EMPIRICAL BAYES ORACLE UNCERTAINTY QUANTIFICATION FOR REGRESSION

BY EDUARD BELITSER AND SUBHASHIS GHOSAL*

VU Amsterdam and North Carolina State University

We propose an empirical Bayes method for high dimensional linear regression models. Following an oracle approach which quantifies the error locally for each possible value of the parameter, we show that the 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 [30]]. 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 adaptive optimal size is far more delicate than estimation accuracy. For instance, in the well-studied infinitely many normal means model

$$X_i \sim N(\theta_i, n^{-1}), \ i = 1, 2, \ldots,$$

$$\sum_{i=1}^{\infty} \theta_i^2 < \infty,$$

it is impossible for any confidence set to have coverage at all parameter values maintaining the optimal diameter adaptively over different classes [Li [18], Baraud [1], Cai and Low [6]]. 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

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

primary 62G15, 62C05, 62H35; secondary 62G99.

Keywords and phrases: credible ball, coverage, empirical Bayes, excessive bias restriction, oracle rate.
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 non-deceptive 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 [23]], polished tail [Szabo et al. [29]] and excessive bias restriction [Belitser [2]]. To the best of our knowledge, the only articles dealing with construction of confidence regions in high dimensional linear regression models are Nickl and van de Geer [22] and Cai and Guo [5]. 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 instead of the parameter as a whole.

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 [10] and Freedman [13]. The problem arises essentially because in nonparametric problems, under optimal smoothing, the order of the bias matches with the order of the posterior variation and hence the coverage may be arbitrarily low. In the 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. [17] obtained desired coverage by undersmoothing, while Castillo and Nickl [7] constructed credible sets with frequentist coverage through Bernstein-von Mises theorems in negative Sobolev spaces using wavelet based priors. Ray [26] pursued the approach of Castillo and Nickl [7] 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. [29], Belitser [2] and Belitser and Nurushev [3]. 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 [8] extended their earlier work to density estimation
problems. Yoo and Ghosal [33] used inflated credible regions to obtain coverage for multivariate nonparametric regression for pointwise, $L_2$- and $L_\infty$-credible sets using priors based on B-spline basis expansion. Sniekers and van der Vaart [28] used scaled Brownian motion to construct credible sets with coverage properties for nonparametric regression while Sniekers and van der Vaart [27] 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 has not been studied thus far.

In this paper we adopt the oracle approach of Belitser [2] and Belitser and Nurushev [3] to quantify estimation accuracy, coverage and precision of an empirical Bayesian procedure for the linear regression problem with a large number of predictors. Suppose that we observe

\begin{equation}
Y = X\theta + \sigma \varepsilon,
\end{equation}

where $X = (X_1, \ldots, X_p)$ is a deterministic $n \times p$ design matrix, $\theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$ is an unknown regression parameter where possibly $p \gg n$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ 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. Like in comparable works in the literature, the noise intensity $\sigma > 0$ will be considered known in our setting. For a predictor of the form $\hat{Y} = X\hat{\theta}$ for some estimator $\hat{\theta}$ of $\theta$, the prediction risk is defined by $R^2(I, \theta) = E\|X\hat{\theta}(I) - X\theta\|^2$, where $\hat{\theta}(I)$ is $\hat{\theta}$ computed on the basis of the model indexed by $I \subseteq \{1, \ldots, p\}$, 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 $X\theta$ at any given value of $\theta$, 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 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 polished tail-ness 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 suitable for ensuring uniformity in uncertainty quantification. 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. In Section 3, a techniques of posterior computation is described. A thorough discussion of the key condition of excessive bias restriction is presented in Section 4. An extension of the results to additive nonparametric regression models is given in Section 5. Proofs are presented in Section 6.

2. Setup and main results. Consider the linear regression model (1.1) with \( \varepsilon \sim N_n(0, I) \), the \( n \)-variate normal distribution with mean vector 0 and dispersion matrix \( I \), the identity matrix. We consider \( X \) to be fixed in our setup and various objects defined in course may depend on \( X \).

We use the following notation and conventions throughout the paper. We denote the probability distribution of \( Y \) from the model (1.1) by \( P_\theta \) and the corresponding expectation is denoted by \( E_\theta \). 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 \( k, l \), \( a_k \) is denoted by \( E \theta \). Let \( 1 - B \leq 0 \) and let \( 0 \) 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 \( A \leq B \) (respectively, \( A < B \)) if \( B - A \) is nonnegative definite (respectively, positive definite). The symbol \( \triangleq \) will refer to equality by definition. Let \( 1 \) stand for the indicator function.

Let \( \mathcal{I} = \{ \emptyset \neq I \subseteq \{1, \ldots, p\} : \text{the columns } (X_i, i \in I) \text{ are linearly independent} \} \), \( X_I \) be the submatrix of \( X \) with columns \( (X_i, i \in I) \). Clearly, \( |I| \leq 2^p \) and \( |I| \leq \nu \) for any \( I \in \mathcal{I} \), where \( \nu = \text{rank}(X) \leq \min(n, p) \). If \( I \in \mathcal{I} \) were known to be the support of \( \theta \), 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

\[
\begin{align*}
\{ \hat{\theta}(I), I \in \mathcal{I} : \hat{\theta}_I(I) = (X_I'X_I)^{-1}X_I'Y, \ \hat{\theta}_{I^c}(I) = 0_{|I^c|} \},
\end{align*}
\]

and introducing the oracle risk relative to this family. Let \( H_I = X_I(X_I'X_I)^{-1}X_I' \) be the orthogonal projection matrix onto the column space \( \text{col}(X_I) \) of \( X_I \). Since \( X\theta = X_I\theta_I + X_{I^c}\theta_{I^c} \) and \( H_I(I - H_I) = 0 \), the quadratic prediction risk of

\[
\begin{align*}
X\hat{\theta}(I) = X_I\hat{\theta}_I(I) = H_IY = H_I(X_I\theta_I + X_{I^c}\theta_{I^c} + \sigma\varepsilon) = X_I\theta_I + H_I(X_{I^c}\theta_{I^c} + \sigma\varepsilon)
\end{align*}
\]

is given by

\[
\begin{align*}
R^2(I, \theta) = & E_\theta \|X\hat{\theta}(I) - X\theta\|^2 = E_\theta \|H_IY - X_I\theta_I - X_{I^c}\theta_{I^c}\|^2 = E_\theta \|(H_I - I)X_{I^c}\theta_{I^c} + \sigma H_I\varepsilon\|^2 \\
= & \|(I - H_I)X_{I^c}\theta_{I^c}\|^2 + \sigma^2\text{tr}[(H_I' H_I)] = \theta' X'(I - H_I)X\theta + \sigma^2 |I|.
\end{align*}
\]
This follows from the standard fact that if $V$ is a random vector with mean $\mu$ and covariance matrix $\Sigma$ and $A$ is a symmetric matrix of constants, then $E(V'AV) = \text{tr}(A\Sigma) + \mu' A \mu$. For a given $\theta$, the minimizer $I_0^R = I_0^R(\theta)$ of the risk in (2.3) with respect to $I \in I$ is called the $R$-oracle. The corresponding minimum $R(\theta) = R(I_0^R, \theta)$ quantifies the level of accuracy of inference about $X\theta$ 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 $\theta$, uniformly like the oracle procedure $\theta_{iR}$ without knowing $I_0^R$, that is, will mimic the $R$-oracle. However, even in the case $X = I$, Donoho and Johnstone [12] and Birgé and Massart [4] 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 [4]). 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 $\tau$-oracle $I_\tau = I_\tau^*(\theta)$ at $\theta$ as a minimizer (with respect to $I \in I$) of

$$r_\tau^2(I, \theta) = \| (I - H_I) X \theta_{iR}^{-1} \|^2 + \tau \sigma^2|I| \log \frac{ep}{|I|} = \| (I - H_I) X \theta \|^2 + \tau \sigma^2|I| \log \frac{ep}{|I|}.$$  

It should be noted that the quantity $r_\tau(I_\tau^*, \theta)$, called the $\tau$-oracle rate, is unique but the $\tau$-oracle $I_\tau^*$ 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 $i = 1, \ldots, p$, is a $\tau$-oracle. Note that the empty set is not allowed to be the oracle by the definition. The rate (2.4) is clearly a modification of the prediction risk (2.3) where the variance part has been inflated by the factor $\tau \log(ep/|I|)$. Denote from now on $r_\tau(\theta) = r_\tau(I_\tau^*, \theta)$, $r(I, \theta) = r_1(I, \theta)$, $I_0 = I_0^1$ and $r(\theta) = r_1(I_0^*, \theta)$.

It is well known that in the normal regression setting, a normal prior for $\theta$ is conjugate, but it overshrinks towards the prior mean. This property led Castillo et al. [9] in their study of posterior contraction to consider priors with tail at least as thick as that of the Laplace distribution for each coefficient selected as non-zero. In spite of this overshrinkage, a normal prior is actually useful if the mean of the prior is not set to a fixed value but instead is chosen by an empirical Bayes technique, as noted earlier by Belitser [2] and Belitser and Nurushev [3] in the context of the normal means problem. 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 coverage of (empirical Bayes) adaptive credible regions of optimal size, in course we also derive posterior contraction results analogous to those in Castillo et al. [9] based on a spike-and-slab conjugate normal prior with hyperparameters chosen by the empirical Bayes technique instead of Laplace prior used by them. The exact expressions that we obtain lead to a significantly shorter derivation of equivalent results, and more importantly, allows obtaining coverage of credible regions.

For the model (1.1) with $\varepsilon \sim N_n(0, I)$, introduce a two-level hierarchical prior

$$\pi_I(\theta) : \theta|I \sim N(\mu(I), \kappa \sigma^2(X_I'X_I)^{-1}) \otimes \delta_{0_{1|\varepsilon}},$$

and

$$\pi_I(\varepsilon) : \varepsilon|I \sim N(0, I) \otimes \delta_{\varepsilon|I}.$$
where $\kappa > 0$, $\kappa > 1$ are chosen constants, vectors $\mu_I = (\mu(I) : I \in \mathcal{I})$ are to be chosen by the empirical Bayes method and $c_\kappa = c_{\kappa, p}$ is the normalizing constant making $\sum_{I \in \mathcal{I}} \lambda_I = 1$. The prior on $\theta$ given $I$ is Zellner’s $g$-prior, which was also used in the same context by Martin et al. [20] along with an empirical Bayes choice for the mean $\mu(I)$ like ours. However, they imposed absolute restriction on the size of the model by a constant value 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 $\lambda_I$ for $I$ given by (2.6) is proper since 

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

which also implies that $c_\kappa \geq e^{\kappa-1} - 1 > 0$. To make our bounds useful, the parameters $\kappa$ and $\kappa$ need to satisfy the conditions

$$\kappa \geq (1 - h)^{-1/h} - 1, \quad \kappa > 2/h \quad \text{for some} \quad h \in (0, 1/3).$$

These conditions will be assumed throughout below.

