Electronic Journal of Statistics
Vol. 8 (2014) 328–354
ISSN: 1935-7524
DOI: 10.1214/14-EJS883

Estimation and variable selection with exponential weights

Ery Arias-Castro

Department of Mathematics
University of California, San Diego
9500 Gilman Drive
La Jolla, California 92093-0112
USA
e-mail: eariasca@ucsd.edu

and

Karim Lounici*

School of Mathematics
Georgia Institute of Technology
Atlanta, Georgia 30332-0160
USA
e-mail: klounici@math.gatech.edu

Abstract: In the context of a linear model with a sparse coefficient vector, exponential weights methods have been shown to be achieve oracle inequalities for denoising/prediction. We show that such methods also succeed at variable selection and estimation under the near minimum condition on the design matrix, instead of much stronger assumptions required by other methods such as the Lasso or the Dantzig Selector. The same analysis yields consistency results for Bayesian methods and BIC-type variable selection under similar conditions.

MSC 2010 subject classifications: Primary 62J99.
Keywords and phrases: Estimation, variable selection, model selection, sparse linear model, exponential weights, Gibbs sampler, identifiability condition.

Received October 2013.

Contents

*Supported in part by NSF Grant DMS-11-06644 and Simons Foundation Grant 209842.

1. Introduction

Consider the standard linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_\star + \boldsymbol{z}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector; $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the regression (or design) matrix, assumed to have normalized columns; $\boldsymbol{\beta}_\star \in \mathbb{R}^p$ is the coefficient vector; and $\boldsymbol{z} \in \mathbb{R}^n$ is white Gaussian noise, i.e., $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$. As in general the model (1) is not identifiable, we let $\boldsymbol{\beta}_\star$ denote one of the coefficient vectors of minimal support size such that $\boldsymbol{X}\boldsymbol{\beta} = \mathbb{E}(\boldsymbol{y})$. Then $J_\star$ and $s_\star$ denote the support and support size of $\boldsymbol{\beta}_\star$. We are most interested in the case where the coefficient vector is sparse, meaning $s_\star$ is much smaller than p. As usual, we want to perform inference based on the design matrix $\boldsymbol{X}$ and the response vector $\boldsymbol{y}$. The four main inference problems are:

- *Denoising:* estimate the mean response vector $\boldsymbol{X}\boldsymbol{\beta}_\star$;
- *Prediction:* estimate $\mathbf{U}\boldsymbol{\beta}_\star$ for a new observation $\mathbf{U} \in \mathbb{R}^p$;
- *Estimation:* estimate the coefficient vector $\boldsymbol{\beta}_\star$;
- *Support recovery:* estimate the support $J_\star$.

These problems are not always differentiated and often referred to jointly as *variable/model selection* in the statistics literature, and *feature selection* in the machine learning literature. Being central to statistics, a large number of papers address these problems. We review the literature with particular emphasis on papers that advanced the theory of model selection. We find [38], who provides necessary conditions and sufficient conditions under which the AIC/Mallows' C_p criteria and the BIC criteria are consistent. For example, AIC/Mallows'

C_p are consistent when there is a unique $\boldsymbol{\beta}$ such that $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$, and this $\boldsymbol{\beta}$ has a support of fixed size as $n, p \to \infty$. Also, BIC is consistent when the dimension p is fixed and the model is identifiable — a condition that appears to be missing in that paper. BIC was recently shown in [14] to be consistent when the model is identifiable, $p = O(n^a)$ with $a < 1/2$ and the true coefficient vector has a support of fixed size as $n, p \to \infty$. They also propose an extended BIC for when a is larger. Closely related is the work of [1], who consider the maximum a posteriori (MAP) estimator for essentially the same sparsity prior. They derive the *denoising* performance of this estimator, together with sharp minimax prediction oracle inequalities for the sparse regression model (1).

Assuming the size of the support of $\boldsymbol{\beta}_\star$ is known, [34] establish *denoising* and *estimation* performance bounds for best subset selection, and obtain information bounds for these problems. Relaxing to the ℓ_1-norm penalty, the Lasso and the closely related Dantzig Selector were shown to be consistent when the design matrix satisfies a restricted isometric property (RIP) or has column vectors with low coherence; see [4, 33, 6, 29, 8, 49, 12, 10] among others. [45] used a different set of conditions called cone invertibility factors or also sensitivity conditions and established oracle inequalities for the *estimation* problem with the Lasso and Dantzig Selector. [22] also exploited this approach to build computationally tractable confidence bands for the Dantzig Selector. With a carefully chosen nonconcave penalty, [21] shows that consistent variable selection is possible when $p = O(n^{1/3})$. This condition on p was weaken in the follow-up paper [20], though with an additional restriction on the coherence (Condition (16) there). The strongest results in that line of work seem to appear in [47, 48], which suggests a minimax concave penalty that leads to consistent variable selection under much weaker assumptions. The classical forward stepwise selection, also known as orthogonal matching pursuit, which is shown in [9] to enable variable selection under an assumption of low coherence on the design matrix. Screening was studied in [19] in the ultra high-dimensional setting, assuming the design is random. A combination of screening and penalized regression is explored in [24, 25], with asymptotic optimality when the Gram matrix $\boldsymbol{X}^\top \boldsymbol{X}/n$ is (mildly) sparse.

A distinct line of research considered the use of exponential weighting in high-dimensional *denoising/prediction* problems [7, 13, 43, 18, 23, 26, 28, 16, 17]. This methodology has the potential of striking a good compromise between statistical accuracy and computational complexity. While computational tractability has only been demonstrated in simulations, a number of sharp statistical results exist for the *denoising/prediction* problems. In particular, [2, 35] propose exponential weights procedures that achieve sharp sparsity oracle inequalities with no assumptions of the design matrix $\boldsymbol{X}$. For a recent survey of the exponential weights literature, see [36]. We emphasize that there exists no result in the literature concerning the *estimation* and *support recovery* problems with an exponential weights approach and that our results are the first of this nature.

