for any $I \in \mathcal{I}$,

where $\nu = \operatorname{rank}(\mathrm{X}) \le \min(n, p)$. If $I \in \mathcal{I}$ were known to be the support of θ, then a natural (minimax) estimator will set $\theta_{I^c} = 0_{|I^c|}$ and estimate the remaining components by the least square estimator. This motivates considering the family of estimators

$$\{\hat\theta(I), I \in \mathcal{I} : \hat\theta_I(I) = (\mathrm{X}_I'\mathrm{X}_I)^{-1}\mathrm{X}_I'Y, \hat\theta_{I^c}(I) = 0_{|I^c|}\}, \tag{2.1}$$

and introducing the oracle risk relative to this family. Let $\mathrm{H}_I = \mathrm{X}_I(\mathrm{X}_I'\mathrm{X}_I)^{-1}\mathrm{X}_I'$ be the orthogonal projection matrix onto the column space $\operatorname{col}(\mathrm{X}_I)$ of X_I. Since $\mathrm{X}\theta = \mathrm{X}_I\theta_I + \mathrm{X}_{I^c}\theta_{I^c}$ and $\mathrm{H}_I(\mathrm{I} - \mathrm{H}_I) = \mathrm{O}$, the quadratic prediction risk of $\mathrm{X}\hat\theta(I) = \mathrm{X}_I\hat\theta_I(I) = \mathrm{H}_I Y$ is

$$R^2(I, \theta) = \mathrm{E}_\theta \|\mathrm{X}\hat\theta(I) - \mathrm{X}\theta\|^2 = \|(\mathrm{I} - \mathrm{H}_I)\mathrm{X}\theta\|^2 + \sigma^2|I|. \tag{2.2}$$

This follows from the standard fact that if V is a random vector with mean μ and covariance matrix Σ and A is a symmetric matrix of constants, then $\mathrm{E}(V'\mathrm{A}V) = \operatorname{tr}(\mathrm{A}\Sigma) + \mu'\mathrm{A}\mu$. For a given θ, the minimizer $I_o^R = I_o^R(\theta)$ of the risk in (2.2) with respect to $I \in \mathcal{I}$ is called the *R-oracle*. The corresponding minimum $R(\theta) = R(I_o^R, \theta)$ quantifies the level of accuracy of inference about $\mathrm{X}\theta$ if we choose the best possible index set I. Note that this "local rate" does not rely on any particular scale of regularity. The goal would be to construct an (empirical) Bayesian procedure which performs, under any true θ, uniformly like the oracle procedure $\hat\theta_{I_o^R}$ without knowing I_o^R, that is, will *mimic* the R-oracle. However, even in the case $\mathrm{X} = \mathrm{I}$, Donoho and Johnstone [15] and Birgé and Massart [5] showed that it is impossible to mimic the R-oracle and a logarithmic factor is the unavoidable price for the uniformity over the whole of $\mathbb{R}^p$ (in fact, even over the scale of sparsity classes, cf. Birgé and Massart [5]). Therefore, only a modification of the risk R where the variance term $\sigma^2|I|$ is inflated with an appropriate logarithmic factor is "mimicable." For $\tau > 0$, define a *τ-oracle* $I_o^\tau = I_o^\tau(\theta)$ at θ as a minimizer (with respect to $I \in \mathcal{I}$) of

$$r_\tau^2(I, \theta) = \|(\mathrm{I} - \mathrm{H}_I)\mathrm{X}\theta\|^2 + \tau\sigma^2|I| \log \frac{ep}{|I|}. \tag{2.3}$$

It should be noted that the quantity $r_\tau(I_o^\tau, \theta)$, called the *τ-oracle rate*, is unique but the τ-oracle I_o^τ itself need not be. If there are multiple minimizers, any one can be chosen for our purpose. For instance, if $\theta = 0$, then $\{i\}$ for any $i = 1, \dots, p$, is a τ-oracle. Note that the empty set is not allowed to be the oracle by the definition. The rate (2.3) is clearly a modification of the prediction risk (2.2) where the variance part has been inflated by the factor $\tau \log(ep/|I|)$. Denote from now on $r_\tau(\theta) = r_\tau(I_o^\tau, \theta)$, and let $r(I, \theta) = r_1(I, \theta)$, $I_o = I_o^1$ and $r(\theta) = r_1(I_o^1, \theta)$ correspond to the case of the *standard oracle* (i.e., $\tau = 1$).

2.2. *Empirical Bayes*: *EBMS and EBMA*. It is well known that in the normal regression setting, a normal prior for θ is conjugate, but it overshrinks toward the prior mean. This property led Castillo et al. [11] in their study of posterior contraction to consider priors with tails at least as thick as that of the Laplace distribution for each coefficient selected as nonzero. Nevertheless, a normal prior is useful if the prior mean is not fixed but instead is chosen by an empirical Bayes technique, as noted in Belitser [2] and Belitser and Nurushev [4]. Below we follow the same path because the explicit expressions resulting from conjugacy allow us deriving useful bounds for the procedure in terms of the local rate. Although our primary interest is in the coverage of (empirical Bayes) adaptive credible regions of optimal size, en route the proof we also derive posterior contraction results analogous to those in Castillo et al. [11] but based on a spike-and-slab conjugate normal prior with hyperparameters chosen by the empirical Bayes technique. The exact expressions that we obtain lead to a significantly shorter derivation of equivalent results, and more importantly, allow obtaining coverage of credible regions.

For the model (1.1) with $\varepsilon \sim \mathrm{N}_n(0, \mathrm{I})$, introduce a two-level hierarchical prior,

$$\pi_I(\theta): \quad \theta | I \sim \mathrm{N}\big(\mu(I), \kappa\sigma^2(\mathrm{X}_I'\mathrm{X}_I)^{-1}\big) \otimes \delta_{0_{|I^c|}}, \tag{2.4}$$

$$\lambda(I): \quad I \sim (\lambda_I : I \in \mathcal{I}), \qquad \lambda_I = c_\varkappa e^{-\varkappa |I| \log(ep/|I|)}, \quad I \in \mathcal{I}, \tag{2.5}$$

where $\kappa > 0$, $\varkappa > 1$ are chosen constants, vectors $\mu_{\mathcal{I}} = (\mu(I) : I \in \mathcal{I})$ are to be chosen by the empirical Bayes method and $c_\varkappa = c_{\varkappa,p}$ is the normalizing constant making $\sum_{I\in\mathcal{I}} \lambda_I = 1$. The prior on θ given I is Zellner's g-prior, which was also used in the same context by Martin et al. [22] along with an empirical Bayes choice for the mean $\mu(I)$ like ours. However, they restricted to the situation where the number of true predictors is less than the sample size and considered the pseudo-posterior distribution obtained from a power of the likelihood less than 1, and they did not study coverage properties of credible sets. We note that the prior λ for I given by (2.5) is proper since $\binom{p}{k} \le (ep/k)^k$ and

$$\sum_{I\in\mathcal{I}} \lambda_I \le c_\varkappa \sum_{k=1}^{p} \binom{p}{k} \Big(\frac{ep}{k}\Big)^{-\varkappa k} \le c_\varkappa \sum_{k=1}^{p} \Big(\frac{ep}{k}\Big)^{-(\varkappa-1)k} \le c_\varkappa \sum_{k=1}^{p} e^{-(\varkappa-1)k} = \frac{c_\varkappa e^{-(\varkappa-1)}}{1 - e^{-(\varkappa-1)}} < \infty,$$

which also implies that $c_\varkappa \ge e^{\varkappa-1} - 1 > 0$. The parameters κ and $\varkappa$ need to satisfy

$$\kappa \ge (1-h)^{-1/h} - 1, \qquad \varkappa > 2/h \quad \text{for some } h \in (0, 1/3). \tag{2.6}$$

These conditions will be assumed throughout below.

Simple calculations involving normal-normal conjugacy lead to explicit expressions for the marginal distribution $\pi_I(Y)$ of Y given I and the posterior distribution $\pi(\theta|I, Y)$ of θ given I and Y dependent on μ. Then μ can be estimated by maximizing the marginal likelihood, which is given by the least square estimate; see Section 1 of the Supplementary Material [3]. Upon substituting the estimator $\hat{\mu}(I)$ in place of $\mu(I)$ for every I, the empirical Bayes marginal distribution $\hat{\pi}_I(Y) = \pi_{I,\hat{\mu}}(Y)$ and the empirical Bayes posterior distribution $\hat{\pi}(\theta|I, Y)$ can be seen to be given by

$$\begin{aligned} \hat{\pi}_I(Y) &= \frac{\exp\{-\frac{1}{2\sigma^2} Y'(\mathrm{I} - \mathrm{H}_I)(\mathrm{I} + \kappa\mathrm{H}_I)^{-1}(\mathrm{I} - \mathrm{H}_I)Y\}}{\sqrt{(2\pi\sigma^2)^n \det(\mathrm{I} + \kappa\mathrm{H}_I)}} \\ &= \frac{\exp\{-\frac{1}{2\sigma^2} Y'(\mathrm{I} - \mathrm{H}_I)Y\}}{(2\pi\sigma^2)^{n/2}(1+\kappa)^{|I|/2}}, \end{aligned} \tag{2.7}$$

$$\hat{\pi}(\theta|I, Y) = \mathrm{N}_{|I|}\Big(\hat{\theta}_I(I), \frac{\kappa\sigma^2}{\kappa+1}(\mathrm{X}_I'\mathrm{X}_I)^{-1}\Big) \otimes \delta_{0_{|I^c|}}. \tag{2.8}$$

Above we have used the fact that, as H_I is a projection matrix with $\operatorname{rank}(\mathrm{H}_I) = |I|$, we have that $\det(\mathrm{I} + \kappa\mathrm{H}_I) = (1+\kappa)^{|I|}$ and

$$(\mathrm{I} - \mathrm{H}_I)(\mathrm{I} + \kappa\mathrm{H}_I)^{-1}(\mathrm{I} - \mathrm{H}_I) = (\mathrm{I} - \mathrm{H}_I)\Big(\mathrm{I} - \frac{\kappa}{1+\kappa}\mathrm{H}_I\Big)(\mathrm{I} - \mathrm{H}_I) = \mathrm{I} - \mathrm{H}_I.$$

To obtain the final posterior distribution of θ given Y, we need to eliminate the hyperparameter I from the expression. We may either estimate I by the empirical Bayes approach and substitute its value in the above expressions to obtain the *empirical Bayes model selection* (EBMS) method, or we can sum over I weighted by its (empirical Bayes) posterior distribution to obtain the *empirical Bayes model averaging* (EBMA) method. In the EBMS approach, the resulting posterior density of θ is given by

$$\check{\pi}(\theta|Y) = \pi_{\hat{I},\hat{\mu}}(\theta|Y) = \mathrm{N}_{|\hat{I}|}\Big(\hat{\theta}_{\hat{I}}(\hat{I}), \frac{\kappa\sigma^2}{\kappa+1}(\mathrm{X}_{\hat{I}}'\mathrm{X}_{\hat{I}})^{-1}\Big) \otimes \delta_{0_{|\hat{I}^c|}}, \tag{2.9}$$

$$\hat{I} = \arg\max_{I\in\mathcal{I}} \lambda_I \hat{\pi}_I(Y). \tag{2.10}$$

If the maximizer is not unique, any maximizer may be chosen. Note that the EBMS estimator $\check{\theta} = \check{\mathrm{E}}(\theta|Y) = \hat{\theta}(\hat{I})$ can also be viewed as a penalized estimator since

$$\hat{I} = \arg\min_{I\in\mathcal{I}}\{\|Y - \mathrm{X}\hat{\theta}(I)\|^2 + |I|\log(1+\kappa)^{1/2} + |I|\varkappa\log(ep/|I|)\}. \tag{2.11}$$

In the EBMA approach, the posterior density of θ is given by $\tilde{\pi}(\theta|Y) = \sum_{I\in\mathcal{I}}\tilde{\pi}_I(\theta|Y)\times\tilde{\pi}(I|Y)$, where

$$\begin{aligned}\tilde{\pi}(I|Y) &= \frac{\lambda_I\hat{\pi}_I(Y)}{\sum_{J\in\mathcal{I}}\lambda_J\hat{\pi}_J(Y)},\\ \tilde{\pi}_I(\theta|Y) &= \mathrm{N}_{|I|}\biggl(\hat{\theta}_I(I), \frac{\kappa\sigma^2}{\kappa+1}(\mathrm{X}_I'\mathrm{X}_I)^{-1}\biggr)\otimes\delta_{0_{|I^c|}},\end{aligned} \tag{2.12}$$

and $\hat{\pi}_I(Y)$ is defined by (2.7). Then the *EBMA posterior mean* is given by $\tilde{\theta} = \tilde{\mathrm{E}}(\theta|Y) = \sum_{I\in\mathcal{I}}\tilde{\mathrm{E}}_I(\theta|Y)\tilde{\pi}(I|Y) = \sum_{I\in\mathcal{I}}\hat{\theta}(I)\tilde{\pi}(I|Y)$.

REMARK 1. Notice that $\hat{I}$ can be seen as minimizing the expression in (2.11) over the whole family of subsets $\bar{\mathcal{I}} = \{I\subseteq\{1,\dots,p\} : I\neq\varnothing\}$, rather than just over $\mathcal{I}$, as the second and the third term for I outside $\mathcal{I}$ only increase the expression without decreasing the first term. Also the oracle rate $r(\theta)$ can be seen as the minimizer of $r(I,\theta)$ over the whole family $\bar{\mathcal{I}}$ because these quantities are defined in terms of projections H_I and cardinalities $|I|$ which are well-defined for any $I\in\bar{\mathcal{I}}$.

2.3. *Prediction, support and signal recovery.* In the sequel, by $\hat{\pi}(\theta|Y)$ (with the corresponding expectation $\hat{\mathrm{E}}(\cdot|Y)$) we denote either the EBMS posterior $\check{\pi}(\theta|Y)$ defined by (2.9) or the EBMA posterior $\tilde{\pi}(\theta|Y)$ defined by (2.12), and $\hat{\theta}$ will stand respectively either for $\check{\theta}$ or $\tilde{\theta}$. Then $\hat{\pi}(I\in\mathcal{G}|Y)$ should be read as $\mathbb{1}\{\hat{I}\in\mathcal{G}\}$ in the case $\hat{\pi} = \check{\pi}$, and as $\tilde{\pi}(I\in\mathcal{G}|Y)$ in the case $\hat{\pi} = \tilde{\pi}$, for all $\mathcal{G}\subseteq\mathcal{I}$ that appear in what follows. Respectively, $\hat{\pi}(I|Y) = \mathbb{1}\{\hat{I} = I\}$ and $\mathrm{E}_{\theta_0}\hat{\pi}(I\in\mathcal{G}|Y) = \mathrm{P}_{\theta_0}(\hat{I}\in\mathcal{G})$ in the former case, and $\hat{\pi}(I|Y) = \tilde{\pi}(I|Y)$ and $\mathrm{E}_{\theta_0}\hat{\pi}(I\in\mathcal{G}|Y) = \mathrm{E}_{\theta_0}\tilde{\pi}(I\in\mathcal{G}|Y)$ in the latter case.