In the model indexed by $I$, the marginal density $\pi_{I, \mu}(Y)$ of $Y$ is given by

$$\pi_{I, \mu}(Y) : Y \mid I \sim N_{n_I}(X_I \mu(I), \sigma^2(1 + \kappa H_I)).$$

The corresponding posterior distribution of $\theta$ is described hierarchically by

$$\pi_{I, \mu}(\theta \mid Y) : \theta \mid Y \sim N_{n_I}(\mu(I) + \kappa \theta(I), \kappa \sigma^2(1 + \kappa H_I)^{-1}) \otimes \delta_{0_\mu|\mathcal{I}},$$

$$\pi_{I}(Y) = \frac{\lambda_I \pi_{I, \mu}(Y)}{\sum_{J \in \mathcal{I}} \lambda_J \pi_{J, \mu}(Y)}, \quad I \in \mathcal{I}.$$

Hence the posterior density of $\theta$ given $Y$ is

$$\pi_{\mu}(\theta \mid Y) = \sum_{I \in \mathcal{I}} \pi_{I, \mu}(\theta \mid Y) \pi_{\mu}(I \mid Y).$$

Further, the marginal of $Y$ is given by $\pi_{\mu}(Y) = \sum_{I \in \mathcal{I}} \lambda_I \pi_{I, \mu}(Y)$. To choose the family of hyper-parameters $\mu = \mu_{\mathcal{I}} = (\mu(I) : I \in \mathcal{I})$ of the prior distribution, we follow the empirical Bayes approach of maximizing the marginal density $\pi_{\mu}(Y)$ with respect to $\mu$. This boils down to maximizing $\pi_{I, \mu}(Y)$ defined by (2.8) with respect to $\mu(I)$, which in
turn reduces to minimizing \((Y - X_I \mu(I))' \sigma^2(I + \kappa H_I)^{-1}(Y - X_I \mu(I))\). The solution is given by the weighted least square estimator: for each \(I \in \mathcal{I}\),

\[
\hat{\mu}(I) = (X_I' \sigma^2(I + \kappa H_I))^{-1} X_I'(I + \kappa H_I)^{-1} Y = (X_I' X_I)^{-1} X_I' Y = \hat{\theta}_I(I).
\]

The second equality holds because for a symmetric invertible matrix \(B\) such that \(BA = bA\) for some \(b \neq 0\), \((A'B^{-1}A)^{-1}A'B^{-1} = (A'A)^{-1}A'\).

With \(\hat{\mu}\) given by (2.12) plugged in the expression \(\pi_\mu(I|Y)\) defined by (2.10), we proceed with the empirical Bayes approach by maximizing \(\pi_\mu(I|Y)\) with respect to \(I \in \mathcal{I}\):

\[
(2.13) \quad \hat{I} = \arg\max_{I \in \mathcal{I}} \pi_\mu(I|Y) = \arg\max_{I \in \mathcal{I}} \lambda I \hat{\pi}_I(Y), \quad \text{where} \quad \hat{\pi}_I(Y) = \pi_{I,\hat{\mu}}(Y).
\]

If the minimizer is not unique, any maximizer may be chosen. In view of (2.12), \(X_I \hat{\mu}(I) = X_I \hat{\theta}_I(I) = H_I Y\). Then, by (2.8) we compute the empirical Bayes marginal density \(\hat{\pi}_I(Y) = \pi_{I,\hat{\mu}}(Y)\) of \(Y\) in model \(I\):

\[
(2.14) \quad \hat{\pi}_I(Y) = \frac{\exp\left(-\frac{1}{2\sigma^2} Y'(I - H_I)(I + \kappa H_I)^{-1}(I - H_I)Y\right)}{\sqrt{(2\pi\sigma^2)^n \det(I + \kappa H_I)}} = \frac{\exp\left(-\frac{1}{2\sigma^2} Y'(I - H_I)Y\right)}{(2\pi\sigma^2)^n/2(1 + \kappa)^{|I|/2}},
\]

as \(H_I\) is a projection matrix with \(\text{rank}(H_I) = |I|\), implying \(\det(I + \kappa H_I) = (1 + \kappa)^{|I|}\) and

\[
(I - H_I)(I + \kappa H_I)^{-1}(I - H_I) = (I - H_I)(I - \frac{\kappa}{1 + \kappa} H_I)(I - H_I) = I - H_I.
\]

Finally, we obtain the following empirical Bayes posterior for \(\theta\):

\[
(2.15) \quad \hat{\pi}(\theta|Y) = \pi_{I,\hat{\mu}}(\theta|Y) = N[I \hat{\mu}(\hat{I}), \frac{\kappa \sigma^2}{\kappa + 1} (X_I' X_I)^{-1}] \otimes \delta_{\theta_0[I|\hat{I}|]};
\]

Let \(\hat{E}(-|Y)\) be the expectations with respect to the empirical Bayes posterior distribution \(\hat{\pi}(-|Y)\) and by \(\hat{\theta} = \hat{E}(\theta|Y) = \hat{\theta}(\hat{I})\) denote the empirical Bayes posterior mean. The following theorem shows that the empirical Bayes posterior distribution of the mean vector \(X\theta\) of \(Y\) concentrates near its true value \(X\theta_0\) in the frequentist sense and the empirical Bayes posterior mean converges to \(X\theta_0\) uniformly over the entire parameter space, at the oracle rate \(r(\theta_0)\).

**Theorem 1** (Denoising). There exists a constant \(C > 0\) such that, for any \(\theta_0 \in \mathbb{R}^p\), \(M > 0\), \(E_{\theta_0}[\hat{\pi}(\|X\theta - X\theta_0\|) \geq Mr(\theta_0)|Y] \leq CM^{-2}\) and \(E_{\theta_0}[\|X\hat{\theta} - X\theta_0\|^2] \leq Cr^2(\theta_0)\).

Notice that \(X\hat{\theta} = H_I Y\). The result is analogous to the first conclusion of Theorem 2 of Castillo et al. [9] for the prior (2.5)–(2.6) (with the empirical Bayes choice for the hyperparameter \((\mu(I), I \in \mathcal{I})\)) replacing their prior. This follows by observing that for \(\theta_0\)
with sparsity \( s(\theta_0) = |I^*(\theta_0)| \geq 1 \), where \( I^* = I^*(\theta_0) = \text{supp}(\theta_0) = \{i : \theta_{0i} \neq 0\} \) is the true active index set, we have that \((I - H_T)X\theta_0 = 0\) and, by the definition of the oracle,

\[
r^2(\theta_0) \leq r^2(I^*, \theta_0) \leq \sigma^2 s(\theta_0) \log(ep/s(\theta_0)).
\]

Note that we do not need the compatibility condition connecting \( \ell_1 \)- and \( \ell_2 \)-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 \( \theta_0 \), i.e., the number of non-zero elements of \( \theta_0 \).

We also obtain the following result as a by-product of the proof of Theorem 1, which states that models yielding high risk are not frequently selected by the empirical Bayes method. In particular, models with substantially higher size than the true one are unlikely to occur under the empirical Bayes posterior. This result and Corollaries 2 and 3 below are analogous to the corresponding results of Castillo et al. \[9\] for the posterior distribution using Laplace priors on active coefficients and also of Martin et al. \[20\] for their empirical Bayes pseudo-posterior distribution.

**Corollary 1 (Risk control and sparsity).** There exist positive constants \( m_0, c', c'', \tau_0 \) such that, for any \( \theta_0 \in \mathbb{R}^p \) and \( m > 0 \),

(i) \( P_{\theta_0}(r^2(\hat{I}, \theta_0) \geq m_0 r^2(\theta_0) + m \sigma^2) \leq e^{-c'm} \),

(ii) \( P_{\theta_0}(|\hat{I}| \geq \tau_0 s(\theta_0)) \leq \exp\{-c'' s(\theta_0) \log(ep/s(\theta_0))\} \).

The above claims, being non-asymptotic and uniform in \( \theta_0 \), can be specialized to certain situations. In particular, assertion (ii) in Corollary 1 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 \( \theta_0 \in \ell_0[s_n] \triangleq \{\theta : |\text{supp}(\theta)| \leq s_n\} \). Further, when \( s_n = o(p_n) \), the constant \( \tau_0 \) 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 \( \theta \) near its true value \( \theta_0 \), measured by the usual Euclidean or the \( \ell_1 \)-distance. Because the dimension \( p \) may be (and generally is) larger than \( n \), the correspondence between \( X\theta \) and \( \theta \) 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 \[31\]), we too assume a condition that lower bounds the norm of \( X\theta \) by a positive multiple of a norm on \( \theta \) for sparse vectors, which is in turn a condition on the design matrix \( X \).

Recall that \( s(\theta) \) denotes the number of non-zero elements of \( \theta \in \mathbb{R}^p \), i.e., the cardinality of the support \( \text{supp}(\theta) = I^*(\theta) = \{i : \theta_i \neq 0\} \) of \( \theta \). Let \( \|\theta\|_1 = \sum_{j=1}^p |\theta_j| \) be the \( \ell_1 \)-norm of \( \theta \). Define \( \|X\|_{\max} = \max_{k=1,...,p} \|X_k\| \). For a positive integer \( m \), define

\[
\phi_1(m) = \inf \left\{ \sqrt{m} \frac{\|X\theta\|}{\|X\|_{\max} \|\theta\|_1} : 0 \neq |\text{supp}(\theta)| \leq m \right\},
\]
(2.18) \[ \phi_2(m) = \inf \left\{ \frac{\|X\theta\|}{\|X\|_{\max} \|\theta\|} : 0 \neq |\text{supp}(\theta)| \leq m \right\}. \]

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

\[ \|\theta_I\| = \|(X'_I X_I)^{-1}X'_I X_I \theta_I\| \leq \|(X'_I X_I)^{-1}\| \|X'_I\| \|X_I \theta_I\| = \frac{\lambda_{\max}(X'_I X_I)}{\lambda_{\min}(X'_I X_I)} \|X_I \theta_I\| \]

for any \( I \in \mathcal{I} \), it follows that \( \phi_2(m) \geq \inf \{\lambda_{\min}(X'_I X_I)/\|X\|_{\max} \sqrt{\lambda_{\max}(X'_I X_I)} |I| \leq m, I \in \mathcal{I} \} \), where \( \lambda_{\min}(M) \) and \( \lambda_{\max}(M) \) 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 \( X'X \). In the result below, if any of \( \phi_1 \) or \( \phi_2 \) is zero at its argument, then the result becomes trivial but remains valid.