Our contribution is the following. We establish performance bounds for the version of exponential weights studied in [2] for the three inference problems of *denoising*, *estimation* and *support recovery*. The methodology developed in the present paper is new and brings novel and interesting results to the sparse

regression literature. The main feature of this methodology is that it only requires comparatively almost minimum assumptions on the design matrix $\boldsymbol{X}$ and the target $\boldsymbol{\beta}_\star$. In particular, for *estimation* and *support recovery*, the conditions are slightly stronger than identifiability. Moreover, when the size of support is known, the exponential weights method is consistent under the minimum identifiability condition as long as the nonzero coefficients are large enough, close in magnitude to what is required by any method, in particular matching the performance of best subset selection [34]. See also [41, 11, 46].

The rest of the paper is organized as follows. In Section 2, we describe in detail the methodology and state the main results concerning estimation and support recovery with our exponential weights procedure. We also apply our methodology to establish novel results in estimation and support recovery for the Bayesian MAP estimator studied in [1], thus completing the theoretical study of the MAP. In Section 3, we present further results and establish some connections with the Bayesian theory and the BIC estimator. We also compare the results we obtained for exponential weights with those established for other methods, in particular the Lasso and MCP. In Section 4, we discuss our results in the light of recent information bounds for model selection. The proofs of our main results are in Section 5.

2. Main results

We consider the version of exponential weights studied in [2], shown there to enjoy optimal oracle performance for the *denoising* problem. The procedure puts a sparsity prior on the coefficient vector and selects the estimates using the posterior distribution. We obtain a new *denoising* performance bound which is based on balancing the sparsity level and the size of the least squares residuals. The result does not assume any conditions on the design matrix. The task of *support recovery*, to be amenable, necessitates additional assumptions. We show that under near-identifiablity conditions on the design matrix, the posterior concentrates on the correct subset of nonzero components with overwhelming probability, provided that these coefficients are sufficiently large — somewhat larger than the noise level. This immediately implies that the maximum a posteriori (MAP) and the randomized exponential weights estimator (REW) are consistent. We then derive *estimation* performance guarantees in Euclidean norm and l_∞-norm for the MAP, REW and the averaged exponential weights estimator (AEW) that can also be interpreted as the posterior mean by analogy with the bayesian theory.

Throughout, we assume the noise variance σ^2 is known. We also assume that $p \geq n$ and remark that similar results hold when $n \geq p$, with p replaced by n in the bounds.

We use some standard notation. For any $\boldsymbol{u} = (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_d)^\top \in \mathbb{R}^d$ with $d \geq 1$ and $q \geq 1$, we define

$$\|\boldsymbol{u}\|_q = \left(\sum_{j=1}^{d} \boldsymbol{u}_j\right)^{1/q}, \quad \|\boldsymbol{u}\|_\infty = \max_{1 \leq j \leq d} |\boldsymbol{u}_j|.$$

Without loss of generality, we assume from now on that the predictors are normalized in the sense that

$$\frac{1}{\sqrt{n}}\|\boldsymbol{X}_j\|_2 = 1, \text{ for all } 1 \le j \le p. \tag{2}$$

For a subset $J \subset [p] := \{1, \dots, p\}$, let $\boldsymbol{X}_J = [\boldsymbol{X}_j, j \in J] \in \mathbb{R}^{n \times |J|}$, where $\boldsymbol{X}_j$ denotes the jth column vector of $\boldsymbol{X}$. For a subset $J \subset [p]$, let M_J be the linear span of $\{\boldsymbol{X}_j, j \in J\}$ and let $\boldsymbol{P}_J$ be the orthogonal projection onto M_J. Then, $\boldsymbol{P}_J^\perp := \boldsymbol{I}_n - \boldsymbol{P}_J$ is the orthogonal projection onto $M_J^\perp$. We say that a vector is s-sparse if its support is of size s.

2.1. Exponential weights

We start with the definition of a sparsity prior on the subsets of $[p]$, which favors subsets with small support. This leads to a pseudo-posterior, which is used in turn to define various exponential weights estimators.

- *The prior* π. Fix an upper bound $\overline{s} \ge 1$ on the support size, and a sparsity parameter $\lambda > 0$. The prior chooses the subset $J \subset [p]$ with probability

$$\pi(J) \propto \binom{p}{|J|}^{-1} e^{-\lambda |J|} \mathbb{1}_{\{|J| \le \overline{s}\}}. \tag{3}$$

- *The posterior* Π. Given that the noise is assumed i.i.d. Gaussian with variance σ^2, given a subset of variables $J \subset [p]$, the coefficient vector that maximizes the likelihood is the least squares estimate $\widehat{\boldsymbol{\beta}}_J$ with a maximum proportional to $\exp(-\|\boldsymbol{P}_J^\perp(\boldsymbol{y})\|_2^2/(2\sigma^2))$. In light of this, we define the following pseudo-posterior, which chooses $J \subset [p]$ with probability

$$\Pi(J) \propto \pi(J) \, \exp\left(-\frac{\|\boldsymbol{P}_J^\perp(\boldsymbol{y})\|_2^2}{2\sigma^2}\right). \tag{4}$$

The prior π enforces sparsity and focuses on subsets of size not exceeding $\overline{s}$. Without additional knowledge, we shall take $\overline{s} = p$. The exponential factor in $\|\boldsymbol{P}_J^\perp(\boldsymbol{y})\|_2^2$ in the posterior enforces fidelity to the observations. Note that Π is not a true posterior because no prior is assumed for $\boldsymbol{\beta}_\star$; we elaborate on this point in Section 3.2. The variance term $2\sigma^2$ corresponds to the temperature T in a standard Gibbs distribution. We will calibrate the procedure via the sparsity exponent λ in (3), though we could have done so via the temperature as well. Remember that we assume that σ^2 is known. When the variance is unknown, we can replace it with a consistent estimator $\hat{\sigma}^2$.