The following theorem shows that most of the $\hat{\pi}$-posterior mass of $\mathrm{X}\theta$ concentrate near the true value $\mathrm{X}\theta_0$ in the frequentist sense and the empirical Bayes posterior mean $\mathrm{X}\hat{\theta}$ converges to $\mathrm{X}\theta_0$ uniformly over the entire parameter space, at the oracle rate $r(\theta_0)$.

THEOREM 1 (Prediction). *There exists a constant $C>0$ such that, for any $\theta_0\in\mathbb{R}^p$, $M>0$, $\mathrm{E}_{\theta_0}\hat{\pi}(\|\mathrm{X}\theta - \mathrm{X}\theta_0\|\geq Mr(\theta_0)|Y)\leq CM^{-2}$ and $\mathrm{E}_{\theta_0}\|\mathrm{X}\hat{\theta} - \mathrm{X}\theta_0\|^2\leq Cr^2(\theta_0)$.*

Notice that $\mathrm{X}\hat{\theta} = \mathrm{H}_{\hat{I}}Y$. The result is analogous to the first conclusion of Theorem 2 of Castillo et al. [11] for the prior (2.4)–(2.5) (with the empirical Bayes choice for the hyperparameter $(\mu(I), I\in\mathcal{I})$) replacing their prior. This follows by observing that for θ_0 with sparsity $s(\theta_0) = |I^*(\theta_0)|\geq 1$, where $I^* = I^*(\theta_0) = \mathrm{supp}(\theta_0) = \{i : \theta_{0i}\neq 0\}$ is the true active index set, we have that $(\mathrm{I} - \mathrm{H}_{I^*})\mathrm{X}\theta_0 = 0$ and, by the definition of the oracle,

$$r^2(\theta_0)\leq r^2(I^*,\theta_0)\leq\sigma^2 s(\theta_0)\log\bigl(ep/s(\theta_0)\bigr). \tag{2.13}$$

Note that we do not need the compatibility condition connecting ℓ_1- and ℓ_2-norms since we do not use Laplace priors. The symbol $s(\theta_0)$ will be reserved throughout for the sparsity level at the true parameter θ_0, that is, the number of nonzero elements of θ_0.

We also obtain the following result as a by-product of the proof of Theorem 1.

COROLLARY 1 (Weak support recovery and sparsity control). *There exist positive constants m_0, c', c'', τ_0 such that, for any $\theta_0\in\mathbb{R}^p$ and $m>0$:*

(i) $\mathrm{E}_{\theta_0}\hat{\pi}(I \in \mathcal{I} : r^2(I,\theta_0) \geq m_0 r^2(\theta_0) + m\sigma^2 | Y) \leq e^{-c'm}$,
(ii) $\mathrm{E}_{\theta_0}\hat{\pi}(I \in \mathcal{I} : |I| \geq \tau_0 s(\theta_0) | Y) \leq \exp\{-c'' s(\theta_0) \log(ep/s(\theta_0))\}$.

One can interpret the first claim (i) of Corollary 1 as some sort of *weak support recovery*: the posterior $\tilde{\pi}(I|Y)$ and the variable selector $\hat{I}$ "live" on those index sets $I \in \mathcal{I}$ whose rate $r^2(I,\theta_0)$ is close to the oracle rate $r^2(\theta_0)$, uniformly in $\theta_0 \in \mathbb{R}^p$. Notice that no conditions are needed. The second claim (ii) provides sparsity control from above: it ensures basically that models with substantially higher size than the true one are unlikely to occur under the empirical Bayes posterior. This property and Corollaries 2 and 3 below are analogous to the corresponding results of Castillo et al. [11] for the posterior distribution using Laplace priors on active coefficients and also of Martin et al. [22] for their empirical Bayes pseudo-posterior distribution. Notice that Lemma 2 (auxiliary result needed in the proof of Theorem 2) provides sparsity control also from below, but in terms of the τ-oracle.

Corollary 1, being nonasymptotic and uniform in θ_0, can be specialized to certain situations. In particular, assertion (ii) leads to an interesting conclusion under the asymptotic setting $p = p_n \to \infty$ and $1 \leq s(\theta_0) \leq s_n = o(p_n)$ as $n \to \infty$. Then the probability bound goes to 0 as $n \to \infty$, uniformly in $\theta_0 \in \ell_0[s_n] \triangleq \{\theta : |\operatorname{supp}(\theta)| \leq s_n\}$. Further, when $s_n = o(p_n)$, the constant τ_0 can be chosen smaller, which makes the conclusions in Corollaries 2 and 3 below stronger.

It is of substantial interest to know if the empirical Bayes procedure also gives concentration for the posterior of θ near its true value θ_0, measured by the usual Euclidean or the ℓ_1-distance. Because the dimension p may be larger than n, the correspondence between $\mathrm{X}\theta$ and θ is not unique, and hence additional conditions are necessary even in the noiseless situation. As is commonly adopted in the literature (see, e.g., van de Geer and Bühlmann [33]), we too assume a condition that lower bounds the norm of $\mathrm{X}\theta$ by a positive multiple of a norm on θ for sparse vectors, which is a condition on the design matrix X.

Recall that $s(\theta)$ denotes the number of nonzero elements of $\theta \in \mathbb{R}^p$, that is, the cardinality of the support $\operatorname{supp}(\theta) = I^*(\theta) = \{i : \theta_i \neq 0\}$ of θ. Let $\|\theta\|_1 = \sum_{j=1}^p |\theta_j|$ be the ℓ_1-norm of θ. Define $\|\mathrm{X}\|_{\max} = \max_{k=1,\ldots,p} \|X_k\|$ (notice that if the design matrix X is normalized so that $\|X_k\|^2 = n$, $k = 1, \ldots, p$, then $\|\mathrm{X}\|_{\max} = \sqrt{n}$). For a positive integer m, define

$$\phi_1(m) = \inf\left\{ \frac{\sqrt{m}\|\mathrm{X}\theta\|}{\|\mathrm{X}\|_{\max}\|\theta\|_1} : 0 \neq |\operatorname{supp}(\theta)| \leq m \right\}, \tag{2.14}$$

$$\phi_2(m) = \inf\left\{ \frac{\|\mathrm{X}\theta\|}{\|\mathrm{X}\|_{\max}\|\theta\|} : 0 \neq |\operatorname{supp}(\theta)| \leq m \right\}. \tag{2.15}$$

Because $\|\theta\|_1 \leq \sqrt{s(\theta)}\|\theta\|$, it follows that $\phi_1(m) \geq \phi_2(m)$. Positivity of ϕ_1 at an argument m is called the compatibility condition, and is stronger if $\phi_1(m)$ is larger. Since

$$\|\theta_I\| = \|(\mathrm{X}_I'\mathrm{X}_I)^{-1}\mathrm{X}_I'\mathrm{X}_I\theta_I\| \leq \|(\mathrm{X}_I'\mathrm{X}_I)^{-1}\|\|\mathrm{X}_I'\|\|\mathrm{X}_I\theta_I\| = \frac{\sqrt{\lambda_{\max}(\mathrm{X}_I'\mathrm{X}_I)}}{\lambda_{\min}(\mathrm{X}_I'\mathrm{X}_I)}\|\mathrm{X}_I\theta_I\|$$

for any $I \in \mathcal{I}$, it follows that $\phi_2(m) \geq \inf\{\lambda_{\min}(\mathrm{X}_I'\mathrm{X}_I)/\|\mathrm{X}\|_{\max}\sqrt{\lambda_{\max}(\mathrm{X}_I'\mathrm{X}_I)}, |I| \leq m, I \in \mathcal{I}\}$, where $\lambda_{\min}(\mathrm{M})$ and $\lambda_{\max}(\mathrm{M})$ refer respectively to the minimal and maximal eigenvalues of a square matrix M. Hence, $\phi_2(m)$ can be bounded in terms of eigenvalues of submatrices of $\mathrm{X}'\mathrm{X}$. In the result below, if any of ϕ_1 or ϕ_2 is zero at its argument, then the result becomes trivial but remains valid.

COROLLARY 2 (Signal recovery). *Let ϕ_1 and ϕ_2 be defined by* (2.14) *and* (2.15), *respectively. For constants C, τ_0, c'' appearing in Theorem* 1 *and Corollary* 1, *we have*

$$\mathrm{E}_{\theta_0}\hat{\pi}\left(\|\theta-\theta_0\|_1 \geq \frac{Mr(\theta_0)\sqrt{(\tau_0+1)s(\theta_0)}}{\|\mathrm{X}\|_{\max}\phi_1((\tau_0+1)s(\theta_0))}\Big|Y\right) \leq CM^{-2}+\exp\bigl[-c''s(\theta_0)\log\bigl(ep/s(\theta_0)\bigr)\bigr],$$

$$\mathrm{E}_{\theta_0}\hat{\pi}\left(\|\theta-\theta_0\| \geq \frac{Mr(\theta_0)}{\|\mathrm{X}\|_{\max}\phi_2((\tau_0+1)s(\theta_0))}\Big|Y\right) \leq CM^{-2}+\exp\bigl[-c''s(\theta_0)\log\bigl(ep/s(\theta_0)\bigr)\bigr].$$

Selecting the correct set of predictors is also relevant, but is clearly impossible without assuming that the signals are sufficiently strong, since estimating a very small coefficient introduces more error than the bias for ignoring it. Assuming that all signals are significantly large, the following result shows that the empirical Bayes posterior (and the variable selector $\hat{I}$) can rarely miss a signal; this is called the "variable screening property" in Bühlmann [6]. Notice that the condition in the corollary is a version of the so-called "beta-min" condition; cf. Castillo et al. [11] and Bühlmann [6].

COROLLARY 3 (Selection). *Let ϕ_2 be defined by* (2.15) *and assume that*

$$|\theta_{0,j}| > M\sigma\|\mathrm{X}\|_{\max}^{-1}\sqrt{s(\theta_0)\log\bigl(ep/s(\theta_0)\bigr)}/\phi_2\bigl((\tau_0+1)s(\theta_0)\bigr),$$

whenever $\theta_{0,j}\neq 0$. Then for constants C, τ_0, c'' appearing in Theorem 1 *and Corollary* 1, *we have* $\mathrm{E}_{\theta_0}\hat{\pi}(\theta:\operatorname{supp}(\theta)\not\supset\operatorname{supp}(\theta_0)|Y) \leq CM^{-2}+\exp[-c''s(\theta_0)\log(ep/s(\theta_0))]$.

From Corollary 3, variable selection consistency follows by an easy argument when the actual support is sufficiently small; see Martin et al. [22].

2.4. *Uncertainty quantification.* While the empirical Bayes method has the optimal accuracy for point estimation (matching with the local oracle rate), the question of frequentist coverage and optimal size of empirical Bayes credible balls is a lot more delicate. As mentioned in the Introduction, it is impossible to construct any procedure, Bayesian or otherwise, which will have the optimal size and required coverage uniformly all over the parameter space. If we wish to obtain the optimal size of a confidence ball with high probability uniformly in the parameter space, the best we can achieve is coverage only after removing a (small) collection of deceptive parameters. The main problem is in the bias term $\|(\mathrm{I}-\mathrm{H}_{I_o^\tau(\theta)})\mathrm{X}\theta\|^2$ of the τ-oracle rate, that results from the fact that the coordinates $\theta_{(I_o^\tau)^c}=(\theta_i, i\notin I_o^\tau)$ are classified as zeros by the τ-oracle, but actually these may not be zeros. Although this is the best choice from the τ-oracle's perspective, some "deceptive" θ values may still have many nonzero coordinates from $(I_o^\tau)^c$ (being classified as zeros), making the bias term large (i.e., "the bias is excessive") relative to the variance term. The idea is to control the excessive bias via the (inflated) oracle variance, which means that the whole oracle rate is then controlled by the variance term.

Define the normalized τ-bias at $\theta_0\in\mathbb{R}^p$ by

$$b_\tau(\theta_0)=\frac{\theta_0'\mathrm{X}'(\mathrm{I}-\mathrm{H}_{I_o^\tau(\theta_0)})\mathrm{X}\theta_0}{\sigma^2|I_o^\tau(\theta_0)|\log(ep/|I_o^\tau(\theta_0)|)}, \tag{2.16}$$

and hence we may write $r_\tau^2(\theta_0)=(b_\tau(\theta_0)+\tau)\log(ep/|I_o^\tau(\theta_0)|)\sigma^2|I_o^\tau(\theta_0)|$. The τ-oracle rate becomes a multiple of the variance term of the rate. However, to make this multiple factor uniform over $\theta_0\in\mathbb{R}^p$, a control over the normalized τ-bias $b_\tau(\theta_0)$ will be needed. This naturally leads to a restriction, which will ensure coverage at θ_0.

DEFINITION 1. We say that the *excessive bias restriction* (EBR) condition with structural parameters (t, τ) holds at a parameter value θ_0 if $b_\tau(\theta_0) \le t$. We denote the corresponding region of the parameter space by $\Theta_{\mathrm{eb}}(t, \tau) = \{\theta \in \mathbb{R}^p : b_\tau(\theta) \le t\}$.

A discussion on the EBR condition is given in Section 3. We shall need the EBR condition satisfied for

$$\tau > \bar{\tau} \triangleq \big[4\varkappa(e+1) + 2e\log(1+\kappa)\big]/(e-2). \tag{2.17}$$

REMARK 2. The origin of the condition (2.17) can be elucidated as follows. Recall that by the definition of the oracle rate, for any $\theta \in \mathbb{R}^p$ we have $|I_o^{\tau_2}(\theta)| \le |I_o^{\tau_1}(\theta)| \le |I^*(\theta)|$ for any $0 \le \tau_1 \le \tau_2$. This means that the EBR condition is weaker if τ is smaller. Interestingly, the "limiting" oracle, as $\tau \downarrow 0$, $I_o^0(\theta) = I^*(\theta)$ recovers the active index set $I^*(\theta)$ and the corresponding limiting EBR set would become the entire space: $\Theta_{\mathrm{eb}}(t, 0) = \mathbb{R}^p$. Besides, checking the proofs reveals that the constants will not deteriorate for smaller τ (in fact, will improve). However, as condition (2.17) shows, Lemma 2 (and hence Theorem 2) hold only if τ is appropriately large, thus ruling out the possibility that arbitrarily weak versions of the EBR condition can ensure coverage.