**Corollary 2 (Recovery).** Let \( \phi_1 \) and \( \phi_2 \) be defined by (2.17) and (2.18) respectively. For constants \( C, \tau_0, c'' \) appearing in Theorem 1 and Corollary 1, we have

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

\[ \mathbb{E}_{\theta_0} \hat{\pi} \left( \|\theta - \theta_0\|_1 \geq \frac{M r(\theta_0)}{\|X\|_{\max} \phi_2((\tau_0 + 1) s(\theta_0))} \left| Y \right| \leq CM^{-2} + \exp[-c'' s(\theta_0) \log(ep/s(\theta_0))]. \]

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 can rarely miss a signal.

**Corollary 3 (Selection).** Let \( \phi_2 \) be defined by (2.18) and assume that

\[ |\theta_{0,j}| > M \sigma |X|_{\min}^{-1} \sqrt{s(\theta_0) \log(ep/s(\theta_0))}/\phi_2((\tau_0 + 1) s(\theta_0)), \]

whenever \( \theta_{0,j} \neq 0 \). Then for constants \( C, \tau_0, c'' \) appearing in Theorem 1 and Corollary 1, we have \( \mathbb{E}_{\theta_0} \hat{\pi}(\theta : \text{supp}(\theta) \nsubseteq \text{supp}(\theta_0)|Y) \leq CM^{-2} + \exp[-c'' s(\theta_0) \log(ep/s(\theta_0))]. \)

**Remark 1.** Plugging-in \( \hat{\mu}(I), I \in \mathcal{I} \) derived in (2.12) in the expressions (2.9), (2.10) and (2.11), we obtain yet another empirical Bayes posterior:

\[ \hat{\pi}(\theta|Y) = \hat{\pi}_{\hat{\mu}}(\theta|Y) = \sum_{I \in \mathcal{I}} \hat{\pi}_{I}(\theta|Y) \hat{\pi}(I|Y), \]

\[ \hat{\pi}(I|Y) = \hat{\pi}_{\hat{\mu}}(I|Y) = \frac{\lambda_{I}\hat{\pi}_{I}(Y)}{\sum_{J \in \mathcal{I}} \lambda_{J}\hat{\pi}_{J}(Y)}, \]

\[ \hat{\pi}_{I}(\theta|Y) = \hat{\pi}_{I,\hat{\mu}}(\theta|Y) = N_{|I|} \left( \hat{\theta}_{I}(I), \frac{\sigma^2}{\kappa+1} (X'_I X_I)^{-1} \right) \otimes \delta_{0_{|I|}}. \]
where \( \hat{\pi}_I(Y) = \pi_{I,h}(Y) \) is defined by (2.14). Let \( \hat{E}(\cdot|Y) \) and \( \hat{E}_I(\cdot|Y) \) be the expectations with respect to the empirical Bayes posterior distributions \( \hat{\pi}(\cdot|Y) \) and \( \hat{\pi}_I(\cdot|Y) \), respectively. Then \( \hat{\theta} = \hat{E}(\theta|Y) = \sum_{I \in \mathcal{I}} \hat{E}_I(\theta|Y)\hat{\pi}(I|Y) = \sum_{I \in \mathcal{I}} \hat{\theta}(I)\hat{\pi}(I|Y) \) is the corresponding empirical Bayes posterior mean.

All the claimed results so far hold exactly in the same form for the empirical Bayes posterior \( \hat{\pi}(\theta|Y) \) and estimator \( \hat{\theta} \) as for \( \hat{\pi}(\theta|Y) \) and \( \hat{\theta} \), where the expressions of the form \( P_{\theta_0}(\hat{I} \in \mathcal{G}) \) for some \( \mathcal{G} \subseteq \mathcal{I} \) should be replaced by expressions \( E_{\theta_0}\hat{\pi}(I: I \in \mathcal{G}|Y) \). Indeed, by the definition of \( \hat{\pi}(I|Y) \) in (2.19), we derive the counterpart of (6.2) ahead as follows: for any \( I, I_0 \in \mathcal{I} \) and any \( h \in [0,1] \),

\[
E_{\theta_0}\hat{\pi}(I|Y) \leq E_{\theta_0}\left[ \frac{\lambda_I\hat{\pi}_I(Y)}{\sum_{J \in \mathcal{I}} \lambda_I\hat{\pi}_J(Y)} \right]^h \leq E_{\theta_0}\left[ \frac{\lambda_I\hat{\pi}_I(Y)}{\lambda_{I_0}\hat{\pi}_{I_0}(Y)} \right]^h.
\]

From that point (relation (6.2)) onwards, the proof of the properties of \( \hat{\pi}(\theta|Y) \) and \( \hat{\theta} \) proceeds exactly in the same way as for \( \hat{\pi}(\theta|Y) \) and \( \hat{\theta} \), with the only difference that everywhere (in the claims and in the proofs) we substitute \( \hat{\pi}(I \in \mathcal{G}|Y) \) instead of \( 1\{\hat{I} \in \mathcal{G}\} \) and \( E_{\theta_0}\hat{\pi}(I \in \mathcal{G}|Y) \) instead of \( P_{\theta_0}(\hat{I} \in \mathcal{G}) \) for all \( \mathcal{G} \subseteq \mathcal{I} \) that appear in the proof.

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 optimal size and required coverage uniformly all over the parameter space. If we assure 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 \( \|I - H_{I^c}(\theta)\|X\theta \|^2 \) of the \( \tau \)-oracle rate, that results from the fact that the coordinates \( \theta(I^c) = (\theta_i, i \notin I^c) \) are classified as zeros by the \( \tau \)-oracle, whereas these may not be zeros. Although this is the best choice from the \( \tau \)-oracle’s perspective, some “deceptive” \( \theta \) values may still have many nonzero coordinates from \( (I^c)^c \) (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 \( \tau \)-bias at \( \theta \in \mathbb{R}^n \) by

\[
(2.20) \quad b_\tau(\theta) = \frac{\theta'X'(I - H_{I^c}(\theta))X\theta}{\sigma^2[I^c_0(\theta)] \log(eP/|I^c_0(\theta)|)},
\]

and hence we may write \( \tau_\tau^2(\theta) = (b_\tau(\theta) + \tau) \log(ep/|I^c_0(\theta)|)\sigma^2[I^c_0(\theta)] \). The \( \tau \)-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 \( \tau \)-bias \( b_\tau(\theta_0) \) will be needed. This naturally leads to a restriction on \( \theta_0 \), as defined below. The restriction will ensure coverage at \( \theta_0 \).
**Definition 1.** We say that the *excessive bias restriction* (EBR) condition with structural parameters \((t, \tau)\) holds at parameter value \(\theta\) if \(b_\tau(\theta) \leq t\). We denote the corresponding region of the parameter space by \(\Theta_{eb}(t, \tau) = \{ \theta \in \mathbb{R}^p : b_\tau(\theta) \leq t \}\).

A discussion on the EBR condition is given in Section 4. We shall need the EBR condition satisfied for \(\tau > \bar{\tau} \triangleq \frac{4\kappa(e + 1) + 2e \log(1 + \kappa)}{(e - 2)}\).

**Remark 2.** The origin of the condition (2.21) can be elucidated as follows. Recall that by the definition of the oracle rate, for any \(\theta \in \mathbb{R}^p\) we have \(|I_\tau^0(\theta)| \leq |I_o^1(\theta)| \leq |I^*(\theta)|\) for any \(0 \leq \tau_1 \leq \tau_2\). This means that the EBR condition is weaker if \(\tau\) is smaller. Interestingly, the “limiting” oracle, as \(\tau \downarrow 0\), \(I_\tau^0(\theta) = I^*(\theta)\) recovers the active index set \(I^*(\theta)\) and the corresponding limiting EBR set would become the entire space: \(\Theta_{eb}(t, 0) = \mathbb{R}^p\). Besides, checking the proofs reveals that the constants will not deteriorate for smaller \(\tau\) (in fact, will improve). However, as condition (2.21) shows, Lemma 2 (and hence Theorem 2) hold only if \(\tau\) 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 \(E_\theta(Y) = X\theta\) using the empirical Bayes posterior distribution \(\hat{\pi}(\theta|Y)\). From (2.15), we have \(X\theta|Y \sim N_n(X\hat{\theta}, \frac{\kappa\sigma^2}{\kappa + 1}H_I)\). Denoting by \(\chi^2_k\) the \((1 - \alpha)\)-quantile of \(\chi^2_k\)-distribution, we have that \(\hat{\pi}(\|X\theta - X\hat{\theta}\|^2 \leq \frac{\kappa\sigma^2}{\kappa + 1} \chi^2_{|I|, \alpha}|Y) = 1 - \alpha\). However \(\chi^2_{|I|, \alpha}\) is bounded by a constant multiple of \(|I|\), and hence for simplicity the latter can replace the former to obtain a credible ball of optimal size up to a constant multiple with guaranteed frequentist coverage. This leads to the set \(B(X\hat{\theta}, M\sigma|\hat{I}|^{1/2})\) as a credible ball for \(X\theta\), 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 \(\theta_0\), with arbitrarily high probability, \(|\hat{I}|\) is of the order \(|I^*_o|\). Since the minimax risk of the class of adaptive estimators is larger than this by a multiplicative factor \(\log(ep/|I^*_o|)\), it is clear that \(B(X\hat{\theta}, M\sigma|\hat{I}|^{1/2})\) 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 \(B(X\hat{\theta}, M\hat{\rho})\), where

\[
\hat{\rho}^2 = \frac{2\sigma^2}{\kappa + 1} |\hat{I}| \log(ep/|\hat{I}|).
\]

The following theorem shows that the logarithmic inflation leads to coverage at all true parameter \(\theta_0\) satisfying the EBR condition.
Theorem 2 (Coverage and oracle size). Let (2.21) 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

\[
\sup_{\theta_0 \in \Theta_{eb}(t, \tau)} P_{\theta_0}(X\theta_0 \notin B(X\hat{\theta}, M\hat{\rho})) \leq \epsilon_1, \quad \sup_{\theta_0 \in \mathbb{R}^p} P_{\theta_0}(\hat{\rho} \geq Lr(\theta_0)) \leq \epsilon_2.
\]

The estimator \( \hat{\theta} \) defined in Remark 1 can also be used in Theorem 2 instead of \( \hat{\theta} \).