Based on the pseudo-prior Π, it is natural to consider the maximum a posteriori (MAP) support estimate, defined as

$$\widehat{J}_{\text{map}} = \arg\max_J \Pi(J). \tag{5}$$

This leads to considering the MAP coefficient estimate. For any $J \subset [p]$, let $\widehat{\boldsymbol{\beta}}_J$ denote the the least squares coefficient vector for the sub-model $(\boldsymbol{X}_J, \boldsymbol{y})$ with minimum Euclidean norm — so that $\widehat{\boldsymbol{\beta}}_J$ is unique even when the columns of $\boldsymbol{X}_J$ are linearly dependent. When the columns of $\boldsymbol{X}_J$ are linearly independent, the standard formula applies

$$\widehat{\boldsymbol{\beta}}_J = (\boldsymbol{X}_J^\top \boldsymbol{X}_J)^{-1} \boldsymbol{X}_J^\top \boldsymbol{y}. \tag{6}$$

The MAP coefficient estimate is then defined as $\widehat{\boldsymbol{\beta}}_{\mathrm{map}} = \widehat{\boldsymbol{\beta}}_{\widehat{J}_{\mathrm{map}}}$.

We can also consider a randomized version of the exponential weights (REW) defined as follows

$$\widehat{\boldsymbol{\beta}}_{\mathrm{rew}} = \widehat{\boldsymbol{\beta}}_{\widehat{J}_{\mathrm{rew}}}, \quad \widehat{J}_{\mathrm{rew}} \sim \Pi. \tag{7}$$

In words, we first draw a subset of variables $\hat{J}$ according to the posterior Π and we compute the standard least squares estimator in the corresponding submodel $(\boldsymbol{X}_{\hat{J}}, \boldsymbol{y})$.

We also want to study the averaged exponential weights estimator (AEW)

$$\widehat{\boldsymbol{\beta}}_{\mathrm{aew}} = \sum_J \Pi(J) \widehat{\boldsymbol{\beta}}_J. \tag{8}$$

In the rest of this section, we establish *denoising*, *support recovery* and *estimation* oracle inequalities for the MAP, REW and AEW estimators.

2.2. A concentration result for the posterior

Our performance bounds for support recovery rely, as they should, on concentration properties of the posterior Π. We first prove that, without any condition on the design matrix $\boldsymbol{X}$, the posterior Π concentrates on subsets of small size.

Proposition 1. *Consider a design matrix $\boldsymbol{X}$ with $p \geq n$ and normalized column vectors (2). For some $\varepsilon > 0$ and $c \geq 1$, take*

$$\lambda = \frac{1+\varepsilon}{\varepsilon}(23 + 5c) \log p. \tag{9}$$

Then, with probability at least $1 - 2p^{-c}$, $\Pi(J) < \Pi(J_\star)$ for all $J \subset [p]$ such that $|J| > (1+\varepsilon)s_\star$, and in fact

$$\Pi\left(J : |J| > (1+\varepsilon)s_\star\right) \leq 4p^{-c}\, \Pi(J_\star). \tag{10}$$

2.3. Identifiability

Actual support recovery requires some additional conditions, the bare minimum being that the model is identifiable.

Condition $\mathbf{I}(s)$: For any subset $J \subset \{1, \dots, p\}$ of size $|J| \leq s$, the submatrix $\boldsymbol{X}_J$ is full-rank.

This condition characterizes the identifiability of the model as stated in the following simple result.

Lemma 1. *Assuming $\boldsymbol{\beta}_\star \in \mathbb{R}^p$ is $s_\star$-sparse, it is identifiable if, and only if, $\mathbf{I}(2s_\star)$ is satisfied.*

In this paper, we establish that exponential weights, and also ℓ_0-penalized variable selection, allow for support recovery and estimation under the condition $\mathbf{I}((2+\varepsilon)s_\star)$ for any $\varepsilon > 0$ fixed, as long as the non-zero entries of the coefficient vector are sufficiently large. In fact, $\mathbf{I}(2s_\star)$ suffices when $s_\star$ is known.

While $\mathbf{I}(s)$ is qualitative, results on estimation and support recovery necessarily require a quantitative measure of correlation in the covariates. The following quantity appears in the performance bounds we derive for exponential weights and related methods: for any integer $s \geq 1$, define

$$\nu_s = \min_{J \subset [p] \,:\, |J| \leq s} \; \min_{\boldsymbol{u} \in \mathbb{R}^{|J|} \,:\, \|\boldsymbol{u}\|_2 = 1} \frac{1}{\sqrt{n}} \|\boldsymbol{X}_J \boldsymbol{u}\|_2. \tag{11}$$

The quantity ν_s is the smallest sparse singular value of among sub-matrices of $\frac{1}{\sqrt{n}}\boldsymbol{X}$ made of at most s columns. Note that, indeed, $\mathbf{I}(s)$ is equivalent to $\nu_s > 0$.

2.4. Support recovery

We now state the main result concerning the support recovery problem. It states that, under $\mathbf{I}((2+\varepsilon)s_\star)$, the posterior distribution Π concentrates sharply on the support of $\boldsymbol{\beta}_\star$ — which we assumed to be $s_\star$-sparse — as long as λ and the nonzero coefficients are sufficiently large.

Theorem 1. *Consider a design matrix $\boldsymbol{X}$, with $p \geq n$ and normalized column vectors (2), that satisfies Condition $\mathbf{I}((2+\varepsilon)s_\star)$ for some fixed $\varepsilon > 0$. Assume that (9) holds and*

$$\min_{j \in J_\star} |\beta_{\star,j}| \geq \rho := \frac{3\sigma\sqrt{\lambda/n}}{\nu_{(2+\varepsilon)s_\star}}. \tag{12}$$

Then, with probability at least $1 - 2p^{-c}$, $\Pi(J_\star) > \Pi(J)$ for all J, and in fact

$$\Pi\left(J_\star\right) \geq 1 - 4p^{-c}.$$

Under the conditions of Theorem 1, some straightforward calculations imply that $\widehat{J}_{\mathrm{rew}} = J_\star$ with probability at least $1 - 6p^{-c}$. In particular, as $p \to \infty$, the REW consistently recovers the support of the coefficient vector. Note that the same is immediately true for $\widehat{J}_{\mathrm{map}}$ with probability at least $1 - 2p^{-c}$. Thus, we have the following corollary.