Now we construct a confidence ball for $\mathrm{E}_\theta(Y) = \mathrm{X}\theta$ using the EBMS posterior distribution $\check{\pi}(\theta|Y)$. From (2.9), we have $\mathrm{X}\theta|Y \sim \mathrm{N}_n(\mathrm{X}\check{\theta}, \frac{\kappa\sigma^2}{\kappa+1}\mathrm{H}_{\hat{I}})$. Denoting by $\chi^2_{k,\alpha}$ the $(1-\alpha)$-quantile of χ^2_k-distribution, we have that $\hat{\pi}(\|\mathrm{X}\theta - \mathrm{X}\check{\theta}\|^2 \le \frac{\kappa\sigma^2}{\kappa+1}\chi^2_{|\hat{I}|,\alpha}|Y) = 1-\alpha$. However, $\chi^2_{|\hat{I}|,\alpha}$ is bounded by a constant multiple of $|\hat{I}|$, and hence for simplicity the latter can replace the former to obtain a credible ball. Instead of EBMS posterior mean $\check{\theta}$, the EBMA posterior mean $\tilde{\theta}$ may also be used as the center of the resulting confidence ball. This leads to the confidence set $B(\mathrm{X}\hat{\theta}, M\sigma|\hat{I}|^{1/2})$ as a credible ball for $\mathrm{X}\theta$, which clearly can be guaranteed to have at least a given level of credibility by choosing a sufficiently large constant M. We shall see that under a true parameter θ_0, with arbitrarily high probability, $|\hat{I}|$ is of the order $|I_o^\tau|$. Since the minimax risk of the class of adaptive estimators is larger than this by a multiplicative factor $\log(ep/|I_o^\tau|)$, it is clear that $B(\mathrm{X}\hat{\theta}, M\sigma|\hat{I}|^{1/2})$ cannot have a guaranteed coverage, since otherwise the center $\hat{\theta}$ will have risk lower than the minimax risk for sparse adaptive estimation. Hence to obtain coverage, the order of the radius of any confidence ball must contain a logarithmic factor. This leads us to the inflated credible ball $B(\mathrm{X}\hat{\theta}, M\hat{\rho})$, where

$$\hat{\rho}^2 = \sigma^2|\hat{I}|\log(ep/|\hat{I}|). \tag{2.18}$$

The following theorem shows that the logarithmic inflation leads to coverage at all true parameter θ_0 satisfying the EBR condition.

THEOREM 2 (Coverage and oracle size). *Let* (2.17) *be fulfilled. Then for any* $t > 0$, $\epsilon_1, \epsilon_2 > 0$, *there exist* $M = M(t, \epsilon_1) > 0$ *and* $L = L(\epsilon_2) > 0$ *such that*

$$\sup_{\theta_0 \in \Theta_{\mathrm{eb}}(t,\tau)} \mathrm{P}_{\theta_0}\big(\mathrm{X}\theta_0 \notin B(\mathrm{X}\hat{\theta}, M\hat{\rho})\big) \le \epsilon_1, \qquad \sup_{\theta_0 \in \mathbb{R}^p} \mathrm{P}_{\theta_0}\big(\hat{\rho} \ge Lr(\theta_0)\big) \le \epsilon_2.$$

REMARK 3. The EBR condition is formulated in terms of τ-oracle I_o^τ defined by (2.3), rather than the "standard" oracle I_o (i.e., for $\tau = 1$). It may be desirable to impose an EBR condition in terms of the oracle I_o. Notice that the τ-oracle can be seen as the standard oracle but in the model with the new variance parameter $\sigma_\tau^2 = \tau\sigma^2$ instead of σ^2. By rewriting the original model (1.1) as $Y\tau^{-1/2} = \mathrm{X}\theta\tau^{-1/2} + \sigma\tau^{-1/2}\varepsilon$, it is not difficult to see that we can construct a confidence ball for $\mathrm{X}\theta$ with the radius $\sqrt{\tau}M\hat{\rho}$ satisfying the coverage property as above, but now uniformly over $\Theta_{\mathrm{eb}}(t, 1)$.

REMARK 4. Interestingly, as is recently shown by Castillo and Szabó [12] for the "signal plus noise" model, the EBR condition turns out to be minimal in a certain sense.

REMARK 5. We can construct a confidence ball with a radius of the order $r(\theta_0)+\sigma n^{1/4}$ with coverage uniformly over the whole space $\theta_0 \in \mathbb{R}^p$. This means that there is no EBR condition needed if the extra information $r_o^2(\theta_0) \geq C\sigma^2\sqrt{n}$ (the parameter θ_0 is "not sparse" in a sense) is available (cf. Nickl and van de Geer [24]). However, the construction of that confidence ball is different from the confidence ball in Theorem 2. It is an interesting problem to investigate whether it is possible to test between "sparse" and "nonsparse" cases.

REMARK 6. While it is appropriate to treat $B(\mathrm{X}\hat\theta, M\hat\rho)$ as a credible and confidence region for $\mathrm{E}_\theta Y = \mathrm{X}\theta$, it is possible to interpret $\{\theta : \mathrm{X}\theta \in B(\mathrm{X}\hat\theta, M\hat\rho)\}$ as a region for θ, which automatically "inherits" the corresponding credibility and coverage. However, as the matrix $\mathrm{X}'\mathrm{X}$ is typically rank deficient, the set $\{\theta : (\theta-\hat\theta)'\mathrm{X}'\mathrm{X}(\theta-\hat\theta) \leq M^2\hat\rho^2\}$ obtained from the corresponding quadratic form will be unbounded in terms of the Euclidean distance. The main difficulty is that the relation $\theta \mapsto \mathrm{X}\theta$ is not invertible. Hence, a bounded diameter credible and confidence region is possible only by imposing an additional restriction on the parameter θ through the compatibility condition that allows the inversion.

COROLLARY 4. *Let* (2.17) *hold. Then for any* $t>0$, $\theta_0 \in \Theta_{\mathrm{eb}}(t,\tau)$, $\epsilon_1,\epsilon_2>0$, *there exist* $M = M(t,\epsilon_1) > 0$, $L = L(\epsilon_2) > 0$ *and positive constants* τ_0, c'' *such that*

$$\mathrm{P}_{\theta_0}\bigl(\|\theta_0-\hat\theta\| \geq M\hat\rho/\bigl[\|\mathrm{X}\|_{\max}\phi_2\bigl((\tau_0+1)s(\theta_0)\bigr)\bigr]\bigr) \leq \epsilon_1 + \exp\biggl\{-c''s(\theta_0)\log\biggl(\frac{ep}{s(\theta_0)}\biggr)\biggr\},$$

$$\begin{aligned}&\mathrm{P}_{\theta_0}\bigl(\|\theta_0-\hat\theta\|_1 \geq M\sqrt{s(\theta_0)}\hat\rho/\bigl[\|\mathrm{X}\|_{\max}\phi_1\bigl((\tau_0+1)s(\theta_0)\bigr)\bigr]\bigr)\\ &\quad\leq \epsilon_1 + \exp\biggl\{-c''s(\theta_0)\log\biggl(\frac{ep}{s(\theta_0)}\biggr)\biggr\},\end{aligned}$$

and $\sup_{\theta_0\in\mathbb{R}^p}\mathrm{P}_{\theta_0}(\hat\rho \geq Lr(\theta_0)) \leq \epsilon_2$.

The first two claims can be made uniform over $\theta_0 \in \Theta_{\mathrm{eb}}(t,\tau)\cap\ell_0[s]$ with $s \geq 1$:

$$\sup_{\theta_0\in\Theta_{\mathrm{eb}}(t,\tau)\cap\ell_0[s]}\mathrm{P}_{\theta_0}\bigl(\|\theta_0-\hat\theta\| \geq M\hat\rho/\bigl[\|\mathrm{X}\|_{\max}\phi_2\bigl((\tau_0+1)s\bigr)\bigr]\bigr) \leq \epsilon_1 + \exp\bigl[-c''\log(ep)\bigr],$$

and similarly for the second relation. The diameter of the confidence ball for θ constructed above is $M\hat\rho/[\|\mathrm{X}\|_{\max}\phi_2((\tau_0+1)s(\theta_0))]$. As $\hat\rho$ is of the oracle order $r(\theta_0)$ with large probability, it follows that the diameter of the confidence set for $\theta \in \Theta_{\mathrm{eb}}(t,\tau)\cap\ell_0[s]$ is at most of the order $r(\theta_0)/[\|\mathrm{X}\|_{\max}\phi_2((\tau_0+1)s)] \leq \sigma\sqrt{s\log(ep/s)}/[\|\mathrm{X}\|_{\max}\phi_2((\tau_0+1)s)]$ with high probability.

To construct a credible ball using the EBMA posterior, we can choose the radius $\tilde\rho$ to be that of the smallest $1-\gamma$ credible ball:

$$\tilde\rho = \inf\bigl\{\rho : \tilde\pi\bigl(\|\mathrm{X}\theta-\mathrm{X}\hat\theta\| \leq \rho|Y\bigr) \geq 1-\gamma\bigr\}. \tag{2.19}$$

Next, we can construct the confidence ball $B(\mathrm{X}\hat\theta, M\tilde\rho)$, and it can even be shown that the quantity $\tilde\rho^2$ is of the order $\sigma^2|I_o|$. As in the EBMS posterior, this concentration rate around its center is faster than the oracle rate $r^2(\theta_0)$, which is the rate at which the EBMA posterior mean concentrates near the truth. Hence in order to ensure coverage, logarithmic inflation of the radius $\tilde\rho$ is needed. In doing so, we attain the full coverage under EBR, sacrificing slightly on the size property. The following theorem describes coverage and size precisely. Its proof is given in the Supplementary Material [3].

THEOREM 3 (EBMA coverage and oracle size). *Let* (2.17) *be fulfilled. Then for any* $t > 0$ *and* $\epsilon_1, \epsilon_2 > 0$, *there exist* $M = M(t, \epsilon_1) > 0$ *and* $L = L(\epsilon_2) > 0$ *such that*

$$\sup_{\theta_0 \in \Theta_{\mathrm{eb}}(t,\tau)} \mathrm{P}_{\theta_0}\big(\mathrm{X}\theta_0 \notin B\big(\mathrm{X}\hat{\theta}, M\big[\log(ep)\big]^{1/2}\tilde{\rho}\big)\big) \le \epsilon_1, \qquad \sup_{\theta_0 \in \mathbb{R}^n} \mathrm{P}_{\theta_0}\big(\tilde{\rho} \ge Lr(\theta_0)\big) \le \epsilon_2.$$

REMARK 7. By analyzing Lemma 3 in the Supplementary Material [3] (and subsequent proof of Theorem 3), we see that we can replace the log factor $[\log(ep)]^{1/2}$ by $[\log\log(ep)]^{1/2}$ in the coverage relation of Theorem 3 if $|I_o^\tau| \ge C\log(ep)/\log\log(ep)$.

One can also show that the coverage relation in Theorem 3 holds uniformly over the class of the very sparse parameters $\theta \in \ell_0[s_n]$ with $s_n \le c\log p$.

3. Discussions and extensions.

Random predictors. In our setting, for posterior concentration and coverage results, we treated the predictors as deterministic. As the posterior distribution is obtained by conditioning on the given data, all Bayesian procedures remain unchanged even if the predictors are random, as long as they have a fixed distribution free of the unknown model parameters. However, frequentist behaviors of Bayesian procedures are different when predictors are considered to be deterministic or random. Clearly, conditional on the observed predictors, the frequentist behavior of a Bayesian procedure is given by the corresponding result for deterministic predictors. Since the oracle approach measures errors "locally at a parameter," this allows deriving concentration and coverage results for random predictors through those for deterministic predictors, as shown below.

When the predictors $\mathrm{X} = (X_1, \ldots, X_p)$ are obtained randomly, we first define the risk function and the oracle rates (see (2.3)) conditional on the realization of X. Let $r(\theta_0|\mathrm{X})$ stand for the oracle rate at parameter θ_0 given X, and let $r^2(\theta_0) = \mathrm{E}_{\mathrm{X}} r^2(\theta_0|\mathrm{X})$, where E_{X} stands for the expectation with respect to the distribution of X. Note that the conclusion of Theorem 1, conditioned on a realized value of a random predictor X, can be strengthened to $\mathrm{E}_{\theta_0}[\hat{\mathrm{E}}\{\|\mathrm{X}\theta - \mathrm{X}\theta_0\|^2|Y\}|\mathrm{X}] \le Cr^2(\theta_0|X)$ for some constant $C > 0$. Then the second part of Theorem 1 holds by Jensen's inequality while the first part follows from Markov's inequality. Clearly, Corollaries 1–3 follow with unconditional probabilities replacing the corresponding conditional probabilities given X, although the compatibility coefficients are also dependent on X. Note that the joint distribution of X can be completely arbitrary.

Extension of Theorem 2 is less immediate since the EBR condition also depends on X. For a structural parameter (t, τ) and $\epsilon > 0$, we say that the ϵ-EBR condition holds at θ_0 if $b_\tau(\theta_0) = b_\tau(\theta_0|\mathrm{X})$ defined by (2.16) satisfies $\mathrm{P}_{\mathrm{X}}(b_\tau(\theta_0|\mathrm{X}) > t) < \epsilon$. Clearly, the condition is implied by $\mathrm{E}_{\mathrm{X}}(b_\tau(\theta_0)|\mathrm{X}) < \epsilon t$. The last condition can be more easily understood if different replications $X^{(1)}, \ldots, X^{(n)}$ of the predictors are independent (or even just uncorrelated) with a common mean vector $M \in \mathbb{R}^p$ and covariance matrix Σ. Then denoting $\mathrm{H}_{I_o^\tau(\theta_0)}$ by H, the expectation of $\theta_0'\mathrm{X}'(\mathrm{I} - \mathrm{H})\mathrm{X}\theta_0$ is given by

$$\begin{aligned}&\theta_0'(M, M, \ldots, M)(\mathrm{I} - \mathrm{H})(M, M, \ldots, M)'\theta_0 + \mathrm{tr}\big[\big((\theta_0'\Sigma\theta_0)\mathrm{I}\big)(\mathrm{I} - \mathrm{H})\big]\\ &\quad = \big(M'\theta_0\big)^2\big\|(\mathrm{I} - \mathrm{H})(1, \ldots, 1)'\big\|^2 + \big(\theta_0'\Sigma\theta_0\big)\big(n - \mathrm{tr}(\mathrm{H})\big).\end{aligned}$$

Thus the ϵ-EBR condition with structural parameter holds at θ_0 if

$$\big(M'\theta_0\big)^2\big\|(\mathrm{I} - \mathrm{H})(1, \ldots, 1)'\big\|^2 + \big(\theta_0'\Sigma\theta_0\big)(n - \mathrm{tr}\,\mathrm{H})) < \epsilon t\sigma^2\big|I_0^\tau(\theta_0)\big|\log\big(ep/\big|I_0^\tau(\theta_0)\big|\big).$$

Clearly, for any θ_0 this holds for t sufficiently large, implying that the inflated credible set we constructed will have adequate coverage if the inflation factor is chosen properly.

Let the set of all θ satisfying the ϵ-EBR condition with structural parameters (t,τ) be denoted by $\Theta_{\epsilon\text{-eb}}(t,\tau)$. Then Theorem 2 can be stated as follows: if τ satisfies (2.17), $t>0$, $\epsilon_1,\epsilon_2>0$, then there exist $M=M(t,\epsilon_1)>0$ and $L=L(\epsilon_2)>0$ such that

$$\sup_{\theta_0\in\Theta_{\epsilon\text{-eb}}(t,\tau)} \mathrm{P}_{\theta_0}\big(\mathrm{X}\theta_0\notin B(\mathrm{X}\hat\theta, M\hat\rho)\big)\le\epsilon+\epsilon_1, \qquad \sup_{\theta_0\in\mathbb{R}^p}\mathrm{P}_{\theta_0}\big(\hat\rho\ge Lr(\theta_0)\big)\le\epsilon_2.$$

The first relation is evident from the original version of Theorem 2. To see the size condition, first obtain L' such that $\sup_{\theta_0\in\mathbb{R}^p}\mathrm{P}_{\theta_0}(\hat\rho\ge L'r(\theta_0|\mathrm{X}))\le\epsilon_2/2$ for all realizations of X. By Markov's inequality, find constant $L''>0$ such that $\mathrm{P}_{\mathrm{X}}(r(\theta_0|\mathrm{X})\ge L''r(\theta_0))\le\epsilon_2/2$. Now define $L=L'L''$ to conclude the proof. Corollary 4 clearly follows as well.