Remark 3. While it is appropriate to interpret \( B(X\hat{\theta}, M\hat{\rho}) \) as a credible and confidence region for \( E_\theta Y = X\theta \), it is possible to interpret \( \{ \theta : X\theta \in B(X\hat{\theta}, M\hat{\rho}) \} \) as a region for \( \theta \), which automatically “inherits” the corresponding credibility and coverage. However, as the matrix \( X'X \) is typically rank deficient, the set \( \{ \theta : (\theta - \hat{\theta})'X'X(\theta - \hat{\theta}) \leq M^2\hat{\rho}^2 \} \) obtained from the corresponding quadratic form will be unbounded in terms of the Euclidean distance. The main difficulty is that the relation \( \theta \mapsto X\theta \) is not invertible. Hence, a bounded diameter credible and confidence region is possible only by imposing an additional restriction on the parameter \( \theta \) through the compatibility condition that allows the inversion.

Corollary 4. Let (2.21) hold. Then for any \( t > 0, \theta_0 \in \Theta_{eb}(t, \tau), \epsilon_1, \epsilon_2 > 0 \), there exist \( M = M(t, \epsilon_1) > 0 \), \( L = L(\epsilon_2) > 0 \) and positive constants \( \tau_0, \epsilon' \) such that

\[
P_{\theta_0}(\|\theta_0 - \hat{\theta}\| \geq M\hat{\rho}/[\|X\|_{\text{max}}\phi_2((\tau_0 + 1)s(\theta_0))]) \leq \epsilon_1 + \exp\{-c''s(\theta_0)\log(\frac{ep}{s(\theta_0)})\},
\]

\[
P_{\theta_0}(\|\theta_0 - \hat{\theta}\| \geq M\sqrt{s(\theta_0)}\hat{\rho}/[\|X\|_{\text{max}}\phi_1((\tau_0 + 1)s(\theta_0))]) \leq \epsilon_1 + \exp\{-c''s(\theta_0)\log(\frac{ep}{s(\theta_0)})\},
\]

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

The first two claims can be made uniform over \( \theta_0 \in \Theta_{eb}(t, \tau) \cap \ell_0[s] \setminus \{0\} \) with \( s \geq 1: \)

\[
\sup_{\theta_0 \neq 0 \in \Theta_{eb}(t, \tau) \cap \ell_0[s]} P_{\theta_0}(\|\theta_0 - \hat{\theta}\| \geq M\hat{\rho}/[\|X\|_{\text{max}}\phi_2((\tau_0 + 1)s)]) \leq \epsilon_1 + \exp[-c''\log(ep)],
\]

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

Remark 4 (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 behavior 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 predictor X is obtained randomly, we first define the risk function and the oracle rates (see (2.4)) conditional on the realization of X. Let \( r(\theta|X) \) stand for the oracle rate at parameter \( \theta \) given X, and let \( r^2(\theta_0) = \mathbb{E}_X r^2(\theta_0|X) \), where \( \mathbb{E}_X \) stands for the expectation with respect to the distribution of the predictors. The second part of Theorem 1 then holds by using the corresponding assertion conditional on X and then integration with respect to X. The first now follows from the second by 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 \((t, \tau)\) and \( \epsilon > 0 \), we say that the \( \epsilon \)-EBR condition holds at \( \theta_0 \) if \( b_\epsilon(\theta_0) = b_\epsilon(\theta_0|X) \) defined by (2.20) satisfies \( \mathbb{P}_X(b_\epsilon(\theta_0|X) > t) < \epsilon \). Let the set of all parameter values satisfying the \( \epsilon \)-EBR condition with structural parameters \((t, \tau)\) be denoted by \( \Theta_{\epsilon-eb}(t, \tau) \). Then Theorem 2 can be stated as follows: If \( \tau \) satisfies (2.21), \( t > 0 \), \( \epsilon_1, \epsilon_2 > 0 \), there exist \( M = M(t, \epsilon_1) > 0 \) and \( L = L(\epsilon_2) > 0 \) such that

\[
\sup_{\theta_0 \in \Theta_{\epsilon-eb}(t, \tau)} \mathbb{P}_{\theta_0}(X \theta_0 \notin B(X \hat{\theta}, M \hat{\rho})) \leq \epsilon + \epsilon_1, \quad \sup_{\theta_0 \in \mathbb{R}^p} \mathbb{P}_{\theta_0}(\hat{\rho} \geq L r(\theta_0)) \leq \epsilon_2.
\]

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

### 3. Computation

The proposed empirical Bayes method can be computed using Monte Carlo sampling. Clearly the main challenge is the computation of \( \tilde{I} \) in (2.13), since once \( \tilde{I} \) is computed, posterior samples for \( \theta \) are easily obtained from (2.15). Since the number of possible index sets is exponential in \( p \), evaluation of the marginal likelihood of \( I \) is generally difficult. We propose evaluation of the marginal posterior \( \pi(I|Y) \) of \( I \) given the data by the following simulated annealing technique (see Kirkpatrick et al. [16]).

Recall that the simulated annealing technique finds a global maximizer of a positive function \( f \) on a domain through stochastic iterations. Like the Metropolis-Hastings algorithm, given a state \( x \), it proposes a move to a neighboring state \( x' \). If \( f(x') > f(x) \), the move is accepted. If \( f(x') \leq f(x) \), the move is accepted with some probability \( p(x'|x, T) \) depending on the function \( f \), the proposed value \( x' \), the current value \( x \) and a “temperature” parameter which goes to zero as the iteration progresses making \( p(x'|x, T) \) more and more “peaked” around \( x \). While various choices are possible, a common idea is to choose
cases (i) and (ii). \( H' \) is then easily obtained in terms of \( (X_I I) \) where
\[
X_I = c_i \text{ for } i \in I.
\]
Instead of choosing randomly (iii) always \( i \notin I \), we can choose the \( i \) cyclically from \( \{1, 2, \ldots, p\} \). Note that if \( I \cup \{i\} \notin \mathcal{I} \) and only if \( (1 - H_I)X_i = 0 \), and hence the condition can be easily checked. The proposed move will be rejected immediately if \( I \cup \{i\} \notin \mathcal{I} \). Suppose that we have computed \( (X_I X_i)^{-1} \) (and hence \( H_I \)) also. To develop a recursion for computing \( (X_I' X_i')^{-1} \) (and hence \( H_I' \)), we consider cases (i) and (ii).

For (i), we can write \( X_{I'} = (X_I, X_i) \). For an invertible block matrix \( M = \begin{pmatrix} A & b \\ b' & c \end{pmatrix} \),
\[
M^{-1} = \begin{pmatrix} A^{-1} + A^{-1}b'bA^{-1}/k & -A^{-1}b/k \\ -bA^{-1}/k & 1/k \end{pmatrix},
\]
where \( k = c - b'A^{-1}b \neq 0 \) since \( M \) is invertible. The inverse of \( X_{I'} X_{I'} = \begin{pmatrix} X_I X_I & X_I X_i \\ X_i X_I & X_i X_i \end{pmatrix} \),
is then easily obtained in terms of \( (X_I X_I)^{-1}, X_I X_i \text{ and } \|X_i\|^2 \), and we can verify that
\[
H_{I'} = H_I + \frac{1}{X_i(I - H_I)X_i} (I - H_I)X_i X_i(I - H_I),
\]
so that (3.23) can easily be evaluated in this case.

For the proposed move (ii), we have \( X_I = (X_{I'}, X_i) \). Denote \( X_i X_I = \begin{pmatrix} X_{I'} X_{I'} & X_{I'} X_i \\ X_i X_{I'} & X_i X_i \end{pmatrix} \).
\[
\begin{pmatrix} A & b' \\ b & c \end{pmatrix} = M \text{ and } (X_i X_i)^{-1} = M^{-1} = \begin{pmatrix} D & e' \\ e & f \end{pmatrix} \]
which is already computed. The goal is to obtain \( (X_{I'} X_{I'})^{-1} = A^{-1} \). The block-matrix inversion formula (3.24) yields the equations
\[
A^{-1} + A^{-1}b'bA^{-1}/k = D, -A^{-1}b/k = e \text{ and } f = 1/k. \]
The solutions are \( k = 1/f, b = - Ae/f \) and \( A^{-1} = D - ee'/f \). Thus \( (X_{I'} X_{I'})^{-1} = D - ee'/f \) is easily obtained and hence \( H_{I'} \) too.
If the alternative empirical Bayes posterior distribution in Remark 1 is to be computed, instead of maximizing the posterior distribution of $I$ given $Y$, one has to sample from it. This can be done by the Metropolis-Hastings algorithm based on the same proposed moves and a slight modification of the acceptance probabilities. In fact, if we fix the temperature parameter $T$ to 1 instead of decaying it slowly, then the successive iterations give Metropolis-Hastings samples from the posterior distribution of $I$ given $Y$, which can be collected after a sufficient burn-in. However, typically the Metropolis chain needs to be run longer compared with the number of iterations in simulated annealing. This is because in the latter method, the probability function gets more peaked and hence the iteration stabilizes more quickly.

4. Discussions and extensions. 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 [22], 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 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.

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 $H_{I_{-i}}$ be the projections onto the column spaces $\text{col}(X_I)$, $\text{col}(X_{I_i})$ and $\text{col}(X_{I_{-i}})$, respectively. Clearly, $H_{I_i} - H_I$ is the projection onto the orthogonal complement of $\text{col}(X_I)$ in $\text{col}(X_{I_i})$ which is zero if $X_i \in \text{col}(X_I)$. Otherwise, by some tedious direct calculations using (3.24),

$$H_{I_i} - H_I = h_{I_i}h'_{I_i}, \quad \text{where} \quad h_{I_i} = \frac{(I - H_I)X_i}{\|(I - H_I)X_i\|}.$$

Similarly, if $X_i \notin \text{col}(X_{I_{-i}})$, then

$$H_I - H_{I_{-i}} = h_{I_{-i}}h'_{I_{-i}}, \quad \text{where} \quad h_{I_{-i}} = \frac{(I - H_{I_{-i}})X_i}{\|(I - H_{I_{-i}})X_i\|}.$$