Corollary 1. *Let the conditions of Theorem 1 be satisfied. Let $\hat{J}$ denote either $\widehat{J}_{\mathrm{map}}$ or $\widehat{J}_{\mathrm{rew}}$. Then we have with probability at least $1 - 6p^{-c}$ that $\hat{J} = J_\star$*

The result applies in the high-dimensional setting $p > n$, as long as the conditions are met. Note that some classes of matrices like random Gaussian, Rademacher or Fourier matrices are known to satisfy our conditions with probability close to 1. See for instance [37] for a review of existing results. Characterizing all design matrices $\boldsymbol{X}$ that satisfy $\mathbf{I}((2+\varepsilon)s_\star)$ in the high-dimensional setting is an interesting open question beyond the scope of this paper.

We mention that, if $s_\star$ is known and we restrict the prior over subsets J of size exactly $s_\star$, then the same conclusions are valid with $\varepsilon = 0$ and $\nu_{(2+\varepsilon)s_\star}$ replaced by $\nu_{2s_\star}$ in (12), yielding consistent support recovery under the minimum identifiability condition $\mathbf{I}(2s_\star)$. In Section 3, we show that the Lasso estimator requires much more restrictive conditions on the design matrix and $\boldsymbol{\beta}_\star$ to ensure it selects the correct variables with high probability.

Finally, we note that the concentration is even stronger. Under the same conditions, if

$$\lambda = \frac{(1+\varepsilon)(23+5c)+m}{\varepsilon}\log p,$$

for some fixed constant $m \geq 1$, then

$$\sum_{J\subset[p]\,:\,J\neq J_\star} |J|^m \Pi(J) \leq 4p^{-c}\Pi(J_\star). \tag{13}$$

We will use this refinement in the proof of Theorem 4.

2.5. Estimation

Armed with results for the *support recovery*, we establish corresponding bounds for the estimation problem. Our first result is a simple consequence of Theorem 5 and Proposition 1.

Theorem 2. *Consider a design matrix $\boldsymbol{X}$ with $p \geq n$ and normalized column vectors (2). Let $\widehat{\boldsymbol{\beta}}$ denote either $\widehat{\boldsymbol{\beta}}_{\mathrm{map}}$ or $\widehat{\boldsymbol{\beta}}_{\mathrm{rew}}$. Assume λ satisfies (9) with $\varepsilon \leq 1/2$. Then with probability at least $1-3p^{-c}$, we have*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\|_2 \leq \sigma\sqrt{\frac{8s_\star\lambda}{n\nu^2_{(2+\varepsilon)s_\star}}}.$$

We continue with bounds on the estimation error, this time in terms of the l_∞-norm. Based on Theorem 1 (and its proof), we deduce the following.

Theorem 3. *Let the conditions of Theorem 1 be satisfied. Let $\widehat{\boldsymbol{\beta}}$ denote either $\widehat{\boldsymbol{\beta}}_{\mathrm{map}}$ or $\widehat{\boldsymbol{\beta}}_{\mathrm{rew}}$. Then, with probability at least $1-7p^{-c}$, we have*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\|_\infty \leq \sigma\sqrt{\frac{2(c+1)\log p}{n\nu^2_{s_\star}}}. \tag{14}$$

We emphasize that this estimator requires only the near minimum condition $\mathbf{I}((2+\varepsilon)s_\star)$ and that the nonzero components of $\boldsymbol{\beta}_\star$ are somewhat larger than the noise level in (12) to achieve the optimal (up to logs) dependence on n, p of the l_∞-norm estimation bound. We will develop this point further in Section 3 below where we compare our procedure to the Lasso and MCP estimators.

We now study the performances of the AEW $\widehat{\boldsymbol{\beta}}_{\mathrm{aew}}$ and that of the following variant

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{aew}} = \sum_{J \subset [p] \,:\, \nu_J > 0} \Pi(J) \widehat{\boldsymbol{\beta}}_J, \qquad \nu_J := \min_{\boldsymbol{u} \in \mathbb{R}^{|J|} \,:\, \|\boldsymbol{u}\|_2 = 1} \frac{1}{\sqrt{n}} \|\boldsymbol{X}_J \boldsymbol{u}\|_2. \tag{15}$$

Define the quantity $\nu_{\min} = \min_{J \subset [p] \,:\, \nu_J > 0} \{\nu_J\}$, and note that $\nu_{\min} > 0$.

Theorem 4. *Let the conditions of Theorem 1 be satisfied and let $c \geq 1$.*

1. *Take $\lambda = \frac{(1+\varepsilon)(23+5c)+1}{\varepsilon} \log p$. Then, with probability at least $1 - 4p^{-c}$,*

$$\|\widetilde{\boldsymbol{\beta}}_{\mathrm{aew}} - \boldsymbol{\beta}_\star\|_\infty \leq \sigma \sqrt{\frac{2(c+1)\log p}{n \nu_{s_\star}^2}} + \frac{3}{\nu_{\min} p^c} \left[\sigma \sqrt{(20+4c)\frac{\log p}{n}} + \frac{\|\boldsymbol{X}\boldsymbol{\beta}_\star\|_2}{\sqrt{n}} + \nu_{\min} \|\boldsymbol{\beta}_\star\|_\infty \right].$$

2. *If in addition $\mathbf{I}(\overline{s})$ is satisfied,*

$$\|\widehat{\boldsymbol{\beta}}_{\mathrm{aew}} - \boldsymbol{\beta}_\star\|_\infty \leq \sigma \sqrt{\frac{2(c+1)\log p}{n \nu_{s_\star}^2}} + \frac{3}{\nu_{\overline{s}} p^c} \left[\sigma \sqrt{(20+4c)\frac{\log p}{n}} + \frac{\|\boldsymbol{X}\boldsymbol{\beta}_\star\|_2}{\sqrt{n}} + \nu_{\overline{s}} \|\boldsymbol{\beta}_\star\|_\infty \right].$$