EBR condition. Next, we discuss the excessive bias restriction (EBR) condition. Let $I_i=I\cup\{i\}$, $I_{-i}=I\setminus\{i\}$ and H_I, H_{I_i} and $\mathrm{H}_{I_{-i}}$ be the projections onto the column spaces $\mathrm{col}(\mathrm{X}_I)$, $\mathrm{col}(\mathrm{X}_{I_i})$ and $\mathrm{col}(\mathrm{X}_{I_{-i}})$, respectively. Clearly, $\mathrm{H}_{I_i}-\mathrm{H}_I$ is the projection onto the orthogonal complement of $\mathrm{col}(\mathrm{X}_I)$ in $\mathrm{col}(\mathrm{X}_{I_i})$, which is zero if $X_i\in\mathrm{col}(\mathrm{X}_I)$. Otherwise, by some tedious direct calculations,

$$\mathrm{H}_{I_i}-\mathrm{H}_I=h_{I_i}h_{I_i}' \quad \text{where } h_{I_i}=\frac{(\mathrm{I}-\mathrm{H}_I)X_i}{\|(\mathrm{I}-\mathrm{H}_I)X_i\|}.$$

Similarly, if $X_i\notin\mathrm{col}(\mathrm{X}_{I,-i})$, then

$$\mathrm{H}_I-\mathrm{H}_{I_{-i}}=h_{I_{-i}}h_{I_{-i}}' \quad \text{where } h_{I_{-i}}=\frac{(\mathrm{I}-\mathrm{H}_{I_{-i}})X_i}{\|(\mathrm{I}-\mathrm{H}_{I_{-i}})X_i\|}.$$

For any $i\in\{1,\dots,p\}$, denote $I^\tau_{o,i}=I^\tau_o\cup\{i\}$. Then by the definition of the τ-oracle we have $r^2_\tau(I^\tau_{o,i},\theta_0)\ge r^2_\tau(I^\tau_o,\theta_0)$, yielding

$$\big\|(\mathrm{H}_{I^\tau_{o,i}}-\mathrm{H}_{I^\tau_o})\mathrm{X}\theta_0\big\|^2\le\tau\sigma^2\big(|I^\tau_o|+1\big)\log\frac{ep}{|I^\tau_o|+1}-\tau\sigma^2|I^\tau_o|\log\frac{ep}{|I^\tau_o|}\le\tau\sigma^2\log(ep).$$

On the other hand, if $|I^\tau_o|>1$, the oracle coordinates must have a relatively significant contribution in the following sense. For any $i\in I^\tau_o$, let $I_{o,-i}=I^\tau_o\setminus\{i\}$, we have by the definition of the τ-oracle that $r^2_\tau(I^\tau_{o,-i},\theta)\ge r^2_\tau(I^\tau_o,\theta)$, yielding

$$\begin{aligned}\big\|(\mathrm{H}_{I^\tau_o}-\mathrm{H}_{I^\tau_{o,-i}})\mathrm{X}\theta\big\|^2&\ge\tau\sigma^2|I^\tau_o|\log\frac{ep}{|I^\tau_o|}-\tau\sigma^2\big(|I^\tau_o|-1\big)\log\frac{ep}{|I^\tau_o|-1}\\&=\tau\sigma^2\bigg(\log(ep)+\log\frac{(|I^\tau_o|-1)^{|I^\tau_o|-1}}{|I^\tau_o|^{|I^\tau_o|}}\bigg)\\&\ge\tau\sigma^2\big(\log(ep)-\log\big[\sqrt{2}\big(|I^\tau_o|-1\big)\big]\big).\end{aligned}$$

If we compare the EBR parameters $\Theta_{\mathrm{eb}}(t,\tau)$ with the traditional sparsity class $\ell_0[s]=\{\theta\in\mathbb{R}^p:|I^*(\theta)|\le s\}$, it is easy to see that a subset of $\ell_0[s]$ with prominent nonzero components $\bar\ell_0(s,\tau_1)=\{\theta\in\ell_0[s]:I^*(\theta)=I^{\tau_1}_o(\theta)\}$ trivially satisfies EBR for $\tau=\tau_1$ and any $t\ge0$. For example, if $\theta\in\ell_0[s]$ is such that $\|(\mathrm{H}_J-\mathrm{H}_{J_{-i}})\mathrm{X}\theta_0\|^2\ge\tau_1\sigma^2\log(ep)$ for all $i\in J$ for some $J\in\mathcal{I}$ with $|J|=s>1$, then $\theta_0\in\bar\ell_0(s,\tau_1)\subseteq\Theta_{\mathrm{eb}}(t,\tau_1)$ for any $t\ge0$.

In this light, the EBR condition can be thought of as a more lenient sparsity on a parameter value θ_0, (to be called the *EBR-sparsity*), compared with the classical "conservative" sparsity requirement $|I^*(\theta_0)|\le s$: the error the oracle makes when setting certain "insignificant" coordinates of θ to zero should not exceed (up to a constant multiple) the error made in recovering the contribution by "significant" nonzero coordinates. The parameter $t\ge0$ measures the EBR-sparsity level, which is weaker for larger t, and becomes no condition as $t\to\infty$. Thus for any fixed $\tau>0$, a new scale, to be called the *EBR scale*, "slices" $\mathbb{R}^p$ as

$\mathbb{R}^p = \bigcup_{t\geq 0} \Theta_{\mathrm{eb}}(t, \tau)$ by varying the other structural parameter t over the positive half-line. Therefore, in view of Theorem 2, at any true parameter θ_0, an optimal-order inflated credible ball has high coverage provided that the constant inflation factor is chosen sufficiently large. This provides a new perspective at the deceptiveness issue mentioned in the Introduction: each parameter value θ_0 is deceptive (or nondeceptive) to some extent as $\theta_0 \in \Theta_{\mathrm{eb}}(t, \tau)$ for some $t \geq 0$ depending on θ_0. It is the structural parameter t of the new EBR-scale that measures the amount of deceptiveness, which in turn determines the effective amount of inflation of the confidence ball needed to provide high coverage.

The EBR condition appears naturally as a bias-variability comparison. Besides giving a sort of minimal condition to suppress the bias to ensure coverage, it is conceptually easier to envision in more general settings. In contrast, the concepts of self-similarity or polished tail are based on diminishing tails in the infinitely many normal means problem, where the bias is handled by controlling the tail of the sequence. The definitions of self-similar or polished tail sequences are presently unclear for sparse sequences as there is no clear conception of a tail in this setting.

Relation to the literature. Although our main goal in this paper is the uncertainty quantification problem in the linear regression model, we do obtain en route posterior contraction, estimation and some more related variable selection results, some of which constitute the necessary ingredients for the main problem of uncertainty quantification. There is a substantial number of results on estimation and variable selection (and posterior contraction) in the Bayesian and frequentist literature for the linear regression, especially in the modern high-dimensional setting with a sparsity structure. Our Theorem 1, Corollaries 1, 2 and 3 are the results of that type (although the weak support recovery result in claim (i) of Corollary 1 seems to be new). We, however, emphasize the main distinctive feature of our approach as compared with the existing literature on the topic: our results are all local and uniform over the entire space. That is, we do not a priori assume any traditional sparsity structure. We rather establish that our method exploits as much intrinsic sparsity, measured by the oracle rate $r(\theta_0)$, present in the underlying θ_0 (and X).

As to the main goal of our study, within our familiarity, the only paper addressing the issue of confidence ball of the whole regression coefficient for sparse high-dimensional linear regression is Nickl and van de Geer [24], who constructed such a set centered at a sparse estimator with radius obtained from an estimate of risk using completely non-Bayesian methods. They showed that it is impossible to adapt to the optimal size for different sparsity without paying in terms of coverage at some parameter values. Then they adapt their confidence region for only two possible sparsity levels at a time, by imposing a restriction on a parameter value in the larger class (i.e., less sparse) to maintain a distance of some appropriate order from the smaller class (i.e., more sparse). We, on the contrary, consider all possible sparsity levels simultaneously and obtain adaptive size and coverage after removing the set of deceptive parameters by the excessive bias restriction condition.

Oracle approach. It is to be noted that the definition of the oracle is dependent on the choice of the class of estimators. In particular, we defined the oracle risk for a given model and parameter value through the risk of the family of least square estimators restricted to the given model at the given parameter value. There are other sensible families of estimators for the problem such as a ridge regression estimator or the Lasso with some chosen tuning parameter, especially when the given model leads to multicollinearity. One advantage of the latter is that the requirement of linear independence of the columns of X_I is no longer required, and apparently, the risk bound may be lower. However, we note that the oracle approach needs only to consider the minimizer of the risk over the entire family, and hence if a lot of multicollinearity is present in X_I (thus making the least square estimator restricted to the model

I have higher risk), the oracle will be given by an appropriately smaller model, and hence the oracle risk will be sufficiently controlled. This intuitive reasoning is supported by the fact that the oracle approach based on the family of least square estimators is already sufficient for obtaining the optimal rates for sparsity classes. The use of the family of least square estimators to construct the oracle allows more explicit expressions, which are instrumental in the derivations.

The oracle approach is elegant since it automatically derives the minimax rates simultaneously for different possible scales through the local rates, but is clearly dependent on the availability of explicit expressions for the posterior density in each model. More specifically, a key step is to bound the posterior probabilities of models corresponding to different selection indexes by strong exponential inequalities as in (5.4). This restricts us essentially to a normal-normal conjugate setting, and hence the specific choice of the prior is important. This prior also makes the empirical Bayes choice for the mean, correcting the overshrinkage, necessary. With the availability of explicit expressions, the construction of credible sets with guaranteed frequentist coverage becomes more manageable. The conjugacy also makes the computations more manageable through Monte Carlo sampling, as explained in the Supplementary Material [3]. Formulating an analogous computational procedure for other priors, such as for the Laplace prior used by Castillo et al. [11], may be difficult to implement since the model posterior probabilities cannot be obtained explicitly due to the lack of conjugacy, although other algorithms such as sequential Monte Carlo may work; see Castillo et al. [11]. Therefore, it appears that in the linear Gaussian setting, the proposed empirical Bayes procedure is very fruitful. While every problem needs to be considered separately to address their unique features, a general path to the proof may be envisioned from the experiences in Belitser [2], Belitser and Nurushev [4] and the present paper. Indeed analogous results may be immediately extended to the problem of linear regression with grouped predictors. These predictors can enter the model only as groups, and the true set of predictors is also assumed to respect the grouping structure. We assume that for each group of predictors G, the submatrix X_G formed by the predictors in G has linearly independent columns. Then in our setting, the grouped predictor problem can be simply interpreted as imposing a restriction on $\mathcal{I}$, the family of possible sets of active predictors in the model. Thus the reduced cardinality of $\mathcal{I}$ leads to reduced complexity and hence $(\lambda_I : I \in \mathcal{I})$ continues to sum to a number bounded independently of p and n. This shows that the key step (5.8) remains valid and hence all the bounds derived for ungrouped predictors will remain in force. An extension to additive nonparametric regression is described in the next section.

4. Additive nonparametric regression. Nonparametric regression in high dimensions suffers from the curse of dimensionality problem because the convergence rate significantly slows down with the dimension. Additive nonparametric regression avoids the curse of dimensionality, but still retains the flexibility of nonparametric models, and is widely used; see Hastie and Tibshirani [18]. An additive nonparametric regression model for a response variable Y_k based on p-dimensional random predictors $X^{(k)} = (X_1^{(k)}, \dots, X_p^{(k)})' \in \mathbb{R}^p$, $k = 1, \dots, n$, is given by the relation

$$Y_k = \mu + \sum_{i=1}^{p} f_i\big(X_i^{(k)}\big) + \varepsilon_k, \quad k = 1, \dots, n, \tag{4.1}$$

where μ is a real-valued parameter, $f_1, \dots, f_p$ are functions with some smoothness properties and satisfy the identifiability restrictions $\mathrm{E} f_i(X_i^{(k)}) = 0$, $i = 1, \dots, p$, and $\varepsilon_k \overset{\text{ind}}{\sim} \mathrm{N}(0, \sigma^2)$, $k = 1, \dots, n$. Common approaches to estimating the regression function use kernel smoothing, penalization or expanding in a convenient basis. When p is large, a variable selection step

needs to be incorporated in the inference, which is typically performed through a penalization approach [Lin and Zhang [21], Ravikumar et al. [27], Meier et al. [23], Raskutti et al. [26]] or by Bayesian variable selection [Curtis et al. [14], Yang and Tokdar [34]]. Representing each component function through basis expansions, the additive nonparametric regression model (4.1) can be reduced to a linear regression model after truncating the basis expansion and allowing for approximation bias. Below we present an argument to show that the results of Section 2 can be adapted in this setting to quantify posterior concentration rates in terms of the oracle rates and derive coverage of the resulting credible sets.