For any $i \in \{1, \ldots, p\}$, denote $I^{'}_{o,i} = I^{'}_o \cup \{i\}$. Then by the definition of the $\tau$-oracle we have $r^2_\tau(I^{'}_{o,i}, \theta) \geq r^2_\tau(I^{'}_o, \theta)$, yielding

$$\| (H_{I^{'}_{o,i}} - H_{I^{'}_o})X\theta \|^2 \leq \tau \sigma^2(|I^{'}_o| + 1) \log \frac{ep}{|I^{'}_o| + 1} - \tau \sigma^2 |I^{'}_o| \log \frac{ep}{|I^{'}_o|} \leq \tau \sigma^2 \log(ep).$$
On the other hand, if $|I^*_0| > 1$, the oracle coordinates must have a relatively significant contribution in the following sense. For any $i \in I^*_0$, let $I^*_{0-i} = I^*_0 \setminus \{i\}$, we have by the definition of the $\tau$-oracle that

$$r^2(I^*_{0-i}, \theta) \geq r^2(I^*_0, \theta),$$

yielding

$$\| (H_{I^*_0} - H_{I^*_{0-i}}) X \theta \|^2 \geq \tau \sigma^2 |I^*_0| \log \frac{ep}{|I^*_0|} - \tau \sigma^2 (|I^*_0| - 1) \log \frac{ep}{|I^*_0| - 1} \geq \tau \sigma^2 (\log(ep) + \log \frac{|I^*_0| - 1}{|I^*_0|}) \geq \tau \sigma^2 (\log(ep) - \log(\sqrt{2}(|I^*_0| - 1))).$$

If we compare the EBR parameters $\Theta_{eb}(t, \tau)$ with the traditional sparsity class $\ell_0[s] = \{ \theta \in \mathbb{R}^p : |I^*(\theta)| \leq s \}$, it is easy to see that a subset of $\ell_0[s]$ with prominent non-zero components $\hat{\ell}_0(s, \tau_1) = \{ \theta \in \ell_0[s] : I^*(\theta) = I^*_0(\theta) \}$ trivially satisfies EBR for $\tau = \tau_1$ and any $t \geq 0$. For example, if $\theta \in \ell_0[s]$ is such that $\| (H_J - H_{J-i}) X \theta \|^2 \geq \tau_1 \sigma^2 \log(ep)$ for all $i \in J$ for some $J \in \mathcal{I}$ with $|J| = s > 1$, then $\theta \in \hat{\ell}_0(s, \tau_1) \subseteq \Theta_{eb}(t, \tau_1)$ for any $t \geq 0$.

In this light, the EBR condition can be thought of as a more lenient type of sparsity on the parameter $\theta$, (to be called the **EBR-sparsity**), compared with the classical “conservative” sparsity requirement $|I^*(\theta)| \leq s$: the error the oracle makes when setting certain “insignificant” coordinates of $\theta$ to zero should not exceed (up to a constant multiple) the error made in recovering the contribution by “significant” non-zero coordinates. The parameter $t \geq 0$ 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 $\mathbb{R}^p = \cup_{t \geq 0} \Theta_{eb}(t, \tau)$ by varying the other structural parameter $t$ over the positive half line. Therefore in view of Theorem 2, at any true parameter $\theta_0$, an optimal order inflated credible ball has high coverage provided that the constant inflating factor is chosen sufficiently large.

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 more generally. 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 controlled 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.

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 (6.4). This restricts us to a normal-normal conjugate setting, and hence the specific choice of 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 Section 3. Formulating an analogous computational procedure for other priors, such as for the Laplace prior used by Castillo et al. [9] 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. [9].

It therefore 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 [3] 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 the set $\mathcal{I}$, the class of the 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 (6.8) 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.

5. 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 [15]. An additive nonparametric regression model for a response variable $Y$ based on $p$-dimensional random predictors $X = (X_1, \ldots, X_p)$, is given by the relation

$$Y = \mu + \sum_{i=1}^{p} f_i(X_i) + \varepsilon,$$

where $\mu$ is a real-valued parameter, $f_1, \ldots, f_p$ are functions with some smoothness properties and satisfy the identifiability restrictions $E f_i(X_i) = 0$, $i = 1, \ldots, p$, and $\varepsilon$ is a random error distributed as $N(0, \sigma^2)$. 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 [19], Ravikumar et al. [25], Lukas et al. [21], Raskutti et al. [24]] or by Bayesian variable selection [Curtis et al. [11], Yang and Tokdar [32]]. Representing each component function through basis expansions, the additive nonparametric regression model (5.1) can be reduced to a linear regression model after truncating the basis expansion and allowing for approximation bias. Below we sketch 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.

We assume that the $i$th predictor $X_i$ is distributed according to a density $q_i$ on its domain $D_i$ and different predictors are independently distributed. Let $\{B_{ij} : j = 1, 2, \ldots\}$ be an orthonormal basis of $L_2(D_i)$ and $f_i \in L_2(D_i)$. Denoting $Z_{ij} = B_{ij}(X_i)$, $j = 1, 2, \ldots$, we can rewrite the model (5.1) as $Y = \mu + \sum_{i=1}^{p} \sum_{j=1}^{\infty} \theta_{ij} Z_{ij} + \varepsilon = \mu + Z \theta + \varepsilon.$ Now, for every $I \subseteq \{1, \ldots, p\}$ and $J \subseteq \{1, 2, \ldots\}$, consider the truncated linear model

$$Y = \mu + \sum_{i \in I} \sum_{j \in J} \theta_{ij} Z_{ij} + \varepsilon = \mu + Z_{I,J} \theta_{I,J} + \varepsilon.$$ 

Let $(X^{(k)}, Y^{(k)})$, $k = 1, \ldots, n$, be the observed samples, $X = (X^{(1)}, \ldots, X^{(n)})', Y = (Y^{(1)}, \ldots, Y^{(n)})', Z^{(k)} = (1, B_{ij}(X_i^{(k)}) : i = 1, \ldots, p, j = 1, 2, \ldots)$ and $Z = (Z^{(1)}, \ldots, Z^{(n)})'$; here we have slightly abused the notation $Y$ to denote it the vector of the observed values of the response variable. For each $I \subseteq \{1, \ldots, p\}$ and $J \subseteq \{1, 2, \ldots\}$, define $Z_{I,J}$ to be the matrix with the $k$th row equal to $Z_{I,J}^{(k)} = (B_{ij}(X_i^{(k)}) : i \in I, j = 1, \ldots, J).$ Let $[J] = \{1, \ldots, J\}$. Observe that in the $(I, J)$-th model (5.2) the least square estimator $\hat{\theta}(I, J) = (\hat{\theta}_{ij}(I, J), i = 1, \ldots, p, j = 1, 2, \ldots)'$ of $\theta = (\theta_{ij}, i = 1, \ldots, p, j = 1, 2, \ldots)'$ is $\hat{\theta}_{I,J}(I, J) = (Z_{I,J}' Z_{I,J})^{-1} Z_{I,J}' Y$, $\hat{\theta}_{I,J'}(I, J) = 0$, $\hat{\theta}_{I,J'}(I, J) = 0$, $\hat{\theta}_{I,J'}(I, J) = 0$, provided that $(I, J) \in J = \{I', J' : \text{the columns of } Z_{I', J'} \text{ are linearly independent}\}$. In other words, this means that the estimator has the $(i, j)$-th component set to zero if either
As before, we define the oracle rate by
\[
\{ \text{the values of } r \} = \text{the empirical Bayes method to set it to the least square estimator } \hat{\theta} \text{ of the marginal posterior probability of } (I, J).
\]
and define the τ-mean parameter τ by
\[
\mu = 1 \in \mathbb{R}
\]
We consider the conjugate prior
\[
\theta_{1,[J]}(I, J) \sim \text{N}(\mu(I, J), \kappa \sigma^2(Z_{I,J}Z_{I,J})^{-1}), \quad \theta_{i,j} = 0 \quad \text{if } i \notin I \text{ or } j > J.
\]
On I, we put the prior given by (2.6), and independently we put a geometric prior \( \lambda(J = j) = (e^{\beta} - 1)e^{-\beta j} \) on \( J, j = 1, 2, \ldots, \beta > 0 \). As before, for each \( (I, J) \), \( \mu(I, J) \) is selected by the empirical Bayes method to set it to the least square estimator \( \hat{\theta}_{I,[J]}(I, J) \). The grand mean parameter \( \mu \) is also given a conjugate normal prior and the posterior is recentered at the sample mean \( \bar{Y} \) following the empirical Bayes selection. Finally, the maximizer \( (I, \hat{J}) \) of the marginal posterior probability of \( (I, J) \) is selected and the corresponding empirical Bayes posterior is denoted by \( \hat{\pi} \). The empirical Bayes posterior then satisfies concentration results similar to Theorem 1: There exists a constant \( C > 0 \) such that, for any \( \mu \in \mathbb{R} \) and \( f_{i,0} \in L_2(D_i) \) with \( \int_{D_i} f_{i,0}(x_i)q_i(x_i)dx_i = 0, \ i = 1, \ldots, p \), and \( M > 0 \), we have that
\[
E_{\mu_0, f_0}(\|Z_\theta - \mu_0 1_n - \sum_{i=1}^p F_{i,0}\| \geq Mr(\mu_0, f_0)|X, Y) \leq CM^{-2},
\]
where \( f_0 = (f_{1,0}, \ldots, f_{p,0}) \), \( 1_n = (1, \ldots, 1)' \) and \( F_{i,0} = (f_{i,0}(X_i^{(1)}), \ldots, f_{i,0}(X_i^{(n)}))' \), \( i = 1, \ldots, p \). Here the true expectation integrates out the \( X \)-values too as they are random; see Remark 4. If \( \hat{\theta} \) denotes the empirical Bayes posterior expectation of \( \theta \), then its convergence bound in \( L_2 \) is also given by
\[
E_{\mu_0, f_0}(\|Z_\hat{\theta} - \mu_0 1_n - \sum_{i=1}^p F_{i,0}\|^2 \leq Cr^2(\mu_0, f_0).
\]
The above concentration result is admittedly abstract, but it can easily lead to well-interpretable results when sparsity conditions for selection of the variables $X_1, \ldots, X_p$ in the true regression function and smoothness of each selected function are assumed. For this we need to bound the oracle risk 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. Let $s$ stand for the true number of active predictors and suppose that each true function $f_{i,0}, \ldots, f_{p,0}$ are Sobolev $\alpha$-smooth in the respective domain, where $\alpha$ is not known to us. Note that an inactive predictor $X_i$ in the true data-generating process corresponds to a true regression function $f_{i,0} = 0$, and hence trivially satisfies the smoothness condition. Now using a convenient orthonormal basis $\{B_1, B_2, \ldots\}$ such as that of Legendre polynomials or Fourier functions for all predictors, it follows that there exists a constant $K$ such that for any $J \geq 1$, there exist coefficients $\theta_{ij,0}, i = 1, \ldots, p$, $j = 1, \ldots, J$, with the approximation property

$$
E|f_{i,0}(X_i) - \sum_{j=1}^{J} \theta_{ij,0} B_j(X_i)|^2 = \sum_{j>J} \theta_{ij,0}^2 \leq KJ^{-2\alpha},
$$