3. *If in addition $\mathbf{I}(s_\star + \overline{s})$ is satisfied and $\lambda \geq (62+4c)\log p$,*

$$\|\widehat{\boldsymbol{\beta}}_{\mathrm{aew}} - \boldsymbol{\beta}_\star\|_\infty \leq \sigma \sqrt{\frac{2(c+1)\log p}{n \nu_{s_\star}^2}} + \frac{2\sqrt{10}\sigma}{\sqrt{n}\nu_{s_\star + \overline{s}}} \left[\frac{2\sqrt{s_\star}}{p^c} + \frac{1}{p^{s_\star}} \right]. \tag{16}$$

We note that $\widehat{\boldsymbol{\beta}}_{\mathrm{aew}}$ requires at least $\mathbf{I}(\overline{s})$. (Recall that we assume $\overline{s}$ is known such that $s_\star \leq \overline{s}$.) In practice, when the sparsity is unknown, we make a conservative choice $\overline{s} \gg 2s_\star$ so that $\mathbf{I}(\overline{s})$ is substantially more restrictive that $\mathbf{I}(2s_\star)$. Typically, we assume that $s_\star = O(\frac{n}{\log p})$ and we take $\overline{s}$ of this order of magnitude. We will see below in Section 2.6 that for Gaussian design, the condition $\mathbf{I}(s_\star + \overline{s})$ is satified with probability close to 1. On the other hand, the estimation result for $\widetilde{\boldsymbol{\beta}}$ holds true under the near minimum condition $\mathbf{I}((2+\varepsilon)s_\star)$. For both estimators, their estimation bounds depend on the quantities $\nu_{\min}$, $\nu_{\overline{s}}$, $\|\boldsymbol{X}\boldsymbol{\beta}_\star\|_2$ and $\|\boldsymbol{\beta}_\star\|_\infty$ which can potentially yield a sub-optimal rate of estimation. Note however the presence of the factor p^{-c} in the bound. In particular, if

the nonzero components of $\boldsymbol{\beta}_\star$ are sufficiently large, then the quantities $\nu_{\min}$, $\nu_{\overline{s}}$, $\|\boldsymbol{X}\boldsymbol{\beta}_\star\|_2$ and $\|\boldsymbol{\beta}_\star\|_\infty$ may be completely cancelled for a sufficiently large $c > 0$. If $\mathbf{I}(s_\star + \overline{s})$ is satisfied, then we can derive a bound that no longer depends on $\|\boldsymbol{X}\boldsymbol{\beta}_\star\|_2$ and $\|\boldsymbol{\beta}_\star\|_\infty$. We will also see below that this bound yields the optimal rate of l_∞-norm estimation (up to logs) for the estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{aew}}$ when the design matrix is Gaussian. Optimality considerations are further discussed in Section 4 based on recent information bounds obtained elsewhere.

2.6. Example: Gaussian design

The quintessential example is that of a random Gaussian design, where the *row* vectors of $\boldsymbol{X}$, denoted $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, are independent Gaussian vectors in $\mathbb{R}^p$ with zero mean and $p \times p$ covariance matrix $\boldsymbol{\Sigma}$. If we assume that $\boldsymbol{\Sigma}$ has 1's on the diagonal, the resulting (random) design is just slightly outside our setting, since the columns vectors are not strictly normalized. Our results apply nevertheless. Therefore, it is of interest to lower-bound ν_s for such a design.

We start by relating $\boldsymbol{X}$ and $\boldsymbol{\Sigma}$. Consider $J \subset [p]$, and let $\boldsymbol{\Sigma}_J$ denote the principal submatrix of $\boldsymbol{\Sigma}$ indexed by J. By [40, Cor. 1.50 and Rem. 1.51], there is a numeric constant $C > 0$ such that, when $n \geq C|J|/\eta^2$, with probability at least $1 - 2\exp(-\eta^2 n/C)$, we have

$$\left\| \frac{1}{n} \boldsymbol{X}_J^\top \boldsymbol{X}_J - \boldsymbol{\Sigma}_J \right\| \leq \eta \|\boldsymbol{\Sigma}_J\|,$$

where $\|\cdot\|$ denotes the matrix spectral norm. When this is the case, by Weyl's theorem [39, Cor. IV.4.9],

$$\lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}_J^\top \boldsymbol{X}_J\right) \geq \lambda_{\min}(\boldsymbol{\Sigma}_J) - \eta \lambda_{\max}(\boldsymbol{\Sigma}_J),$$

where $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote the smallest and largest eigenvalues of a symmetric matrix $\boldsymbol{A}$. Define

$$\eta_{\boldsymbol{\Sigma}}(s) = \max_{J:|J|\leq s} \frac{\lambda_{\max}(\boldsymbol{\Sigma}_J)}{\lambda_{\min}(\boldsymbol{\Sigma}_J)}, \qquad \lambda_{\boldsymbol{\Sigma}}(s) = \min_{J:|J|\leq s} \lambda_{\min}(\boldsymbol{\Sigma}_J).$$

Assume that

$$n \geq \frac{aCs\log p}{\eta_{\boldsymbol{\Sigma}}(s)^2},$$

for some $a \geq 2$. Then, with probability at least $1 - 2p^{-a/2}$,

$$\nu_s \geq \frac{\lambda_{\boldsymbol{\Sigma}}(s)^{1/2}}{2}.$$

For example, in standard compressive sensing where $\boldsymbol{\Sigma}$ is the identity matrix, we have $\eta_{\boldsymbol{\Sigma}}(s) = \lambda_{\boldsymbol{\Sigma}}(s) = 1$ for all s, in which case with high probability $\nu_s \geq 1/2$ when $n \geq 2Cs\log p$. Consequently, the l_∞-norm estimation bounds in (14) and (16) are of the order $b\sigma\sqrt{\log(p)/n}$ for some numerical constant $b > 0$. Again, the constants are loose in this discussion.