Let $(X^{(k)}, Y^{(k)}), k = 1, \ldots, n$, be the observed sample, $Y = (Y_1, \ldots, Y_n)'$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ and $\mathrm{X} = (X^{(1)}, \ldots, X^{(n)})'$. Denote for brevity $\mathcal{J} = \{(i, j) : i \in [p], j \in \mathbb{N}\}$, where $[k] = \{1, \ldots, k\}$ for $k \in \mathbb{N}$. Assume that the ith predictor $X_i^{(k)}$ is distributed according to a density q_i on its domain D_i and different predictors are independently distributed. For each $i \in [p]$, let $\{1, B_{ij} : j \in \mathbb{N}\}$ be an orthonormal basis of $L_2(D_i, q_i)$. This, in particular, means that $\int B_{ij} q_i = 0$ for all $i \in [p]$ and $j \in \mathbb{N}$. Let $f_i \in L_2(D_i, q_i)$ be such that $\int f_i q_i = 0$ for all $i \in [p]$. Denoting $Z^{(k)} = (1, (B_{ij}(X_i^{(k)}) : (i, j) \in \mathcal{J}))'$, $\mathrm{Z} = \mathrm{Z}(\mathrm{X}) = (Z^{(1)}, \ldots, Z^{(n)})'$ and $\theta = (\mu, (\theta_{ij}, (i, j) \in \mathcal{J}))'$, we can rewrite the model (4.1) as $Y = \mathrm{Z}\theta + \varepsilon$. Now, for every $I \subseteq [p]$ and $J \in \mathbb{N}$, consider the (I, J)-truncated linear model

$$Y = \mathrm{Z}_{I,J}\theta_{I,J} + \varepsilon, \tag{4.2}$$

where $\mathrm{Z}_{I,J}$ is the $(n \times (|I|J + 1))$-matrix with the kth row equal to $Z_{I,J}^{(k)} = (1, (B_{ij}(X_i^{(k)}) : i \in I, j \in [J]))'$ and $\theta_{I,J} = (\mu, (\theta_{ij} : i \in I, j \in [J]))'$. Observe that in the (I, J)-th model (4.2), the least square estimator $\hat{\theta}(I, J) = (\hat{\mu}, (\hat{\theta}_{ij}(I, J), (i, j) \in \mathcal{J}))'$ of θ is

$$\hat{\theta}_{I,J}(I, J) = \big(\mathrm{Z}_{I,J}'\mathrm{Z}_{I,J}\big)^{-1}\mathrm{Z}_{I,J}'Y, \qquad \hat{\theta}_{i,j}(I, J) = 0, \quad (i, j) \notin \big(I, [J]\big),$$

provided that $(I, J) \in \mathcal{I} \triangleq \{(I', J') : \text{the columns of } \mathrm{Z}_{I',J'} \text{ are linearly independent}\}$. In other words, this means that the estimator has the (i, j)-th component set to zero if either $i \notin I$ or $j > J$, and the least square estimator is applied to the surviving coefficients. Then, as in (2.2), the quadratic prediction risk of $\mathrm{Z}\hat{\theta}(I, J)$ is given by

$$R^2\big((I, J), \theta\big) = \theta'\mathrm{E}\big[\mathrm{Z}'(\mathrm{I} - \mathrm{H}_{I,J})\mathrm{Z}\big]\theta + \sigma^2\big(|I|J + 1\big),$$

where $\mathrm{H}_{I,J} = \mathrm{Z}_{I,J}(\mathrm{Z}_{I,J}'\mathrm{Z}_{I,J})^{-1}\mathrm{Z}_{I,J}'$ is the projection on the column space of $\mathrm{Z}_{I,J}$. The R-oracle is the minimizer of this risk over all $(I, J) \in \mathcal{I}$. The objective is then to mimic the risk of the oracle based estimator without knowing it. As in linear regression, it is impossible to match the risk of the R-oracle based estimator by any other estimator. Following (2.3), we need to include an additional term to define the τ-oracle rate $r_\tau((I, J), \theta)$ by

$$r_\tau^2\big((I, J), \theta\big) = \theta'\mathrm{E}\big[\mathrm{Z}'(\mathrm{I} - \mathrm{H}_{I,J})\mathrm{Z}\big]\theta + \tau\sigma^2\big(|I|J + 1\big) + \tau\sigma^2|I|\log\frac{ep}{|I|}, \tag{4.3}$$

and define the τ-oracle $(I_o^\tau(\theta), J_o^\tau(\theta))$ to be the minimizer of (4.3). We note that a term with the additional logarithmic factor $\log(ep/|I|)$ is needed only for the first index I, since the values of J are linearly ordered, unlike I, which requires considering all possible subsets of $\{1, \ldots, p\}$. The role of the index J is similar to the truncating parameter in Belitser [2]. As before, we define the oracle rate by $r(\theta) = r_1((I_o^1(\theta), J_o^1(\theta)), \theta)$.

We consider the conjugate prior: given (I, J),

$$\theta_{I,J}|(I, J) \sim \mathrm{N}\big(\mu(I, J), \kappa\sigma^2\big(\mathrm{Z}_{I,J}'\mathrm{Z}_{I,J}\big)^{-1}\big), \qquad \theta_{ij} = 0 \quad \text{if } i \notin I \text{ or } j > J.$$

It may be noted that the functions $\sum_{j=1}^J \theta_{ij}B_{ij}$ used in modeling f_i are automatically centered with respect to q_i for all for any choice of coefficients $\{\theta_{ij}\}$ and $i \in [p]$, since the basis

functions are centered. On I, we put the prior given by (2.5), and independently we put a geometric prior $\lambda(J=j)=(e^{\beta}-1)e^{-\beta j}$ on J, $j\in\mathbb{N}$, $\beta>0$. The resulting joint prior for (I,J) will be denoted by $\lambda_{I,J}$. This leads to the posteriors $\pi_{\mu}(\theta|\mathrm{X},Y,(I,J))=\pi_{(I,J),\mu}(\theta|\mathrm{X},Y)$ and $\pi_{\mu}((I,J)|\mathrm{X},Y)$. As before, for each (I,J), $\mu(I,J)$ is selected by the empirical Bayes method: $\hat{\mu}(I,J)=\hat{\theta}_{I,J}(I,J)$, leading to the empirical Bayes posteriors $\pi_{(I,J),\hat{\mu}}(\theta|\mathrm{X},Y)$ and $\pi_{\hat{\mu}}((I,J)|\mathrm{X},Y)$. Now we can again determine the maximizer $(\hat{I},\hat{J})$ of the marginal empirical Bayes posterior probability $\pi_{\hat{\mu}}((I,J)|\mathrm{X},Y)$, and construct two versions of empirical Bayes posteriors: the EBMS posterior $\breve{\pi}(\theta|\mathrm{X},Y)$ with the corresponding EBMS mean $\breve{\theta}$, and the EBMA posterior $\tilde{\pi}(\theta|\mathrm{X},Y)$ with the corresponding EBMA mean $\tilde{\theta}$. Let, as before, $\hat{\pi}$ denote either $\breve{\pi}$ or $\tilde{\pi}$ and, respectively, $\hat{\theta}$ denote either $\breve{\theta}$ or $\tilde{\theta}$.

Denote $\theta_0=(\mu_0,(\theta_{ij,0},(i,j)\in\mathcal{J}))$ and $f_{i,0}(x)=\sum_{j=1}^{\infty}\theta_{ij,0}B_{ij}(x)$, $i=1,\dots,p$. We shall write P_{θ_0} and E_{θ_0}, respectively, for the probability and the expectation under the distribution induced by the parameter value θ_0. The empirical Bayes posterior $\hat{\pi}$ and the estimator $\hat{\theta}$ then satisfy concentration results similar to Theorem 1. A proof is provided in the Supplementary Material [3].

THEOREM 4. *There exists a constant $C>0$ such that, for any $\mu_0\in\mathbb{R}$ and $f_{i,0}\in L_2(D_i)$ with $\int_{D_i}f_{i,0}(x_i)q_i(x_i)\,dx_i=0$, $i=1,\dots,p$, and any $M>0$, we have that*

$$\mathrm{E}_{\theta_0}\hat{\pi}\big(\|\mathrm{Z}\theta-\mathrm{Z}\theta_0\|\ge Mr(\theta_0)|\mathrm{X},Y\big)\le CM^{-2},\qquad \mathrm{E}_{\theta_0}\|\mathrm{Z}\hat{\theta}-\mathrm{Z}\theta_0\|^2\le Cr^2(\theta_0). \tag{4.4}$$

Here, the true expectation integrates out the X-values too as they are random; see Remark 3. Note that, when modeling additive nonparametric regression by linear regression, we are only interested in the prediction problem, that is, an inference on $\mathrm{Z}\theta$, rather than on θ.

In addition, as a consequence of the oracle approach, we obtain that a credible ball with squared radius a large multiple of $\hat{\rho}^2=1+|\hat{I}|\hat{J}+|\hat{I}|\log(ep/|\hat{I}|)$ has frequentist coverage at all true parameters which meet the ϵ-EBR condition for a sufficiently small $\epsilon>0$, and the size of the credible set is minimax optimal with high probability. Precisely, in this case, the normalized τ-bias at θ_0 is defined as follows:

$$b_{\tau}(\theta_0)=b_{\tau}(\theta_0|\mathrm{X})=\frac{\theta_0'\mathrm{Z}'(\mathrm{I}-\mathrm{H}_{I_o^{\tau},J_o^{\tau}})\mathrm{Z}\theta_0}{\sigma^2(|I_o^{\tau}|J_o^{\tau}+1)+\tau\sigma^2|I_o^{\tau}|\log\frac{ep}{|I_o^{\tau}|}}.$$

For a structural parameter (t,τ) and $\epsilon>0$, recall that the ϵ-EBR condition holds at θ_0 if the normalized τ-bias $b_{\tau}(\theta_0)=b_{\tau}(\theta_0|\mathrm{X})$ satisfies $\mathrm{P}_{\mathrm{X}}(b_{\tau}(\theta_0|\mathrm{X})>t)<\epsilon$. Let the set of all parameter values satisfying the ϵ-EBR condition with structural parameters (t,τ) be denoted by $\Theta_{\epsilon\text{-eb}}(t,\tau)$. The following theorem claims the uncertainty quantification results for the additive nonparametric regression model under the ϵ-EBR condition. The proof is given in the Supplementary Material [3].

THEOREM 5. *Let τ satisfy (2.17), $t>0$, $\epsilon_1,\epsilon_2>0$. Then there exist $M=M(t,\epsilon_1)>0$ and $L=L(\epsilon_2)>0$ such that*

$$\sup_{\theta_0\in\Theta_{\epsilon\text{-eb}}(t,\tau)}\mathrm{P}_{\theta_0}\big(\mathrm{Z}\theta_0\notin B(\mathrm{Z}\hat{\theta},M\hat{\rho})\big)\le\epsilon+\epsilon_1,\qquad \sup_{\theta_0\in\mathbb{R}^p}\mathrm{P}_{\theta_0}\big(\hat{\rho}\ge Lr(\theta_0)\big)\le\epsilon_2.$$

The above results, Theorem 4 and 5, are admittedly abstract, but they can easily lead to well-interpretable results when sparsity conditions for selection of the variables $X_1,\dots,X_p$ in the true regression function and smoothness of each selected function are assumed. This would deliver adaptive minimax results over an appropriate functional scale. For this, we only need to upper bound the oracle risk by a multiple of the corresponding minimax rate under these assumptions.

Let each component X_i of the predictor variable X be distributed uniformly on the unit interval $[0,1]$ independently of other components. Fix a convenient orthonormal basis $\{1, B_1, B_2, \ldots\}$ in $L_2[0,1]$ (such as that of Legendre polynomials or trigonometric functions). Then an additive regression function $\mu + \sum_{i=1}^{p} f_i(X_i)$ with $f_i \in L_2[0,1]$, $\int f_i = 0$, can be equivalently represented by the infinite dimensional vector $\theta = (\mu, (\theta_j(f_i), (i,j) \in \mathcal{J}))$, where for $f \in L_2[0,1]$, $\theta_j(f) = \int_0^1 B_j(x) f(x)\,dx$. Observe that the property $\int f_i = 0$ for all $i \in [p]$ is needed for the expansion since the basis functions are centered. Writing the additive nonparametric model as an infinite dimensional linear model $Y = \mathrm{Z}\theta + \varepsilon$ using these basis expansions and letting $F_0(\mathrm{X}) = \mathrm{Z}\theta_0$ stand for the true regression function and θ_0 be the corresponding true coefficient vector, we can quantify posterior contraction and coverage of credible sets in terms of the metric $\|x\|_n = n^{-1/2}\|x\|$, $x \in \mathbb{R}^n$, and the ball $B_n(x,r) = \{y \in \mathbb{R}^n : \|x - y\|_n \le r\}$, $r > 0$.

Introduce the tail class $\mathcal{T}(\alpha) = \{\theta \in \ell_2 : \sum_{j>J} \theta_j^2 \le Q J^{-2\alpha}, J \in \mathbb{N}\}$ which contains the Sobolev class $\mathcal{S}(\alpha) = \{\theta \in \ell_2 : \sum_{j \in \mathbb{N}} j^{2\alpha}\theta_j^2 \le Q\}$. Then for the additive regression function, we define the functional classes

$$\mathcal{F}(s,\alpha) = \Big\{ \mu + \sum_{i \in I} f_i(x_i) : \mu \in \mathbb{R}, I \subseteq [p], 1 \le |I| \le s, \int_0^1 f_i(x)\,dx = 0, \theta(f_i) \in \mathcal{T}(\alpha), i \in [p] \Big\}$$

for $s \in [p]$ and $\alpha > 0$. Note that for inactive predictors, the regression functions are zero functions, which trivially satisfy the smoothness assumption.

COROLLARY 5. *There exist a constant $C > 0$ such that*

$$\sup_{F_0 \in \mathcal{F}(s,\alpha)} \mathrm{E}_{F_0} \hat{\pi}\big(\|\mathrm{Z}\theta - F_0(\mathrm{X})\|_n \ge M\epsilon_n | \mathrm{X}, Y\big) \le C M^{-2},$$

$$\sup_{F_0 \in \mathcal{F}(s,\alpha)} \mathrm{E}_{F_0} \|\mathrm{Z}\hat{\theta} - F_0(\mathrm{X})\|_n^2 \le C\epsilon_n^2,$$

where $\epsilon_n^2 = \max\{s n^{-2\alpha/(1+2\alpha)}, s n^{-1}\log(p/s)\}$.

Moreover, let τ satisfy (2.17), $t > 0$, $\epsilon_1, \epsilon_2 > 0$, $\hat{\rho}^2 = n^{-1}\{1 + |\hat{I}|\hat{J} + |\hat{I}|\log(ep/|\hat{I}|)\}$, where $(\hat{I}, \hat{J})$ is the EBMS selector. Then there exist $M = M(t,\epsilon_1) > 0$ and $L = L(\epsilon_2) > 0$ such that

$$\sup_{\theta_{F_0} \in \Theta_{\epsilon\text{-eb}}(t,\tau)} \mathrm{P}_{F_0}\big(F(\mathrm{X}) \notin B_n(\mathrm{Z}\hat{\theta}, M\hat{\rho})\big) \le \epsilon + \epsilon_1, \qquad \sup_{F_0 \in \mathcal{F}(s,\alpha)} \mathrm{P}_{F_0}(\hat{\rho} \ge L\epsilon_n) \le \epsilon_2.$$

A proof of this corollary is provided in the Supplementary Material [3].

The obtained rate ϵ_n was shown to be the minimax rate for estimation in Raskutti et al. [26] and Yang and Tokdar [34]. Thus the empirical Bayes procedure achieves the optimal minimax rate of estimation in the stated regime. Yang and Tokdar [34] also showed, using the general theory of posterior contraction (see Ghosal and van der Vaart [17]), that the posterior based on a Gaussian process prior contracts at the optimal minimax rate up to a logarithmic factor. By using the oracle approach, we obtain the same result without the additional logarithmic factor for the empirical Bayes posterior based on the orthogonal series prior. The same rate for the convergence of the empirical Bayes posterior mean is automatically obtained. The obtained coverage result for credible sets is new.

5. Proofs. In this section, we present the proofs of the results stated in Section 2.

The following is a version of a maximal inequality under a bounded exponential moment assumption and will be used in the proof of Theorem 1.