$i = 1, \ldots, p$, uniformly for functions $f_{i,0}, \ldots, f_{p,0}$ lying in an $\alpha$-Sobolev ball. Let $I_0 = I_0(\mu_0, f_0)$ be the true set of active predictors and $J_0 = J_0(f_0)$ be the integer part of $n^{1/(1+2\alpha)}$. Clearly the oracle rate is bounded by the rate at $(I_0, J_0)$ and the latter is bounded as follows:

$$
\begin{align*}
r_1^2((I_0, J_0), (\mu_0, f_0)) & \leq nE\left[\sum_{i \in I_0} \sum_{j > J_0} \theta_{ij,0} B_j(X_i)\right]^2 + \sigma^2 s J_0 + \sigma^2 s \log \frac{ep}{s} \\
& = n \sum_{i \in I_0} E\left[\sum_{j > J_0} \theta_{ij,0} B_j(X_i)\right]^2 + \sigma^2 s J_0 + \sigma^2 s \log \frac{ep}{s} \\
& = n \sum_{i \in I_0} \sum_{j > J_0} \theta_{ij,0}^2 + \sigma^2 s J_0 + \sigma^2 s \log \frac{ep}{s} \\
& \leq ns((K + \sigma^2)n^{-2\alpha/(1+2\alpha)} + \frac{\sigma^2}{n} \log \frac{ep}{n})
\end{align*}
$$

using the independence of the components of the predictor, the norm-decreasing property of a projection operator, (5.5) and the choice of $J_0$. Hence the posterior contraction rate for the normalized distance $n^{-1} \sum_{k=1}^n |\mu + \sum_{i=1}^p f_i(X^{(k)}_i) - \mu_0 - \sum_{i=1}^p f_{i,0}(X^{(k)}_i)|^2$ is bounded by $\epsilon_n^2 = \max(sn^{-2\alpha/(1+2\alpha)}, s(\log p)/n)$ for the sparsity regime $s$ and the smoothness regime $\alpha$. The latter was shown to be the minimax rate for estimation in Raskutti et al. [24] and Yang and Tokdar [32]. Thus the empirical Bayes procedure achieves the optimal minimax rate of estimation in the stated regime. Yang and Tokdar [32] also showed, using the general theory of posterior contraction (see Ghosal and van der Vaart [14]), 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.

In addition, as a consequence of the oracle approach, we obtain that a credible ball with squared radius a large multiple of $n^{-1} \max(\hat{I}, \hat{J}, |\hat{I}| \log(ep/|\hat{I}|))$ has frequentist coverage at all true parameters which meet the $\epsilon$-EBR condition for a sufficiently small $\epsilon > 0$, and the size of the credible set is minimax optimal with high probability.

6. 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 $\eta_1, \ldots, \eta_N$ be nonnegative random variables and $t > 0$ be such that for all $i = 1, \ldots, N$, $E e^{t\eta_i} \leq A_t$ for some $0 < A_t < \infty$. Then $E \max_{1 \leq i \leq N} \eta_i \leq t^{-1} \log N + \log A_t$.

**Proof.** By Jensen’s inequality,

$$
\exp \left\{ t E \max_{1 \leq i \leq N} \eta_i \right\} \leq E \exp \left\{ t \max_{1 \leq i \leq N} \eta_i \right\} \leq \sum_{i=1}^{N} E e^{t\eta_i} \leq NA_t,
$$

which is equivalent with the assertion. \(\square\)

**Proof of Theorem 1.** According to the empirical Bayes posterior distribution $\hat{\pi}(\cdot|Y)$ given in (2.15), it follows that $X|\theta \sim N_n(H_{\hat{I}}Y, \kappa\sigma^2\kappa + 1H_{\hat{I}})$. Then as $H_{\hat{I}}(I - H_{\hat{I}}) = O$ for any $I$,

$$
\hat{E}(\|X\theta - X\theta_0\|^2|Y) = \frac{\kappa\sigma^2}{\kappa + 1} \text{tr}(H_{\hat{I}}) + \|H_{\hat{I}}Y - X\theta_0\|^2
$$

$$
= \frac{\kappa\sigma^2}{\kappa + 1} |\hat{I}| + \|H_{\hat{I}}(I - I)X_{\hat{I},\theta_0,\hat{I}} + \sigma H_{\hat{I}}\|\|^2 \leq r^2(I, \theta_0) + \sigma^2\|H_{\hat{I}}\|^2.
$$

Therefore by Markov’s inequality we obtain

$$
E_{\theta_0}\hat{\pi}(\|X\theta - X\theta_0\| \geq Mr(\theta_0)|Y) = E_{\theta_0} \sum_{I \in \mathcal{I}} \hat{\pi}(\|X\theta - X\theta_0\| \geq Mr(\theta_0)|Y) \mathbb{1}\{I = I\}
$$

$$
\leq E_{\theta_0} \sum_{I \in \mathcal{I}} \frac{\hat{E}(\|X\theta - X\theta_0\|^2|Y, I)}{M^2r^2(\theta_0)} \mathbb{1}\{I = I\}
$$

$$
\leq \sum_{I \in \mathcal{I}} \frac{r^2(I, \theta_0) \mathbb{P}_{\theta_0}(\hat{I} = I)}{M^2r^2(\theta_0)} + \frac{\sigma^2E_{\theta_0} \sum_{I \in \mathcal{I}} \|H_{\hat{I}}\|^2 \mathbb{1}\{I = I\}}{M^2r^2(\theta_0)}.
$$

(6.1)
Using (2.13) and (2.14), we obtain that, for any $h \in [0, 1]$ and any $I, I_0 \in \mathcal{I}$,

$$P_{\hat{\theta}_0}(\hat{I} = I) \leq P_{\theta_0}\left(\frac{\lambda_I I(Y)}{\lambda_{I_0} I_0(Y)} \geq 1\right)$$

$$\leq E_{\theta_0}\left[\frac{\lambda_I I(Y)}{\lambda_{I_0} I_0(Y)}\right]^h$$

$$= \left(\frac{\lambda_I}{\lambda_{I_0}}\right)^h E_{\theta_0}\left\{\frac{1}{2\pi} \frac{Y'(H_I - H_{I_0})Y}{(1 + \kappa)^{h(I - |I_0|)/2}}\right\}.$$  (6.2)

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

$$E \exp\{Y'AY/2\} = \frac{\exp\left\{\frac{1}{2} \mu'(I - A\Sigma)^{-1}A\mu\right\}}{\sqrt{\det(I - A\Sigma)}}.$$  (6.3)

Denote for brevity $H_{I,I_0} = H_I - H_{I_0}$. Using (6.2) and (6.3) with $\Sigma = \sigma^2 I$ and $A = \frac{h}{\sigma^2} H_{I,I_0}$, we obtain the bound

$$P_{\theta_0}(\hat{I} = I) \leq \left(\frac{\lambda_I}{\lambda_{I_0}}\right)^h E_{\theta_0}\left\{\frac{1}{2\pi} \frac{\theta'X'(I - hH_{I,I_0})^{-1}H_{I,I_0}X\theta}{(1 + \kappa)^{h(I - |I_0|)/2}}\right\},$$

provided that $I - hH_{I,I_0}$ is invertible.

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

$$(I - hH_{I,I_0})^{-1}H_{I,I_0} = H_{I,I_0} + hH_{I,I_0}(I - hH_{I,I_0})^{-1}H_{I,I_0}, \quad h \in [0, 1),$$

as $I - hH_{I,I_0}$ is invertible for any $h \in [0, 1)$. Denoting for brevity $y = X\theta$, we then derive

$$y'(I - hH_{I,I_0})^{-1}H_{I,I_0}y$$

$$\leq y'H_{I,I_0}y + h|y'H_{I,I_0}(I - hH_{I,I_0})^{-1}H_{I,I_0}y|$$

$$\leq \|(I - H_{I_0})y\|^2 - \|(I - H_I)y\|^2 + \frac{h}{1 - \kappa} \|H_{I,I_0}y\|^2$$

$$\leq \|(I - H_{I_0})y\|^2 - \|(I - H_I)y\|^2 + \frac{2h}{1 - \kappa} \|(I - H_I)y\|^2 + \|(I - H_{I_0})y\|^2$$

$$= \frac{1 + h}{1 - \kappa} \|(I - H_{I_0})y\|^2 - \frac{1 + 2h}{1 - \kappa} \|(I - H_I)y\|^2,$$  (6.5)

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

$$\det(I - h(H_I - H_{I_0})) \geq (1 - h)^{|I|}.$$  (6.6)

Inserting (2.6), (6.5) and (6.6) in (6.4), the bound for $P_{\theta_0}(\hat{I} = I)$ reduces to

$$(\varepsilon p/|I|)^{-c_1|I|}\exp\left\{\frac{h(1+h)}{2\pi^2(1-h)} \theta'X'(I - H_{I_0})X\theta + \varepsilon h |I_0| \log \frac{ep}{|I_0|} + \frac{h}{2}|I_0| \log(1 + \kappa)$$
By the Cauchy-Schwarz inequality and (6.9), (1 + \kappa)^{-1/2} \geq 1, h \in (0, 1/3) and \kappa > 2/h. Using this and applying the last relation with I_0 = I_o(\theta_0), we obtain

\begin{equation}
\mathcal{P}_{\theta_0}(\hat{I} = I) \leq (ep/|I|)^{-c_1|I|} \exp \left\{-c_2\sigma^{-2}(r^2(I, \theta_0) - c_3r^2(\theta_0))\right\},
\end{equation}

where

\begin{align*}
&c_1 = 1 + \frac{h\kappa}{2}, \quad c_2 = \min\left\{\frac{h(1 - 3h)}{2(1 - h)}, \frac{h\kappa}{2} - 1\right\}, \quad c_3 = \frac{h}{2c_2}\max\left\{1 + \frac{h}{1 - \frac{\kappa}{c_3}}, 2\kappa + \log(1 + \kappa)\right\}.
\end{align*}