3. A comparison with the literature

3.1. Denoising and prediction for exponential weights

The existing literature on exponential weights focuses entirely on the *denoising/prediction* problem. In particular, sharp oracle inequalities are available. See the recent survey [36]. Our general approach, and what we established in the previous section, allows us to also quantify the performance of exponential weights at *denoising*. Indeed, as a direct consequence of Proposition 1, we establish new sparsity oracle inequalities in probability for the *denoising* problem. In particular, we show without any assumption on $\boldsymbol{X}$ and $\boldsymbol{\beta}_\star$ that the MAP, REW and AEW come within a log factor of that of the oracle estimator $\widehat{\boldsymbol{\beta}}_{J_\star}$ in terms of *denoising* performance:

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{J_\star} - \boldsymbol{X}\boldsymbol{\beta}_\star\|_2 = \|\boldsymbol{P}_{J_\star}\boldsymbol{z}\|_2 = O_P(\sigma\sqrt{s_\star}).$$

Theorem 5. *Consider a design matrix $\boldsymbol{X}$ with $p \geq n$ and normalized column vectors (2). Assume $\lambda = (62 + 12c)\log p$ for some $c > 0$. Then with probability at least $1 - p^{-c}$,*

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{map}} - \boldsymbol{X}\boldsymbol{\beta}_\star\|_2 \leq \sigma\sqrt{8s_\star\lambda}, \tag{17}$$

and

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{aew}} - \boldsymbol{X}\boldsymbol{\beta}_\star\|_2 \leq \sigma\sqrt{12s_\star\lambda}. \tag{18}$$

Note that here, and anywhere else in the paper, what is true of $\widehat{\boldsymbol{\beta}}_{\mathrm{map}}$ is true of $\widehat{\boldsymbol{\beta}}_J$ for any J such that $\Pi(J) \geq \Pi(J_\star)$.

In [2, 35], a sharp sparsity oracle inequality for the *prediction* problem is established in expectation for the AEW using the approach by Stein's Lemma from [27]. Note that [2] also established an oracle inequality in probability for a different version of the REW that requires the knowledge of $\|\boldsymbol{\beta}_\star\|_1$. Here, we use instead the concentration property of the posterior Π and derive a sparsity oracle inequality for *denoising* in probability that is minimax optimal up to a logarithmic factor without knowing $\|\boldsymbol{\beta}_\star\|_1$. An open question is to determine whether our approach can be used to establish a similar result for the *prediction* problem without the knowledge of $\|\boldsymbol{\beta}_\star\|_1$.

Finally, note that there is no contradiction between Theorems 1 and 5 established in the high-dimensional setting $p \geq n$ and Theorem 1 in [44] which states the impossibility to achieve simultaneously consistent variable selection and denoising. More precisely, [44] proved that for any procedure $\widehat{\boldsymbol{\beta}}$ such that $\mathbb{P}(J(\widehat{\boldsymbol{\beta}}) = J_\star) \to 1$ as $n \to \infty$, then we also have $\lim_{n\to\infty} \mathbb{E}[\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}_\star\|_2] = \infty$. Indeed, the bound in (17)–(18) satisfies $\sigma\sqrt{8s_\star\lambda} \to \infty$ as $n \to \infty$ in the high-dimensional setting $p \geq n$.

3.2. Bayesian model selection and BIC estimator

Many Bayesian techniques for model selection have proposed in the literature; see [15] for a comprehensive review. That same paper suggests a procedure

similar to ours, except that it is a bonafide Bayesian model and they use the following independence sparsity prior

$$\tilde{\pi}(J) = \omega^{|J|}(1-\omega)^{p-|J|},$$

where $\omega \in (0,1)$ controls the sparsity level. Roughly, λ for our prior corresponds to $\log(1-1/\omega)$ for this prior. Our main results remain valid under this prior.

[14] studied the performance of BIC in high-dimensional settings. Not only showed that BIC was consistent when $p < \sqrt{n}$ (under some mild conditions on the design matrix), they also suggested a modification of the penalty term to yield a method that is consistent for larger values of p when the number of variables in the true (i.e., sparsest) model $s_\star$ is bounded independently of n or p. [48] also proposed a variable selection result for the BIC estimator essentially under the condition $\mathbf{I}((2+\epsilon)s_\star)$. By a simple modification of our arguments, we can also recover these results under the same condition $\mathbf{I}((2+\epsilon)s_\star)$. Indeed, the results we established for $\widehat{J}_{\text{map}}$— in particular, Corollary 1 — apply to

$$\widehat{J}_{\text{bic}} = \underset{J:|J|\leq \overline{s}}{\arg\min}\ \frac{1}{\sigma^2}\boldsymbol{y}^\top(\boldsymbol{I}-\boldsymbol{P}_J)\boldsymbol{y} + \lambda|J|.$$

[1] considered a prior very similar to (3) and studied the MAP in the context of the *denoising* problem. They established an oracle inequality as well as a performance bound similar to Theorem 5, as well as minimax lower bounds for the denoising problem for model (1).

3.3. The Lasso

The Lasso estimator is the solution of the convex minimization problem

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\left\{\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + 2\lambda\|\boldsymbol{\beta}\|_1\right\},$$

where $\lambda = A\sigma\sqrt{\log(p)/n}$, $A>0$ and $\|\cdot\|_1$ is the l_1-norm. The Lasso has received considerable attention in the literature over the last few years [3, 6, 8, 32, 33, 49]. It is not our goal to make here an exhaustive presentation of all existing results. We refer to Chapter 4 in [30] and the references cited therein for a comprehensive overview of the literature.

Concerning the l_∞-norm estimation and support recovery problems, the most popular assumption is the Irrepresentable Condition [3, 30, 33, 42, 49] denoted from now on by $\mathbf{IC}(s_\star)$. See for instance Assumption 4.2 in [30]. The condition $\mathbf{IC}(s_\star)$ is strictly more restrictive than the identifiability $\mathbf{I}(2s_\star)$ and does not hold true in general when the columns of the design matrix $\boldsymbol{X}$ are not weakly correlated.