LEMMA 1. *Let $t > 0$ and $\eta_1, \dots, \eta_N$ be random variables such that $\mathrm{E}e^{t\eta_i} \le A_t$ for some $0 < A_t < \infty$ and all $i = 1, \dots, N$. Then $\mathrm{E}(\max\{\eta_i : 1 \le i \le N\}) \le t^{-1}(\log N + \log A_t)$.*

PROOF. By Jensen's inequality, $\exp\{t\mathrm{E}\max_{1\le i\le N}\eta_i\} \le \mathrm{E}\exp\{t\max_{1\le i\le N}\eta_i\} \le \sum_{i=1}^N \mathrm{E}e^{t\eta_i} \le NA_t$, which is equivalent with the assertion. □

PROOF OF THEOREM 1. Recall our notational convention for all $\mathcal{G} \subseteq \mathcal{I}$: when $\hat\pi = \check\pi$, $\hat\pi(I \in \mathcal{G}|Y) = \mathbb{1}\{\hat I \in \mathcal{G}\}$ (in particular, $\hat\pi(I|Y) = \mathbb{1}\{\hat I = I\}$) and $\mathrm{E}_{\theta_0}\hat\pi(I \in \mathcal{G}|Y) = \mathrm{P}_{\theta_0}(\hat I \in \mathcal{G})$; while if $\hat\pi = \tilde\pi$, $\hat\pi(I \in \mathcal{G}|Y) = \tilde\pi(I \in \mathcal{G}|Y)$ (in particular, $\hat\pi(I|Y) = \tilde\pi(I|Y)$) and $\mathrm{E}_{\theta_0}\hat\pi(I \in \mathcal{G}|Y) = \mathrm{E}_{\theta_0}\tilde\pi(I \in \mathcal{G}|Y)$.

According to the empirical Bayes posterior distribution $\hat\pi(\cdot|Y)$ given in (2.9), it follows that $\mathrm{X}\theta|(Y, I) \sim \mathrm{N}_n(\mathrm{H}_I Y, \frac{\kappa\sigma^2}{\kappa+1}\mathrm{H}_I)$. Then as $\mathrm{H}_I(\mathrm{I} - \mathrm{H}_I) = \mathrm{O}$ for any I, we have that

$$\begin{aligned}
\hat{\mathrm{E}}\big(\|\mathrm{X}\theta - \mathrm{X}\theta_0\|^2|Y\big) &= \sum_{I\in\mathcal{I}}\bigg(\frac{\kappa\sigma^2}{\kappa+1}\operatorname{tr}(\mathrm{H}_I) + \|\mathrm{H}_I Y - \mathrm{X}\theta_0\|^2\bigg)\hat\pi(I|Y) \\
&= \sum_{I\in\mathcal{I}}\bigg(\frac{\kappa\sigma^2}{\kappa+1}|I| + \big\|(\mathrm{H}_I - \mathrm{I})\mathrm{X}_{I^c}\theta_{0,I^c} + \sigma\mathrm{H}_I\varepsilon\big\|^2\bigg)\hat\pi(I|Y) \\
&\le \sum_{I\in\mathcal{I}}\big(r^2(I,\theta_0) + \sigma^2\|\mathrm{H}_I\varepsilon\|^2\big)\hat\pi(I|Y).
\end{aligned}$$

Therefore, by Markov's inequality we obtain

$$\begin{aligned}
&\mathrm{E}_{\theta_0}\hat\pi\big(\|\mathrm{X}\theta - \mathrm{X}\theta_0\| \ge Mr(\theta_0)|Y\big) \\
&\quad= \mathrm{E}_{\theta_0}\sum_{I\in\mathcal{I}}\hat\pi_I\big(\|\mathrm{X}\theta - \mathrm{X}\theta_0\| \ge Mr(\theta_0)|Y\big)\hat\pi(I|Y) \\
&\quad\le \mathrm{E}_{\theta_0}\sum_{I\in\mathcal{I}}\frac{\hat{\mathrm{E}}(\|\mathrm{X}\theta - \mathrm{X}\theta_0\|^2|Y, I)}{M^2r^2(\theta_0)}\hat\pi(I|Y) \\
&\quad\le \frac{\sum_{I\in\mathcal{I}} r^2(I,\theta_0)\mathrm{E}_{\theta_0}\hat\pi(I|Y)}{M^2r^2(\theta_0)} + \frac{\sigma^2\mathrm{E}_{\theta_0}\sum_{I\in\mathcal{I}}\|\mathrm{H}_I\varepsilon\|^2\hat\pi(I|Y)}{M^2r^2(\theta_0)}.
\end{aligned} \tag{5.1}$$

Henceforth, we consider two cases $\hat\pi = \check\pi$ and $\hat\pi = \tilde\pi$ separately. If $\hat\pi = \check\pi$, by using (2.10) and (2.7), we obtain that, for any $h \in [0, 1]$ and any $I, I_0 \in \mathcal{I}$,

$$\mathrm{E}_{\theta_0}\hat\pi(I|Y) = \mathrm{P}_{\theta_0}(\hat I = I) \le \mathrm{P}_{\theta_0}\bigg(\frac{\lambda_I\hat\pi_I(Y)}{\lambda_{I_0}\hat\pi_{I_0}(Y)} \ge 1\bigg) \le \mathrm{E}_{\theta_0}\bigg[\frac{\lambda_I\hat\pi_I(Y)}{\lambda_{I_0}\hat\pi_{I_0}(Y)}\bigg]^h.$$

When $\hat\pi = \tilde\pi$, by the definition of $\tilde\pi(I|Y)$ in (2.12), we derive the same as follows: for any $h \in [0, 1]$ and any $I, I_0 \in \mathcal{I}$,

$$\mathrm{E}_{\theta_0}\hat\pi(I|Y) = \mathrm{E}_{\theta_0}\tilde\pi(I|Y) \le \mathrm{E}_{\theta_0}\bigg[\frac{\lambda_I\hat\pi_I(Y)}{\sum_{J\in\mathcal{I}}\lambda_J\hat\pi_J(Y)}\bigg]^h \le \mathrm{E}_{\theta_0}\bigg[\frac{\lambda_I\hat\pi_I(Y)}{\lambda_{I_0}\hat\pi_{I_0}(Y)}\bigg]^h.$$

In either case, we derive the bound

$$\mathrm{E}_{\theta_0}\hat\pi(I|Y) \le \mathrm{E}_{\theta_0}\bigg[\frac{\lambda_I\hat\pi_I(Y)}{\lambda_{I_0}\hat\pi_{I_0}(Y)}\bigg]^h = \bigg(\frac{\lambda_I}{\lambda_{I_0}}\bigg)^h\frac{\mathrm{E}_{\theta_0}\exp\{\frac{h}{2\sigma^2}Y'(\mathrm{H}_I - \mathrm{H}_{I_0})Y\}}{(1+\kappa)^{h(|I|-|I_0|)/2}}. \tag{5.2}$$

Recall the fact that if $Y \sim \mathrm{N}(\mu, \Sigma)$ and $\mathrm{A}\Sigma < \mathrm{I}$ for a symmetric matrix A, then

$$\mathrm{E}\exp[Y'\mathrm{A}Y/2] = \frac{\exp[\frac{1}{2}\mu'(\mathrm{I}-\mathrm{A}\Sigma)^{-1}\mathrm{A}\mu]}{\sqrt{\det(\mathrm{I}-\mathrm{A}\Sigma)}}. \tag{5.3}$$

Denote for brevity $\mathrm{H}_{I,I_0} = \mathrm{H}_I - \mathrm{H}_{I_0}$. Using (5.2) and (5.3) with $\Sigma = \sigma^2\mathrm{I}$ and $\mathrm{A} = \frac{h}{\sigma^2}\mathrm{H}_{I,I_0}$,

$$\mathrm{E}_{\theta_0}\hat{\pi}(I|Y) \le \left(\frac{\lambda_I}{\lambda_{I_0}}\right)^h \frac{\exp\{\frac{h}{2\sigma^2}\theta_0'\mathrm{X}'(\mathrm{I}-h\mathrm{H}_{I,I_0})^{-1}\mathrm{H}_{I,I_0}\mathrm{X}\theta_0\}}{(1+\kappa)^{h(|I|-|I_0|)/2}\sqrt{\det(\mathrm{I}-h\mathrm{H}_{I,I_0})}}, \tag{5.4}$$

provided that $\mathrm{I} - h\mathrm{H}_{I,I_0}$ is invertible.

Next, for any $h \in [0, 1]$ and any symmetric matrix A such that $\mathrm{A} \le \mathrm{I}$ and $\mathrm{I} - h\mathrm{A}$ is invertible, we have that $(\mathrm{I}-h\mathrm{A})^{-1}\mathrm{A} = \mathrm{A}(\mathrm{I}-h\mathrm{A})^{-1} = \mathrm{A} + h\mathrm{A}(\mathrm{I}-h\mathrm{A})^{-1}\mathrm{A}$. Thus

$$(\mathrm{I}-h\mathrm{H}_{I,I_0})^{-1}\mathrm{H}_{I,I_0} = \mathrm{H}_{I,I_0} + h\mathrm{H}_{I,I_0}(\mathrm{I}-h\mathrm{H}_{I,I_0})^{-1}\mathrm{H}_{I,I_0}, \quad h \in [0,1),$$

as $\mathrm{I}-h\mathrm{H}_{I,I_0}$ is invertible for any $h \in [0, 1)$. For brevity, denote $y = \mathrm{X}\theta_0$. Observe that $\frac{1}{1-h}(\mathrm{I}-h\mathrm{H}_{I,I_0}) = \mathrm{I} + \frac{h}{1-h}(\mathrm{I}-\mathrm{H}_{I,I_0}) \ge \mathrm{I}$, so that $(\mathrm{I}-h\mathrm{H}_{I,I_0})^{-1} \le \frac{1}{1-h}\mathrm{I}$. Hence

$$\begin{aligned}
&y'(\mathrm{I}-h\mathrm{H}_{I,I_0})^{-1}\mathrm{H}_{I,I_0}y \\
&\quad = y'\mathrm{H}_{I,I_0}y + hy'\mathrm{H}_{I,I_0}(\mathrm{I}-h\mathrm{H}_{I,I_0})^{-1}\mathrm{H}_{I,I_0}y \\
&\quad \le \|(\mathrm{I}-\mathrm{H}_{I_0})y\|^2 - \|(\mathrm{I}-\mathrm{H}_I)y\|^2 + \frac{h}{1-h}\|\mathrm{H}_{I,I_0}y\|^2 \\
&\quad \le \|(\mathrm{I}-\mathrm{H}_{I_0})y\|^2 - \|(\mathrm{I}-\mathrm{H}_I)y\|^2 + \frac{2h}{1-h}\big(\|(\mathrm{I}-\mathrm{H}_I)y\|^2 + \|(\mathrm{I}-\mathrm{H}_{I_0})y\|^2\big) \\
&\quad = \frac{1+h}{1-h}\|(\mathrm{I}-\mathrm{H}_{I_0})y\|^2 - \frac{1-3h}{1-h}\|(\mathrm{I}-\mathrm{H}_I)y\|^2,
\end{aligned} \tag{5.5}$$

for any $h \in [0, 1)$. Besides, we have that for any $h \in [0, 1)$

$$\det\big(\mathrm{I}-h(\mathrm{H}_I-\mathrm{H}_{I_0})\big) \ge (1-h)^{|I|}. \tag{5.6}$$

Inserting (2.5), (5.5) and (5.6) in (5.4), the bound for $\mathrm{E}_{\theta_0}\hat{\pi}(I)$ reduces to

$$\begin{aligned}
&\big(ep/|I|\big)^{-c_1|I|}\exp\bigg\{\frac{h(1+h)}{2\sigma^2(1-h)}\theta_0'\mathrm{X}'(\mathrm{I}-\mathrm{H}_{I_0})\mathrm{X}\theta_0 + \varkappa h|I_0|\log\frac{ep}{|I_0|} + \frac{h}{2}|I_0|\log(1+\kappa) \\
&\quad - \frac{h(1-3h)}{2\sigma^2(1-h)}\theta_0'\mathrm{X}'(\mathrm{I}-\mathrm{H}_I)\mathrm{X}\theta_0 - (h\varkappa - c_1)|I|\log\frac{ep}{|I|} - |I|\log\big[(1+\kappa)^{h/2}(1-h)^{1/2}\big]\bigg\},
\end{aligned}$$

where $c_1 = 1 + \frac{h\varkappa}{2}$. By (2.6), $(1+\kappa)^{h/2}(1-h)^{1/2} \ge 1$, $h \in (0, 1/3)$ and $\varkappa > 2/h$. Using this and applying the last relation with $I_0 = I_o(\theta_0)$, we obtain

$$\mathrm{E}_{\theta_0}\hat{\pi}(I|Y) \le \big(ep/|I|\big)^{-c_1|I|}\exp\big\{-c_2\sigma^{-2}\big(r^2(I,\theta_0) - c_3 r^2(\theta_0)\big)\big\}, \tag{5.7}$$

where

$$c_2 = \min\bigg\{\frac{h(1-3h)}{2(1-h)}, \frac{h\varkappa}{2} - 1\bigg\}, \qquad c_3 = \frac{h}{2c_2}\max\bigg\{\frac{1+h}{1-h}, 2\varkappa + \log(1+\kappa)\bigg\}.$$

By the assumed conditions (2.6) on κ, $\varkappa$ and h, we have $c_1 > 2$, $c_2 > 0$ and $c_3 > 0$.