By the assumed conditions (2.7) on \kappa, \kappa 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 \(\mathcal{O}(\tau_0, \theta_0) = \{I \in \mathcal{I} : r^2(I, \theta_0) \leq \tau_0r^2(\theta_0)\}\). Choosing \(\tau_0 > c_3\), where \(c_3\) is defined in (6.7), it follows that for any \(I \in \mathcal{O}^c(\tau_0, \theta_0)\), we have \(r^2(I, \theta_0) - c_3r^2(\theta_0) \geq (1 - c_3/\tau_0)r^2(I, \theta_0)\. Using the estimates \(xe^{-cx} \leq (ce)^{-1}\) for any \(x \geq 0, c > 0\) and \((\frac{c}{k}) \leq (ep/k)^{c}\), with \(B = c_2(\tau_0 - c_3)/(2\tau_0)\), we obtain from (6.7) that

\begin{align}
\sum_{I \in \mathcal{O}^c(\tau_0, \theta_0)} r^2(I, \theta_0) [\mathcal{P}_{\theta_0}(\hat{I} = I)]^{1/2} &\leq \sum_{I \in \mathcal{O}^c(\tau_0, \theta_0)} \frac{(ep/|I|)^{-c_1|I|/2}r^2(I, \theta_0)e^{-Br^2(I, \theta_0)/\sigma^2}}{Bc}\nonumber
\end{align}

\begin{equation}
\leq \frac{\sigma^2}{Be} \sum_{k=1}^p \left(\frac{p}{k}\right)^{c_1/k} \leq \frac{\sigma^2}{Be(e^{c_1/2} - 1)}. \tag{6.8}
\end{equation}

As \(\|H_I\|^{2} \sim \chi^2_{|I|}\), we have that \(\mathbb{E}\|H_I\|^{4} = |I|^2 + 2|I| \leq 3|I|^2 \leq 3\tau_0r^2(I, \theta_0)/\sigma^4\). Therefore by the Cauchy-Schwarz inequality and (6.8), we obtain

\begin{align}
\sigma^2\mathcal{E}_{\theta_0} \sum_{I \in \mathcal{O}^c(\tau_0, \theta_0)} \|H_I\|^2 1\{\hat{I} = I\} &\leq \sum_{I \in \mathcal{O}^c(\tau_0, \theta_0)} \sigma^2[\mathbb{E}\|H_I\|^4]^{1/2}[\mathcal{P}_{\theta_0}(\hat{I} = I)]^{1/2} 
onumber
\end{align}

\begin{align*}
&\leq \sqrt{3} \sum_{I \in \mathcal{O}^c(\tau_0, \theta_0)} r^2(I, \theta_0)[\mathcal{P}_{\theta_0}(\hat{I} = I)]^{1/2} 
onumber
&\leq \frac{\sqrt{3}\sigma^2}{Be(e^{c_1/2} - 1)}. \tag{6.9}
\end{align*}

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

\begin{equation}
\sigma^2m \log(ep/m) \leq \tau_0r^2(\theta_0), \tag{6.10}
\end{equation}

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

\begin{equation}
\sum_{I \in \mathcal{O}(\tau_0, \theta_0)} \|H_I\|^2 1\{\hat{I} = I\} \leq \max_{I \in \mathcal{O}(\tau_0, \theta_0)} \|H_I\|^2 = \max\{\|H_I\|^2 : I \in \mathcal{O}(\tau_0, \theta_0), |I| = m\}. \tag{6.10}
\end{equation}
Now $\|H_I \varepsilon\|^2 \sim \chi^2_{|I|}$, and hence for any $I \in \mathcal{I}$ with $|I| = m$, $Ee^{t\|H_I \varepsilon\|^2} = (1 - 2t)^{-m/2} \leq e^m$ provided that $t \leq (1 - e^{-2})/2 < 0.43$. As the cardinality of $\{I \in \mathcal{O} (\tau_0, \theta_0) : |I| = m\}$ is at most $(\frac{m}{e}) \leq (ep/m)^m$, it follows from Lemma 1 with $t = 0.4$ that

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

This leads to the bound

$$\sigma^2 E_{\theta_0} \sum_{I \in \mathcal{O} (\tau_0, \theta_0)} \|H_I \varepsilon\|^2 1\{\hat{I} = I\} \leq 5\sigma^2 m \log(ep/m) \leq 5\tau_0 r^2 (\theta_0).$$

Combining this with (6.9) yields a bound for the second term in (6.1):

(6.11) $$\sigma^2 E_{\theta_0} \sum_{I \in \mathcal{I}} \|H_I \varepsilon\|^2 1\{\hat{I} = I\} \leq 5\tau_0 r^2 (\theta_0) + \frac{\sqrt{3}\sigma^2}{Be(e^{(c_1-2)/2} - 1)}.$$

For the first term in (6.1), 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 $r^2 (I, \theta_0) \leq \tau_0 r^2 (\theta_0)$ and $\sum_{I \in \mathcal{I}} 1\{\hat{I} = I\} = 1$. For the sum over $\mathcal{O}^c (\tau_0, \theta_0)$, we import the bound in (6.8) by noting that $P_{\theta_0}(\hat{I} = I) \leq [P_{\theta_0}(\hat{I} = \hat{I})]^{1/2}$. Using these relations, we obtain

(6.12) $$\sum_{I \in \mathcal{I}} r^2 (I, \theta_0) P_{\theta_0} (\hat{I} = I) \leq \tau_0 r^2 (\theta_0) + \frac{\sigma^2}{Be(e^{(c_1-2)/2} - 1)}.$$

Note that $\sigma^2 \leq r^2 (\theta_0)$ by the definition of risk. Thus all terms in (6.11) and (6.12) can be bounded in terms of $r^2 (\theta_0)$ and hence the bound in (6.1) 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 (6.1) and the bounds (6.11) and (6.12), we get a slightly stronger conclusion:

$$E_{\theta_0} \hat{E}(\|\theta - \theta_0\|^2 | Y) = E_{\theta_0} \sum_{I \in \mathcal{I}} \hat{E}_I \|\theta - \theta_0\|^2 1\{\hat{I} = I\} \leq Cr^2 (\theta_0).$$

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

Proof of Corollary 1. Denote $\mathcal{G} = \{I : r^2 (I, \theta_0) \geq m_0 r^2 (\theta_0) + m\sigma^2\}$, for $m_0 = c_3$, where the constants $c_1 > 2$, $c_2 > 0$, $c_3 > 0$ are as in (6.7). Using (6.7) and the fact that $\sum_{I \in \mathcal{I}} (ep/|I|)^{-c_1 |I|} = (e^{c_1-1} - 1)^{-1} \leq 1$ for $c_1 > 2$, it follows that

(6.13) $$P_{\theta_0} (\hat{I} \in \mathcal{G}) = \sum_{I \in \mathcal{G}} P_{\theta_0} (\hat{I} = I) \leq e^{-c_2 m} \sum_{I \in \mathcal{I}} (ep/|I|)^{-c_1 |I|} \leq e^{-c_2 m},$$
which leads to 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), \( |I| \geq \tau_0's(\theta_0) \) implies that

\[
r^2(I, \theta_0) \geq \sigma^2|I| \log(ep/|I|) \geq \tau_0'\sigma^2 s(\theta_0)[\log(ep/s(\theta_0)) - \log \tau_0'] \geq \frac{\tau_0'}{2}\sigma^2 s(\theta_0) \log(ep/s(\theta_0))
\]

provided that \( s(\theta_0) < ep/(\tau_0')^2 \). Using (2.16), the relation above implies that \( r^2(I, \theta_0) \geq m_0\sigma^2 + m \sigma^2 \), 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 \( c' = \tau_0'/2 - m_0 \) whenever \( s(\theta_0) < ep/(\tau_0')^2 \), 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 \( \theta_0 \).

**Proof of Corollary 2.** By the definition of the compatibility coefficient, on models \( I \) with \( |I| \leq \tau_0s(\theta_0) \), \( ||\theta - \theta_0||_1 \) is bounded by \( \sqrt{(\tau_0 + 1)s(\theta_0)||X(\theta - \theta_0)||/||X||_{\max}\phi_1((\tau_0 + 1)s(\theta_0))} \) since the cardinality of \( \text{supp}(\theta - \theta_0) \) is at most \((\tau_0 + 1)s(\theta_0) \). By Theorem 1, \( ||X(\theta - \theta_0)|| > Mr(\theta_0) \) has probability bounded by \( CM^{-2} \) while by Corollary 1, the event \( \{\hat{I} \geq \tau_0s(\theta_0)\} \) has probability bounded by \( \exp[-c's(\theta_0) \log(ep/s(\theta_0))] \). Thus the first assertion follows. The proof of the second one is similar.

**Proof of Corollary 3.** If \( |\theta_{0j}| > M\sigma||X||_{\max}^{-1}s(\theta_0)/(\tau_0 + 1)s(\theta_0) \) and the posterior does not select the \( j \)-th predictor (i.e., sets \( \theta_j = 0 \)), then clearly \( ||\theta - \theta_0|| \geq |\theta_{0j}| \geq M\sigma(1 + \kappa)s(\theta_0) \) by (2.16). 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 \( \theta \) chosen from the posterior can rarely be much smaller than the cardinality of a \( \tau \)-oracle \( I_0^* (\theta_0) \) defined by (2.4) for some sufficiently large \( \tau > 0 \).

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

\[
P_{\theta_0}(\hat{I} \leq \varrho|I_0^* (\theta_0)) \leq \left(e^{\varrho-1} - 1\right)^{-1} \exp\left\{ -\alpha|I_0^* (\theta_0)| \log(ep/|I_0^* (\theta_0)|) \right\},
\]

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

In particular, if \( \tau > \bar{\tau}(e^{-1}) \), then, for all \( \varrho \in [0, e^{-1}] \), \( P_{\theta_0}(\hat{I} \leq \varrho|I_0^* (\theta_0)) \leq \varrho^{\alpha_0} \leq \varrho^{\alpha_0} \), where \( \alpha_0 = \alpha(\tau, e^{-1}) > 0 \).

**Proof.** For each \( \theta_0 \in \mathbb{R}^p \) and \( I \in \mathcal{I} \) such that \( |I| \leq \varrho |I_0^*| \), let \( I_0 = I \cup I_0^* \) for some \( I_0^* \subset I^* \) such that \( I_0 \in \mathcal{I} \) and \( \text{col}(X_I) + \text{col}(X_{I_0^*}) = \text{col}(X_{I_0}) \). For brevity we have suppressed the dependence of \( I_0 \) on \( I \) and \( \theta_0 \) and the dependence of \( I_0^* = I_0^*(\theta_0) \) on \( \theta_0 \). Clearly,

\[
(6.14) \quad \max(|I|, |I_0^*|) \leq |I_0| \leq |I| + |I_0^*| \leq (1 + \varrho)|I_0^*|.
\]
As the function $x \mapsto x \log(ep/x)$ is increasing on $[0,p]$, for $|I| \leq \varrho |I_0^x|$,  
\begin{equation} 
|I| \log \frac{ep}{|I|} \leq \varrho |I_0^x| \log \frac{ep}{\varrho |I_0^x|} \leq \varrho (1 + \log(1/\varrho)) |I_0^x| \log \frac{ep}{|I_0^x|}. 
\end{equation}  

Because $\text{col}(X_{I_0})$ contains $\text{col}(X_{I_0^x})$, the difference $H_{I_0} - H_{I_0^x}$ of the corresponding projections $H_{I_0}$ and $H_{I_0^x}$ is nonnegative definite. Therefore  
\[ \sigma^{-2}\theta'X'(H_{I_0} - H_{I})X\theta = \sigma^{-2}(\theta'X'(I - H_{I})X\theta - \theta'X'(I - H_{I_0})X\theta) \]  
\[ \geq \sigma^{-2}(\theta'X'(I - H_{I})X\theta - \theta'X'(I - H_{I_0})X\theta) \geq \tau \left( |I_0^x| \log \frac{ep}{|I_0^x|} - |I| \log \frac{ep}{|I|} \right), \]  
where the last relation holds because $r_2^2(I_0^x, \theta_0) \leq r_2^2(I, \theta_0)$ by the definition of $I_0^x$ as the minimizer of (2.4). In view of (6.15), this gives the bound  
\begin{equation} 
\sigma^{-2}\theta'X'(H_{I_0} - H_{I})X\theta \geq \tau \alpha(\varrho) |I_0^x| \log(ep/|I_0^x|), 
\end{equation}  
where $\alpha(\varrho) = 1 - \varrho(1 + \log(1/\varrho))$.  