We say that a l_∞-norm estimation rate is optimal if it is of the form $\alpha\sigma\sqrt{\log(p)/n}$ where $\alpha>0$ is an absolute constant as in the case of a Gaussian sequence model where $n=p$, $\boldsymbol{X}=\boldsymbol{I}_n$ is the $n\times n$ identity matrix and

$\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$. On the one hand, the best available estimation bound for the LASSO with Gaussian design matrices and l_∞-norm error is of the order $\sigma\sqrt{s_\star(\log p)/n}$ and it is not clear whether this bound can be improved. Based on this best available l_∞-norm estimation bound, we can only guarantee that the LASSO will be consistent for the support recovery problem when $\rho \gtrsim \sigma\sqrt{s_\star(\log p)/n}$. See Section 4.3 in [30] for more details. On the other hand, we established that our exponential weights procedure attains the optimal l_∞-norm estimation rate and achieves support recovery provided that $\rho \gtrsim \sigma\sqrt{(\log p)/n}$ and Condition $\mathbf{I}((2+\varepsilon)s_\star)$ holds true, a condition that allows for non-negligible correlations between the columns of $\boldsymbol{X}$.

3.4. *The* MCP

The MCP estimator initially proposed by [47] is the solution of the following nonconvex minimization problem:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{mcp}} = \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} \Upsilon(|\boldsymbol{\beta}_j|, \lambda, \gamma) \right\}, \tag{19}$$

where $\lambda, \gamma > 0$ and the MCP penalty function Υ is nonconvex, equal to 0 outside a compact neighborhood of 0 and admits a nonzero right derivative at 0. See equations (2.1)-(2.3) in [47] for more details. The performance of this estimator is established in Theorem 1 and Corollary 1 of [47] for the following choice of parameters:

$$d^* = \mathrm{argmax}\{d \geq 1 \,:\, \nu_d > 0\}, \qquad \lambda = \sigma\sqrt{\frac{2\log p}{n}}, \qquad \gamma \asymp \frac{1}{\nu_{d^*}}.$$

Define

$$\kappa_s := \max_{J\subset[p]\,:\,|J|\leq s} \min_{\|\boldsymbol{u}\|=1} \frac{1}{\sqrt{n}}\|\boldsymbol{X}_J\,\boldsymbol{u}\|_2. \tag{20}$$

If

$$d^*/(\kappa_{d^*}/\nu_{d^*} + 1/2) \geq s_\star \to \infty, \quad \text{and} \quad \rho \gtrsim \sigma\sqrt{\frac{\log p}{\nu_{d^*}^2 n}},$$

then $\mathbb{P}(J(\widehat{\boldsymbol{\beta}}_{mcp}) = J_\star) \to 1$. [48] consider a slightly different choice of the parameters in Corollary 3 and the discussion below, $\gamma > \frac{1}{\nu(s_\star+m)}$ where m is potentially large number depending on the characteristics of the model and the penalty function Υ (See equations (23) and (24) there).

We establish our support recovery result for the random exponential weights estimator under the near minimum conditions $\nu_{(2+\epsilon)s_\star} > 0$ and

$$\rho \gtrsim \sigma\sqrt{\frac{\log p}{\nu_{(2+\epsilon)s_\star}^2 n}},$$

where $\epsilon > 0$ is some small absolute constant. Our conditions are less restrictive since for small $s_\star$ we will have $(2+\epsilon)s_\star \leq d^*$ and consequently $\nu_{(2+\epsilon)s_\star} \geq \nu_{d^*}$.

In the high-dimensional setting, we often have $(2+\epsilon)s_\star \ll d^*$ and very ill-posed design matrices $\boldsymbol{X}$ whose ν_s decrease extremely fast as a function of s. Consequently, we will have $\nu_{(2+\epsilon)s_\star} \gg \nu_{d^*}$. In that situation, the randomized exponential weights estimator will achieve support recovery under much weaker conditions than those required in [47] to establish the support recovery consistency of the MCP estimator.

We also note that Corollary 1 in [47] is obtained under the asymptotic setting $p \gg n > s_\star \to \infty$ while our results hold for any settings of $p, n, s_\star$ as long as $\nu_{(2+\epsilon)s_\star} > 0$ and the above condition on ρ is met. This includes in particular the setting of [47].

Finally, we remark that the optimal theoretical choice of γ for the MCP estimator in Corollary 1 in [47] depends on d^* through ν_{d^*}. For arbitrary design matrices $\boldsymbol{X}$ where no theoretical bound on d^* and ν_{d^*} are available, we may need to compute these quantities. This is a delicate combinatorial problem to solve in high-dimension since it requires considering a very large number of submatrices of $\boldsymbol{X}$. The same parameter γ in Corollary 3 and the discussion below in [48] depends on $s_\star$ which is typically unknown. Note that our exponential weights estimators do not present the same limitation. Indeed, the tuning of λ in (3) does not require computing any restricted eigenvalues and the parameter $\bar{s}$ can be chosen conservatively (for example, $\bar{s} = [n/2]$ if no other information is available). The randomized exponential weights and MAP estimators will achieve support recovery under the near minimum condition $\mathbf{I}((2+\varepsilon)s_\star)$ even if $\bar{s}$ is chosen conservatively (say $\bar{s} = \frac{n}{2}$).

4. Discussion

We established some performance bounds for exponential weights when applied to solving the problems of *denoising*, *estimation* and *support recovery*, and deduced similar results for a slightly different Bayesian model selection procedure [15] and ℓ_0-penalized (BIC-type) variable selection. How sharp are these bounds? We did not optimize the numerical constants appearing in our results, simply because we believe our bounds are loose and also because there are no known sharp information bounds for theses problems, except in specific cases [25]. That said, there are some results available in the literature [41, 34, 31, 1] and our bounds come close to these. For example, from [34] we learn that, when $\mathbf{I}(2s_\star)$ holds, there is a universal constant $C > 0$ such that, for any estimator $\widehat{\boldsymbol{\beta}}$ that knows $s_\star$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\|_2 \geq C\sigma\sqrt{\frac{s_\star \log(p/s_\star)}{n\kappa_{2s_\star}^2}}$$

with probability at least $1/2$, where we recall that κ_s is defined in (20) and from [41], we learn that, for another universal constant $C' > 0$,

$$\mathbb{E}\,\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\|_2^2 \geq C'\sigma^2\left(\frac{s_\star \log(ep/s_\star)}{\kappa_{2s_\star}^2} \vee \frac{1}{\nu_{2s_\star}^2}\right).$$

Thus we see that our estimation bounds (14) and (16) come quite close to these information bounds. See also the detailed discussion in [1].