For a $\tau_0 > 0$ and $\theta_0 \in \mathbb{R}^p$, let $\mathcal{O}(\tau_0, \theta_0) = \{I \in \mathcal{I} : r^2(I, \theta_0) \le \tau_0 r^2(\theta_0)\}$. Choosing $\tau_0 > c_3$, where c_3 is as in (5.7), it follows that for any $I \in \mathcal{O}^c(\tau_0, \theta_0)$, we have $r^2(I, \theta_0) - c_3 r^2(\theta_0) \ge$

$(1 - c_3/\tau_0)r^2(I,\theta_0)$. Using the estimates $xe^{-cx} \le (ce)^{-1}$ for any $x \ge 0$, $c > 0$ and $\binom{p}{k} \le (ep/k)^k$, with $B = c_2(\tau_0 - c_3)/(2\tau_0)$, we obtain from (5.7) that

$$\begin{aligned}\sum_{I\in\mathcal{O}^c(\tau_0,\theta_0)} r^2(I,\theta_0)\big[\mathrm{E}_{\theta_0}\hat\pi(I|Y)\big]^{1/2} &\le \sum_{I\in\mathcal{O}^c(\tau_0,\theta_0)} \left(\frac{ep}{|I|}\right)^{-c_1|I|/2} r^2(I,\theta_0)e^{-Br^2(I,\theta_0)/\sigma^2} \\ &\le \frac{\sigma^2}{Be}\sum_{k=1}^{p}\binom{p}{k}\left(\frac{ep}{k}\right)^{-c_1k/2} \le \frac{\sigma^2}{Be(e^{(c_1-2)/2}-1)}.\end{aligned} \tag{5.8}$$

As $\|\mathrm{H}_I\varepsilon\|^2 \sim \chi^2_{|I|}$, we have that $\mathrm{E}\|\mathrm{H}_I\varepsilon\|^4 = |I|^2 + 2|I| \le 3|I|^2 \le 3r^4(I,\theta_0)/\sigma^4$. Therefore, by the Cauchy–Schwarz inequality and (5.8), we obtain

$$\begin{aligned}\sigma^2\mathrm{E}_{\theta_0}\sum_{I\in\mathcal{O}^c(\tau_0,\theta_0)}\|\mathrm{H}_I\varepsilon\|^2\hat\pi(I|Y) &\le \sum_{I\in\mathcal{O}^c(\tau_0,\theta_0)}\sigma^2\big[\mathrm{E}\|\mathrm{H}_I\varepsilon\|^4\big]^{1/2}\big[\mathrm{E}_{\theta_0}\hat\pi(I|Y)\big]^{1/2} \\ &\le \sqrt{3}\sum_{I\in\mathcal{O}^c(\tau_0,\theta_0)} r^2(I,\theta_0)\big[\mathrm{E}_{\theta_0}\hat\pi(I|Y)\big]^{1/2} \\ &\le \frac{\sqrt{3}\sigma^2}{Be(e^{(c_1-2)/2}-1)}.\end{aligned} \tag{5.9}$$

To estimate the contributions from $I \in \mathcal{O}(\tau_0,\theta_0)$, note that the cardinality $|I|$ of I is necessarily bounded by $m = \max\{|J|, J \in \mathcal{O}(\tau_0,\theta_0)\}$ which satisfies

$$\sigma^2 m\log(ep/m) \le \tau_0 r^2(\theta_0), \tag{5.10}$$

by the definitions of $\mathcal{O}(\tau_0,\theta_0)$ and $r^2(\theta_0)$. In view of the norm decreasing properties of projection operators, we have that

$$\sum_{I\in\mathcal{O}(\tau_0,\theta_0)}\|\mathrm{H}_I\varepsilon\|^2\hat\pi(I|Y) \le \max_{I\in\mathcal{O}(\tau_0,\theta_0)}\|\mathrm{H}_I\varepsilon\|^2 = \max\big\{\|\mathrm{H}_I\varepsilon\|^2 : I\in\mathcal{O}(\tau_0,\theta_0), |I| = m\big\}.$$

Now $\|\mathrm{H}_I\varepsilon\|^2 \sim \chi^2_{|I|}$, and hence for any $I \in \mathcal{I}$ with $|I| = m$, $\mathrm{E}e^{t\|\mathrm{H}_I\varepsilon\|^2} = (1-2t)^{-m/2} \le e^m$ provided that $t \le (1 - e^{-2})/2 \approx 0.43$. As the cardinality of $\{I \in \mathcal{O}(\tau_0,\theta_0) : |I| = m\}$ is at most $\binom{p}{m} \le (ep/m)^m$, it follows from Lemma 1 with $t = 0.4$ that

$$\mathrm{E}\big(\max\big\{\|\mathrm{H}_I\varepsilon\|^2 : I\in\mathcal{O}(\tau_0,\theta_0), |I| = m\big\}\big) \le \frac{5}{2}\big(m\log(ep/m) + m\big) \le 5m\log(ep/m).$$

This leads to the bound

$$\sigma^2\mathrm{E}_{\theta_0}\sum_{I\in\mathcal{O}(\tau_0,\theta_0)}\|\mathrm{H}_I\varepsilon\|^2\hat\pi(I|Y) \le 5\sigma^2 m\log(ep/m) \le 5\tau_0 r^2(\theta_0).$$

Combining this with (5.9) yields a bound for the second term in (5.1):

$$\sigma^2\mathrm{E}_{\theta_0}\sum_{I\in\mathcal{I}}\|\mathrm{H}_I\varepsilon\|^2\hat\pi(I|Y) \le 5\tau_0 r^2(\theta_0) + \frac{\sqrt{3}\sigma^2}{Be(e^{(c_1-2)/2}-1)}. \tag{5.11}$$

For the first term in (5.1), we proceed similarly by splitting $\mathcal{I}$ in $\mathcal{O}(\tau_0,\theta_0)$ and its complement. For the sum over $\mathcal{O}(\tau_0,\theta_0)$, we use the bounds $r^2(I,\theta_0) \le \tau_0 r^2(\theta_0)$ and $\sum_{I\in\mathcal{I}}\hat\pi(I|Y) = 1$. For the sum over $\mathcal{O}^c(\tau_0,\theta_0)$, we import the bound in (5.8) by noting that $\mathrm{E}_{\theta_0}\hat\pi(I|Y) \le [\mathrm{E}_{\theta_0}\hat\pi(I|Y)]^{1/2}$. Using these relations, we obtain

$$\sum_{I\in\mathcal{I}} r^2(I,\theta_0)\mathrm{E}_{\theta_0}\hat\pi(I|Y) \le \tau_0 r^2(\theta_0) + \frac{\sigma^2}{Be(e^{(c_1-2)/2}-1)}. \tag{5.12}$$

Note that $\sigma^2 \le r^2(\theta_0)$ by the definition of the risk. Thus all terms in (5.11) and (5.12) can be bounded in terms of $r^2(\theta_0)$ and hence the bound in (5.1) reduces to C/M^2, where $C = 6\tau_0 + (\sqrt{3}+1)/(Be(e^{(c_1-2)/2}-1))$. This proves the first part of the theorem.

Observe that from (5.1) and the bounds (5.11) and (5.12), we get a stronger conclusion:

$$\mathrm{E}_{\theta_0}\hat{\mathrm{E}}\big(\|\mathrm{X}\theta - \mathrm{X}\theta_0\|^2|Y\big) = \mathrm{E}_{\theta_0}\sum_{I\in\mathcal{I}}\hat{\mathrm{E}}_I\|\mathrm{X}\theta - \mathrm{X}\theta_0\|^2\hat{\pi}(I|Y) \le Cr^2(\theta_0).$$

Now the second part of the theorem follows from Jensen's inequality. □

PROOF OF COROLLARY 1. Denote $\mathcal{G} = \{I : r^2(I,\theta_0) \ge m_0 r^2(\theta_0) + m\sigma^2\}$, for $m_0 = c_3$, where the constants $c_1 > 2$, $c_2 > 0$, $c_3 > 0$ are as in (5.7). Using (5.7) and the fact that $\sum_{I\in\mathcal{I}}(ep/|I|)^{-c_1|I|} \le (e^{c_1-1}-1)^{-1} \le 1$ for $c_1 > 2$, it follows that

$$\mathrm{E}_{\theta_0}\hat{\pi}(I\in\mathcal{G}|Y) = \sum_{I\in\mathcal{G}}\mathrm{E}_{\theta_0}\hat{\pi}(I|Y) \le e^{-c_2 m}\sum_{I\in\mathcal{I}}\big(ep/|I|\big)^{-c_1|I|} \le e^{-c_2 m}, \tag{5.13}$$

which leads to the assertion (i) with $m_0 = c_3$ and $c' = c_2$.

To prove (ii), note that for any $\tau_0' > 2m_0$, where m_0 is obtained from part (i), $|I| \ge \tau_0' s(\theta_0)$ implies that

$$\begin{aligned} r^2(I,\theta_0) &\ge \sigma^2|I|\log(ep/|I|) \ge \tau_0'\sigma^2 s(\theta_0)\big[\log\big(ep/s(\theta_0)\big) - \log\tau_0'\big] \\ &\ge \frac{\tau_0'}{2}\sigma^2 s(\theta_0)\log\big(ep/s(\theta_0)\big) \end{aligned}$$

provided that $s(\theta_0) < ep/(\tau_0')^2$. Using (2.13), the relation above implies that $r^2(I,\theta_0) \ge m_0 r^2(\theta_0) + m\sigma^2$, where $m = (\tau_0'/2 - m_0)s(\theta_0)\log(ep/s(\theta_0))$. Hence by part (i), the assertion holds for $\tau_0 = \tau_0'$ and $c'' = c'(\tau_0'/2 - m_0)$ whenever $s(\theta_0) < ep/(\tau_0')^2$. If $s(\theta_0) \ge ep/(\tau_0')^2$, the result trivially holds by choosing $\tau_0 = (\tau_0')^2/e$. Hence the choice $\tau_0 = \max\{\tau_0', (\tau_0')^2/e\}$ ensures the result for any θ_0. □

PROOF OF COROLLARY 2. By the definition of the compatibility coefficient, if I with $|I| \le \tau_0 s(\theta_0)$, then $\|\theta - \theta_0\|_1 \le \sqrt{(\tau_0+1)s(\theta_0)}\|\mathrm{X}(\theta-\theta_0)\|/\|\mathrm{X}\|_{\max}\phi_1((\tau_0+1)s(\theta_0))$ since the cardinality of $\mathrm{supp}(\theta - \theta_0)$ is at most $(\tau_0+1)s(\theta_0)$. By Theorem 1 and Corollary 1, respectively, we obtain $\mathrm{E}_{\theta_0}\hat{\pi}(\|\mathrm{X}(\theta-\theta_0)\| > Mr(\theta_0)|Y) \le CM^{-2}$ and

$$\mathrm{E}_{\theta_0}\hat{\pi}\big(|I| \ge \tau_0 s(\theta_0)|Y\big) \le \exp\big[-c''s(\theta_0)\log\big(ep/s(\theta_0)\big)\big].$$

Thus the first assertion follows. The proof of the second one is similar. □

PROOF OF COROLLARY 3. If $|\theta_{0j}| > M\sigma\|\mathrm{X}\|_{\max}^{-1}\sqrt{s(\theta_0)\log(ep/s(\theta_0))}/\phi_2((\tau_0+1)\times s(\theta_0))$ and the posterior does not select the jth predictor (i.e., sets $\theta_j = 0$), then clearly $\|\theta - \theta_0\| \ge |\theta_{0j}| \ge Mr(\theta_0)\|\mathrm{X}\|_{\max}^{-1}/\phi_2((\tau_0+1)s(\theta_0))$ by (2.13). The result now follows from the second part of Corollary 2. □

To prove Theorem 2, we need to establish a bound which assures that the cardinality of the support of θ chosen from the posterior can rarely be much smaller than the cardinality of a τ-oracle $I_o^\tau(\theta_0)$ defined by (2.3) for some sufficiently large $\tau > 0$.

LEMMA 2. *If $\varrho \in [0,1)$ and $\tau > \bar{\tau}(\varrho) \triangleq [4\varkappa(1+\varrho)+2\log(1+\kappa)]/(1-\varrho(1+\log(1/\varrho))$, then for any $\theta_0 \in \mathbb{R}^p$,*

$$\mathrm{E}_{\theta_0}\hat{\pi}\big(|I| \le \varrho|I_o^\tau(\theta_0)\|Y\big) \le \big(e^{\varkappa-1}-1\big)^{-1}\exp\big\{-\alpha|I_o^\tau(\theta_0)|\log\big(ep/|I_o^\tau(\theta_0)|\big)\big\},$$

where $\alpha = \alpha(\tau,\varrho) \triangleq \tau[1-\varrho(1+\log(1/\varrho))]/4 - \varkappa(1+\varrho) - \log(1+\kappa)/2 > 0$.

In particular, if $\tau > \bar{\tau}(e^{-1})$, then, for all $\varrho \in [0, e^{-1}]$, $\mathrm{E}_{\theta_0}\hat{\pi}(|I| \le \varrho|I_o^\tau(\theta_0)|Y) \le \varrho^{\alpha_0}$, where $\alpha_0 = \alpha(\tau, e^{-1}) > 0$.

PROOF. For each $\theta_0 \in \mathbb{R}^p$ and $I \in \mathcal{I}$ such that $|I| \le \varrho |I_o^\tau|$, let $I_0 = I \cup I_o^*$ for some $I_o^* \subseteq I_o^\tau$ such that $I_0 \in \mathcal{I}$ and $\operatorname{col}(\mathrm{X}_I) + \operatorname{col}(\mathrm{X}_{I_o^\tau}) = \operatorname{col}(\mathrm{X}_{I_0})$. For brevity, we have suppressed the dependence of I_0 on I and θ_0 and the dependence of $I_o^\tau = I_o^\tau(\theta_0)$ on θ_0. Clearly,

$$\max\bigl(|I|, |I_o^\tau|\bigr) \le |I_0| \le |I| + |I_o^\tau| \le (1+\varrho)|I_o^\tau|. \tag{5.14}$$

As the function $x \mapsto x \log(ep/x)$ is increasing on $[0, p]$, for $|I| \le \varrho |I_o^\tau|$,

$$|I| \log \frac{ep}{|I|} \le \varrho |I_o^\tau| \log \frac{ep}{\varrho |I_o^\tau|} \le \varrho(1 + \log(1/\varrho)) |I_o^\tau| \log \frac{ep}{|I_o^\tau|}. \tag{5.15}$$

Because $\operatorname{col}(\mathrm{X}_{I_0})$ contains $\operatorname{col}(\mathrm{X}_{I_o^\tau})$, the difference $H_{I_0} - H_{I_o^\tau}$ of the corresponding projections H_{I_0} and $H_{I_o^\tau}$ is nonnegative definite. Therefore,

$$\begin{aligned}
\sigma^{-2}\theta_0' \mathrm{X}'(\mathrm{H}_{I_0} - \mathrm{H}_I)\mathrm{X}\theta_0 &= \sigma^{-2}\bigl(\theta_0' \mathrm{X}'(\mathrm{I} - \mathrm{H}_I)\mathrm{X}\theta_0 - \theta_0' \mathrm{X}'(\mathrm{I} - \mathrm{H}_{I_0})\mathrm{X}\theta_0\bigr) \\
&\ge \sigma^{-2}\bigl(\theta_0' \mathrm{X}'(\mathrm{I} - \mathrm{H}_I)\mathrm{X}\theta_0 - \theta_0' \mathrm{X}'(\mathrm{I} - \mathrm{H}_{I_o^\tau})\mathrm{X}\theta_0\bigr) \\
&\ge \tau\bigl(|I_o^\tau| \log(ep/|I_o^\tau|) - |I| \log(ep/|I|)\bigr),
\end{aligned}$$

where the last relation holds because $r_\tau^2(I_o^\tau, \theta_0) \le r_\tau^2(I, \theta_0)$ by the definition of I_o^τ as the minimizer of (2.3). In view of (5.15), this gives the bound

$$\sigma^{-2}\theta_0' \mathrm{X}'(\mathrm{H}_{I_0} - \mathrm{H}_I)\mathrm{X}\theta_0 \ge \tau a(\varrho) |I_o^\tau| \log(ep/|I_o^\tau|), \tag{5.16}$$

where $a(\varrho) = 1 - \varrho(1 + \log(1/\varrho))$.