Since $H_{I_0} - H_{I}$ is also a projection, $(I + H_{I_0} - H_{I})^{-1} = I - \frac{1}{2}(H_{I_0} - H_{I}).$ With $h = 1$ and our choice for $I_0$ in (6.4), the expression for $\lambda_I$, (6.14) and (6.16), it follows that for  
\[ |I| \leq \varrho |I_0^x|, \]  
$P_0_0(\hat{I} = I)$ is bounded by  
\[ \frac{\lambda_I}{\lambda_{I_0}} E_{\theta_0} \exp \left\{ -\frac{1}{2\varrho^2} Y'(H_{I_0} - H_{I})Y \right\} \]  
\[ \leq \frac{\lambda_I}{\varrho \sigma} \exp \left\{ -\frac{1}{2} \tau \alpha(\varrho) |I_0^x| \log \frac{ep}{|I_0^x|} + \alpha(1 + \varrho) \log \frac{ep}{|I_0^x|} + \frac{1}{2} \log(1 + \varrho) |I_0^x| \log \frac{ep}{|I_0^x|} \right\} \]  
\[ \leq (e^{\varrho - 1} - 1)^{-1} \lambda_I \exp \left\{ \left[ -\frac{1}{4} \tau \alpha(\varrho) + \frac{1}{4} \log(1 + \varrho) \right] |I_0^x| \log \frac{ep}{|I_0^x|} \right\} \].  
As $\sum_I \lambda_I = 1$ and $(e^{\varrho - 1} - 1)^{-1} \leq 1$ since $\varrho > 2$ by the choice in (2.7), the result follows by summing over $I \in \mathcal{I}$: $|\hat{I}| \leq \varrho |I_0^x|$.  

To show the second assertion of the lemma, consider two cases $|I_0^x| \leq \log \varrho^{-1}$ and $|I_0^x| > \log \varrho^{-1}$, with $\varrho \in [0,e^{-1}]$. In the first case, as $\varrho \leq e^{-1}$, we have  
\[ P_0_0(\hat{I} \leq \varrho |I_0^x|) \leq P_0_0(|\hat{I}| \leq \varrho \log(1/\varrho)) \leq P_0_0(|\hat{I}| \leq e^{-1}) = 0. \]  
Consider the second case $|I_0^x| > \log \varrho^{-1}$. Since $\alpha = \alpha(\tau, \varrho)$ is a decreasing function of $\varrho \geq 0$, we have $\alpha_0 = \alpha(\tau, e^{-1}) \leq \alpha(\tau, \varrho)$ for each $\varrho \in [0,e^{-1}]$. Then by the first assertion of the lemma and $|I_0^x| > \log \varrho^{-1}$,  
\[ P_0_0(\hat{I} < \varrho |I_0^x|) \leq e^{-\alpha(\tau, \varrho) |I_0^x| \log(ep/|I_0^x|)} \leq e^{-\alpha_0 |I_0^x|} \leq \varrho^{\alpha_0}. \]  
\[ \square \]
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 $M > 0$,

$$\text{P}_{\theta_0}(\|X\theta_0 - X\hat{\theta}\| \geq Mr(\theta_0)) \leq CM^{-2}. \quad (6.17)$$

Since $\tau \geq 1$ and $\theta_0 \in \Theta_{eb}(t, \tau)$, it follows that $r^2(\theta_0) \leq \tau^2(\theta_0) \leq (1+t)\sigma^2|I_0'| \log(ep/|I_0'|)$, where $I_0' = I_0'(\theta_0)$. This combined with the definition (2.22) of $\hat{\rho}$ and (6.17), leads to

$$P_{\theta_0}(X\theta_0 \notin B(X\hat{\theta}, M\hat{\rho})) \leq P_{\theta_0}(\|X\theta_0 - X\hat{\theta}\| > M\hat{\rho}, \hat{\rho} \geq \delta r(\theta_0)) + P_{\theta_0}(\hat{\rho} < \delta r(\theta_0))$$

$$\leq P_{\theta_0}(\|X\theta_0 - X\hat{\theta}\| > M\delta r(\theta_0)) + P_{\theta_0}\left(\hat{I}|\log \frac{ep}{|I|} < \delta(1+t)|I_0'| \log \frac{ep}{|I_0'|}\right)$$

$$\leq \frac{C}{M^2 \delta^2} + P_{\theta_0}\left(\hat{I}|\log \frac{ep}{|I|} < \delta(1+t)|I_0'| \log \frac{ep}{|I_0'|}\right).$$

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

$$P_{\theta_0}\left(\hat{I}|\log \frac{ep}{|I|} < \delta|I_0'| \log \frac{ep}{|I_0'|}\right) \leq P_{\theta_0}\left(\hat{I}|\log \frac{ep}{|I|} < \delta|I_0'| \log \frac{ep}{\delta|I_0'|}\right) = P_{\theta_0}(\hat{I} \leq \delta|I_0'|).$$

The second assertion of Lemma 2 gives $P_{\theta_0}(\hat{I} \leq \delta|I_0'|) \leq (\delta(1+t))^{\alpha_0}$ for all $\delta \leq 1/(e(1+t))$. This and the last two displays entail that, for $\delta \leq 1/(e(1+t))$, 

$$P_{\theta_0}(X\theta_0 \notin B(X\hat{\theta}, M\hat{\rho})) \leq \frac{C}{M^2 \delta^2} + (\delta(1+t))^{\alpha_0},$$

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

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

This establishes the coverage relation as the right hand side can be bounded by $\epsilon_1$ uniformly in $\theta_0 \in \Theta_{eb}(t, \tau)$ by choosing $M$ sufficiently large (depending on $t$ and $\epsilon_1$ only).

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

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

From (6.7) and the last relation it follows that for any $I \in G(L)$

$$P_{\theta_0}(\hat{I} = I) \leq (ep/|I|)^{-c_1 |I|} \exp\{ - c_2 (L - c_3) \sigma^{-2} r^2(\theta_0) \}.$$


Note that $r^2(\theta_0) \geq \tau^{-1}r^2_\circ(\theta_0) \geq \rho^2|I^*_\circ| \log(ep/|I^*_\circ|)$. Using the last relation and reasoning as in (6.13), we derive that, for any $\theta_0 \in \mathbb{R}^p$,

$$
P_{\theta_0}(\hat{\rho} \geq Lr(\theta_0)) = \sum_{I \in \mathcal{G}(L, \theta_0)} P_{\theta_0}(\hat{I} = I) \leq \exp\{-c_2(L^2 - c_3)\sigma^{-2}r^2(\theta_0)\} \sum_{I \in \mathcal{I}} (ep/|I|)^{-c_1|I|} \leq \exp\{-c_2(L^2 - c_3)|I_0| \log(ep/|I_0|)\} \leq e^{c_2c_3} \exp\{-c_2L^2\},$$

for all $L \geq L_0 = \sqrt{c_3}$. Clearly, a sufficiently large $L \geq L_0$ makes the bound smaller than any given $\epsilon_2 > 0$. \qed

**Proof of Corollary 4.** Note that $\hat{\theta}$ has support $\hat{I}$. By part (ii) of Corollary 1, $\hat{I}$ has cardinality $|\hat{I}| \leq \tau_0 s(\theta_0) \leq \tau_0 s$ 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 positive 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 $\phi_2$, it follows that $\|X(\hat{\theta} - \theta_0)\| \geq \phi_2((\tau_0 + 1)s(\theta_0))\|X\|_{\max} \|\hat{\theta} - \theta_0\|$. This immediately leads to the first claim of the corollary. The proof for the $\ell_1$-case is similar. The last relation is from Theorem 2. \qed

**Remark 5.** To prove (5.4), we can proceed as in the proof of Theorem 1 with appropriate modifications of the estimates. In this case, the projection $H_{I,J}$ involves two indexes $I$ and $J$, of which $J$ is linearly ordered. This is the reason why the prior for $J$ does not need an extra logarithmic factor in the exponent. A careful examination of the expression preceding (6.7) shows the occurrences of the terms that are constant multiples of $|I| \log(ep/|I|), |I_0| \log(ep/|I_0|), J$ and $J_0$, and can be condensed to a bound

$$P_{\mu_0, f_0}((\hat{I}, \hat{J}) = (I, J)) \leq \exp\{-c_1|I| \log \frac{ep}{|I|} - c_2J - c_3[r^2((I, J), (\mu_0, f_0)) - c_4r^2(\mu_0, f_0)]\}$$

for some appropriate constants $c_1, c_2, c_3, c_4$. The rest of the arguments can then go through in a similar manner.

A similar remark applies to the chain of estimates following (6.16), which allows an estimate of the form

$$P_{\theta_0}(|\hat{I}| \leq \varrho[I^*_\circ(\mu_0, f_0)]) + P_{\theta_0}(\hat{J} \leq \varrho J^*_\circ(f_0)) \leq C \exp\{-\alpha[I^*_\circ(\mu_0, f_0)] \log(ep/|I^*_\circ(\mu_0, f_0)|)\} + C \exp\{-\beta J^*_\circ(f_0)\},$$

making the remaining steps in the proof of Theorem 2 go through under the $\epsilon$-EBR condition, as mentioned in Remark 4.
REFERENCES


---

**Department of Mathematics**  
Vrije Universiteit Amsterdam  
The Netherlands  
E-mail: e.n.belitser@vu.nl

**Department of Statistics**  
North Carolina State University  
Raleigh, North Carolina 27695-8203  
U.S.A.  
E-mail: sghosal@stat.ncsu.edu