Since $\mathrm{H}_{I_0} - \mathrm{H}_I$ is also a projection, we have $(\mathrm{I} + \mathrm{H}_{I_0} - \mathrm{H}_I)^{-1} = \mathrm{I} - \frac{1}{2}(\mathrm{H}_{I_0} - \mathrm{H}_I)$ and $\det(\mathrm{I} + (\mathrm{H}_{I_0} - \mathrm{H}_I)) \ge 1$. With $h = 1$ and our choice for I_0 in (5.4), the expression for λ_I, (5.14) and (5.16), it follows that for $|I| \le \varrho |I_o^\tau|$, $\mathrm{E}_{\theta_0} \hat{\pi}(I|Y)$ is bounded by

$$\begin{aligned}
&\frac{\lambda_I}{\lambda_{I_0}} \mathrm{E}_{\theta_0} \frac{\exp\{-\frac{1}{2\sigma^2} Y'(\mathrm{H}_{I_0} - \mathrm{H}_I)Y\}}{(1+\kappa)^{(|I| - |I_0|)/2}} \\
&\quad = \frac{\lambda_I \exp\{-\frac{1}{4\sigma^2}\theta_0' \mathrm{X}'(\mathrm{H}_{I_0} - \mathrm{H}_I)\mathrm{X}\theta_0\}}{\lambda_{I_0}(1+\kappa)^{(|I| - |I_0|)/2} \sqrt{\det(\mathrm{I} + (\mathrm{H}_{I_0} - \mathrm{H}_I))}} \\
&\quad \le \frac{\lambda_I}{c_\varkappa} \exp\biggl\{-\frac{1}{4}\tau a(\varrho) |I_o^\tau| \log \frac{ep}{|I_o^\tau|} + \varkappa |I_0| \log \frac{ep}{|I_0|} + \frac{1}{2}\bigl(|I_0| - |I|\bigr)\log(1+\kappa)\biggr\} \\
&\quad \le \bigl(e^{\varkappa - 1} - 1\bigr)^{-1} \lambda_I \exp\biggl\{-\frac{1}{4}\tau a(\varrho) |I_o^\tau| \log \frac{ep}{|I_o^\tau|} + \varkappa(1+\varrho) |I_o^\tau| \log \frac{ep}{|I_o^\tau|} \\
&\qquad + \frac{1}{2}\log(1+\kappa) |I_o^\tau|\biggr\} \\
&\quad \le \bigl(e^{\varkappa - 1} - 1\bigr)^{-1} \lambda_I \exp\biggl\{\biggl[-\frac{1}{4}\tau a(\varrho) + \varkappa(1+\varrho) + \frac{1}{2}\log(1+\kappa)\biggr] |I_o^\tau| \log \frac{ep}{|I_o^\tau|}\biggr\}.
\end{aligned}$$

As $\sum_I \lambda_I = 1$ and $(e^{\varkappa - 1} - 1)^{-1} \le 1$ since $\varkappa > 2$ by the choice in (2.6), the result follows by summing over $I \in \mathcal{I}$: $|I| \le \varrho |I_o^\tau|$.

To show the second assertion of the lemma, consider two cases $|I_o^\tau| \le \log \varrho^{-1}$ and $|I_o^\tau| > \log \varrho^{-1}$, with $\varrho \in [0, e^{-1}]$. In the first case, as $\varrho \le e^{-1}$, we have

$$\mathrm{E}_{\theta_0} \hat{\pi}\bigl(|I| \le \varrho |I_o^\tau| | Y\bigr) \le \mathrm{E}_{\theta_0} \hat{\pi}\bigl(|I| \le \varrho \log(1/\varrho) | Y\bigr) \le \mathrm{E}_{\theta_0} \hat{\pi}\bigl(|I| \le e^{-1} | Y\bigr) = 0.$$

Consider the second case $|I_o^\tau| > \log \varrho^{-1}$. Since $\alpha = \alpha(\tau, \varrho)$ is a decreasing function of $\varrho \ge 0$, we have that $\alpha_0 = \alpha(\tau, e^{-1}) \le \alpha(\tau, \varrho)$ for each $\varrho \in [0, e^{-1}]$. Then by the first assertion of the

lemma and $|I_o^\tau| > \log \varrho^{-1}$,

$$\mathrm{E}_{\theta_0}\hat{\pi}\big(|I| < \varrho |I_o^\tau| | Y\big) \le e^{-\alpha(\tau,\varrho)|I_o^\tau| \log(ep/|I_o^\tau|)} \le e^{-\alpha_0 |I_o^\tau|} \le \varrho^{\alpha_0}. \qquad \square$$

PROOF OF THEOREM 2. We first establish the coverage property. According to Theorem 1, we have that, for any $\theta_0 \in \mathbb{R}^p$ and $M > 0$,

$$\mathrm{P}_{\theta_0}\big(\|\mathrm{X}\theta_0 - \mathrm{X}\hat{\theta}\| \ge M r(\theta_0)\big) \le C M^{-2}. \tag{5.17}$$

Since $\tau \ge 1$ and $\theta_0 \in \Theta_{\mathrm{eb}}(t,\tau)$, it follows that $r^2(\theta_0) \le r_\tau^2(\theta_0) \le (1+t)\sigma^2 |I_o^\tau| \log(ep/|I_o^\tau|)$, where $I_o^\tau = I_o^\tau(\theta_0)$. This combined with the definition (2.18) of $\hat{\rho}$ and (5.17), leads to

$$\begin{aligned}
&\mathrm{P}_{\theta_0}\big(\mathrm{X}\theta_0 \notin B(\mathrm{X}\hat{\theta}, M\hat{\rho})\big) \\
&\quad \le \mathrm{P}_{\theta_0}\big(\|\mathrm{X}\theta_0 - \mathrm{X}\hat{\theta}\| > M\hat{\rho}, \hat{\rho} \ge \delta r(\theta_0)\big) + \mathrm{P}_{\theta_0}\big(\hat{\rho} < \delta r(\theta_0)\big) \\
&\quad \le \mathrm{P}_{\theta_0}\big(\|\mathrm{X}\theta_0 - \mathrm{X}\hat{\theta}\| > M\delta r(\theta_0)\big) + \mathrm{P}_{\theta_0}\bigg(|\hat{I}| \log\frac{ep}{|\hat{I}|} < \delta^2(1+t)|I_o^\tau| \log\frac{ep}{|I_o^\tau|}\bigg) \\
&\quad \le \frac{C}{M^2\delta^2} + \mathrm{P}_{\theta_0}\bigg(|\hat{I}| \log\frac{ep}{|\hat{I}|} < \delta^2(1+t)|I_o^\tau| \log\frac{ep}{|I_o^\tau|}\bigg).
\end{aligned}$$

Let $\tilde{\delta} \triangleq \delta^2(1+t) \le e^{-1}$. As the function $x \mapsto x\log(ep/x)$ is increasing on $[0, p]$,

$$\mathrm{P}_{\theta_0}\bigg(|\hat{I}| \log\frac{ep}{|\hat{I}|} < \tilde{\delta}|I_o^\tau| \log\frac{ep}{|I_o^\tau|}\bigg) \le \mathrm{P}_{\theta_0}\bigg(|\hat{I}| \log\frac{ep}{|\hat{I}|} < \tilde{\delta}|I_o^\tau| \log\frac{ep}{\tilde{\delta}|I_o^\tau|}\bigg) = \mathrm{P}_{\theta_0}\big(|\hat{I}| < \tilde{\delta}|I_o^\tau|\big).$$

The second assertion of Lemma 2 gives $\mathrm{P}_{\theta_0}(|\hat{I}| < \tilde{\delta}|I_o^\tau|) \le (\delta^2(1+t))^{\alpha_0}$ for all $\delta \le 1/\sqrt{e(1+t)}$. This and the last two displays entail that, for $\delta \le 1/\sqrt{e(1+t)}$,

$$\mathrm{P}_{\theta_0}\big(\mathrm{X}\theta_0 \notin B(\mathrm{X}\hat{\theta}, M\hat{\rho})\big) \le \frac{C}{M^2\delta^2} + \big(\delta^2(1+t)\big)^{\alpha_0},$$

for all $\theta_0 \in \Theta_{\mathrm{eb}}(t,\tau)$. Take $\delta = M^{-1/(1+\alpha_0)}/\sqrt{1+t}$, so that for all $M \ge M_0 = e^{(1+\alpha_0)/2}$,

$$\sup_{\theta_0 \in \Theta_{\mathrm{eb}}(t,\tau)} \mathrm{P}_{\theta_0}\big(\mathrm{X}\theta_0 \notin B(\mathrm{X}\hat{\theta}, M\hat{\rho})\big) \le \big(C(1+t)^2 + 1\big) M^{-2\alpha_0/(1+\alpha_0)}.$$

This establishes the coverage relation as the right-hand side can be bounded by ϵ_1 uniformly in $\theta_0 \in \Theta_{\mathrm{eb}}(t,\tau)$ by choosing M sufficiently large (depending on t and ϵ_1 only).

We now show the size property. Introduce the set $\mathcal{G}(L) = \mathcal{G}(L,\theta_0) = \{I \in \mathcal{I} : \sigma^2|I| \times \log(ep/|I|) \ge L^2 r^2(\theta_0)\}$. Then for any $\theta_0 \in \mathbb{R}^p$ and all $I \in \mathcal{G}(L)$,

$$\sigma^{-2}\big(r^2(I,\theta_0) - c_3 r^2(\theta_0)\big) \ge |I| \log\big(ep/|I|\big) - c_3\sigma^{-2} r^2(\theta_0) \ge \big(L^2 - c_3\big)\sigma^{-2} r^2(\theta_0).$$

From (5.7) and the last relation, it follows that for any $I \in \mathcal{G}(L)$,

$$\begin{aligned}
\mathrm{P}_{\theta_0}(\hat{I} = I) &\le \big(ep/|I|\big)^{-c_1|I|} \exp\big\{-c_2\sigma^{-2}\big(r^2(I,\theta_0) - c_3 r^2(\theta_0)\big)\big\} \\
&\le \big(ep/|I|\big)^{-c_1|I|} \exp\big\{-c_2\big(L^2 - c_3\big)\sigma^{-2} r^2(\theta_0)\big\}.
\end{aligned}$$

Note that $r^2(\theta_0) \ge \tau^{-1} r_\tau^2(\theta_0) \ge \sigma^2 |I_o^\tau| \log(ep/|I_o^\tau|)$. Using the last relation and reasoning as in (5.13), we derive that, for any $\theta_0 \in \mathbb{R}^p$,

$$\begin{aligned}
\mathrm{P}_{\theta_0}\big(\hat{\rho} \ge L r(\theta_0)\big) &= \sum_{I \in \mathcal{G}(L,\theta_0)} \mathrm{P}_{\theta_0}(\hat{I} = I) \\
&\le \exp\big\{-c_2\big(L^2 - c_3\big)\sigma^{-2} r^2(\theta_0)\big\} \sum_{I \in \mathcal{I}} \big(ep/|I|\big)^{-c_1|I|} \\
&\le \exp\big\{-c_2\big(L^2 - c_3\big)|I_o| \log\big(ep/|I_o|\big)\big\} \le e^{c_2 c_3} \exp\big\{-c_2 L^2\big\}.
\end{aligned}$$

Clearly, a sufficiently large $L \ge L_0$ makes the bound smaller than any given $\epsilon_2 > 0$. $\square$

PROOF OF COROLLARY 4. Consider the case $\hat{\pi} = \check{\pi}$, the other case is similar. Note that $\hat{\theta}$ has support $\hat{I}$. By part (ii) of Corollary 1, $\hat{I}$ has cardinality $|\hat{I}| \le \tau_0 s(\theta_0) \le \tau_0 s$ for some constant $\tau_0 > 0$ with probability at least $1 - \exp\{-c'' s(\theta_0) \log(ep/s(\theta_0))\}$, where $c'' > 0$ is a constant. Since $(\hat{\theta} - \theta_0)$ is supported within an index set of cardinality at most $(\tau_0 + 1)s(\theta_0)$, by the definition of the compatibility coefficient ϕ_2, it follows that $\|\mathrm{X}(\hat{\theta} - \theta_0)\| \ge \phi_2((\tau_0 + 1)s(\theta_0))\|\mathrm{X}\|_{\max}\|\hat{\theta} - \theta_0\|$. This leads to the first claim of the corollary. The proof for the ℓ_1-case is similar. The last relation is inherited from Theorem 2. □

Acknowledgments. Research of the second author is partially supported by National Science Foundation (NSF) grant number DMS-1510238. A visit to VU Amsterdam was supported by visitor STAR grant from the Netherlands Organisation for Scientific Research (NWO).

SUPPLEMENTARY MATERIAL

Supplement to "Empirical Bayes oracle uncertainty quantification for regression" (DOI: 10.1214/19-AOS1845SUPP; .pdf). More elaboration on empirical Bayes interpretation of the procedures, computational strategy and proofs of some results are provided in Supplement [3].

REFERENCES

[1] BARAUD, Y. (2004). Confidence balls in Gaussian regression. *Ann. Statist.* **32** 528–551. MR2060168 https://doi.org/10.1214/009053604000000085

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

[3] BELITSER, E. and GHOSAL, S. (2020). Supplement to "Empirical Bayes oracle uncertainty quantification for regression." https://doi.org/10.1214/19-AOS1845SUPP

[4] BELITSER, E. and NURUSHEV, N. (2020). Needles and straw in a haystack: Robust empirical Bayes confidence for possibly sparse sequences. *Bernoulli* **26** 191–225. MR403632 https://doi.org/10.3150/19-BEJ1122

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

[6] BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549 https://doi.org/10.3150/12-BEJSP11

[7] CAI, T. T. and GUO, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395 https://doi.org/10.1214/16-AOS1461

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

[9] 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

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

[11] CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874 https://doi.org/10.1214/15-AOS1334

[12] CASTILLO, I. and SZABÓ, B. (2020). Spike and slab empirical Bayes sparse credible sets. *Bernoulli* **26** 127–158. MR4036030 https://doi.org/10.3150/19-BEJ1119

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

[14] CURTIS, S. M., BANERJEE, S. and GHOSAL, S. (2014). Fast Bayesian model assessment for nonparametric additive regression. *Comput. Statist. Data Anal.* **71** 347–358. MR3131975 https://doi.org/10.1016/j.csda.2013.05.012

[15] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089 https://doi.org/10.1093/biomet/81.3.425

[16] FREEDMAN, D. (1999). On the Bernstein–von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* **27** 1119–1140. MR1740119 https://doi.org/10.1214/aos/1017938917

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

[18] HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. *Monographs on Statistics and Applied Probability* **43**. CRC Press, London. MR1082147

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

[20] LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** 1001–1008. MR1015135 https://doi.org/10.1214/aos/1176347253

[21] LIN, Y. and ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34** 2272–2297. MR2291500 https://doi.org/10.1214/009053606000000722

[22] MARTIN, R., MESS, R. and WALKER, S. G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* **23** 1822–1847. MR3624879 https://doi.org/10.3150/15-BEJ797

[23] MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779–3821. MR2572443 https://doi.org/10.1214/09-AOS692

[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] PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28** 298–335. MR1762913 https://doi.org/10.1214/aos/1016120374

[26] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.* **13** 389–427. MR2913704

[27] RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009). Sparse additive models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 1009–1030. MR2750255 https://doi.org/10.1111/j.1467-9868.2009.00718.x

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

[29] SNIEKERS, S. and VAN DER VAART, A. (2015). Credible sets in the fixed design model with Brownian motion prior. *J. Statist. Plann. Inference* **166** 78–86. MR3390135 https://doi.org/10.1016/j.jspi.2014.07.008

[30] SNIEKERS, S. and VAN DER VAART, A. (2015). Adaptive Bayesian credible sets in regression with a Gaussian process prior. *Electron. J. Stat.* **9** 2475–2527. MR3425364 https://doi.org/10.1214/15-EJS1078

[31] 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

[32] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[33] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316 https://doi.org/10.1214/09-EJS506

[34] YANG, Y. and TOKDAR, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43** 652–674. MR3319139 https://doi.org/10.1214/14-AOS1289

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