Of course, there is a trade-off with computational tractability, as computing the exponential weights estimates (of even approximating them) in polynomial time remains an open problem. That said, numerical experiments in [35] show that these methods are promising.

5. Proofs

For the sake of brevity, we let $\|\cdot\| = \|\cdot\|_2$ throughout this section. The proofs are rather lengthy but the driving idea is to show that $\sum_{J\in\mathcal{J}\,:\,J\neq J_\star} \Pi(J)$ is negligible in front of $\Pi(J_\star)$ under suitable conditions. To this end, we study the quantities $\frac{\Pi(J)}{\Pi(J_\star)}$ for all $J \neq J_\star$. An important part of the proofs consists in deriving a sharp control on the magnitude of the noisy perturbations. For the sake of clarity, this part is done separately in Lemmas 2–4 whose proofs are given in the appendix. Armed with these intermediate results, we can tune the parameter λ accurately in order to obtain the desired properties for the various exponential weights estimators considered in this paper.

5.1. Proof of Theorem 5

Define $\boldsymbol{\xi}_J = \boldsymbol{P}_J(\boldsymbol{y}) - \boldsymbol{X}\boldsymbol{\beta}_\star$. For $J \subset [p]$ with $|J| = s$, we have

$$\frac{\Pi(J)}{\Pi(J_\star)} = \frac{\binom{p}{s_\star}}{\binom{p}{s}} \exp\left(\lambda(s_\star - s) + \frac{1}{2\sigma^2}(\|\boldsymbol{P}_{J_\star}^\perp(\boldsymbol{z})\|^2 - \|\boldsymbol{P}_J^\perp(\boldsymbol{y})\|^2)\right) \tag{21}$$

with

$$\|\boldsymbol{P}_{J_\star}^\perp(\boldsymbol{z})\|^2 - \|\boldsymbol{P}_J^\perp(\boldsymbol{y})\|^2 = 2\boldsymbol{z}^T(\boldsymbol{\xi}_J - \boldsymbol{\xi}_{J_\star}) + \|\boldsymbol{\xi}_{J_\star}\|^2 - \|\boldsymbol{\xi}_J\|^2. \tag{22}$$

For the inner product on the RHS, note that $\boldsymbol{\xi}_J \in \mathrm{span}(\boldsymbol{X}_{J\cup J_\star})$ and $\boldsymbol{\xi}_{J_\star} \in \mathrm{span}(\boldsymbol{X}_{J_\star})$, so that

$$\left|2\boldsymbol{z}^T(\boldsymbol{\xi}_J - \boldsymbol{\xi}_{J_\star})\right| = \left|2(\boldsymbol{P}_{J\cup J_\star}\boldsymbol{z})^T(\boldsymbol{\xi}_J - \boldsymbol{\xi}_{J_\star})\right| \le 2\|\boldsymbol{P}_{J\cup J_\star}\boldsymbol{z}\|\ \|\boldsymbol{\xi}_J - \boldsymbol{\xi}_{J_\star}\|, \tag{23}$$

by Cauchy-Schwarz's inequality.

Lemma 2. *For any $c > 0$, with probability at least $1 - p^{-c}$,*

$$\|\boldsymbol{P}_J\,\boldsymbol{z}\|^2 \le (20 + 4c)\sigma^2|J|\log p, \quad \forall J \subset [p]. \tag{24}$$

Set $\zeta_J = \sqrt{(20 + 4c)(|J| + s_\star)\log p}$. Using Lemma 2 in (23), from (22) we have

$$\begin{aligned}
\|\boldsymbol{P}_{J_\star}^\perp(\boldsymbol{z})\|^2 - \|\boldsymbol{P}_J^\perp(\boldsymbol{y})\|^2 &\le \sigma\zeta_J\,\|\boldsymbol{\xi}_J - \boldsymbol{\xi}_{J_\star}\| + \|\boldsymbol{\xi}_{J_\star}\|^2 - \|\boldsymbol{\xi}_J\|^2 \\
&\le \sigma\zeta_J\left(\|\boldsymbol{\xi}_J\| + \|\boldsymbol{\xi}_{J_\star}\|\right) + \|\boldsymbol{\xi}_{J_\star}\|^2 - \|\boldsymbol{\xi}_J\|^2 \\
&\le 4\sigma^2\zeta_J^2 + \frac{3}{2}\|\boldsymbol{\xi}_{J_\star}\|^2 - \frac{1}{2}\|\boldsymbol{\xi}_J\|^2 \\
&\le 6\sigma^2\zeta_J^2 - \frac{1}{2}\|\boldsymbol{\xi}_J\|^2,
\end{aligned} \tag{25}$$

where we used the identity $ab \le 2a^2 + b^2/2$ in the third inequality, and Lemma 2 to bound $\|\boldsymbol{\xi}_{J_\star}\|^2$ in the last inequality.

We tackle the bound in (17). By definition, $\Pi(\widehat{J}_{\text{map}}) \ge \Pi(J_\star)$. Take any J such that $\Pi(J) \ge \Pi(J_\star)$ and let $s = |J|$. Plugging in the bound (25) into (21), and using some crude bounds, we have

$$\begin{aligned} 1 \le \frac{\Pi(J)}{\Pi(J_\star)} &\le \exp\left(s_\star \log p + \lambda(s_\star - s) + 3(s+s_\star)(20+4c)\log p - \frac{1}{4\sigma^2}\|\boldsymbol{\xi}_J\|^2 \right) \\ &\le \exp\left(s_\star\big(\lambda + (61+12c)\log p\big) - \frac{1}{4\sigma^2}\|\boldsymbol{\xi}_J\|^2 \right), \end{aligned}$$

where we used the facts that $\binom{p}{s_\star} \le p^{s_\star}$ and $\lambda \ge (60+12c)\log p$. This in turn implies

$$\|\boldsymbol{\xi}_J\|^2 \le 4\sigma^2 \cdot \big(\lambda s_\star + (61+12c)\log p\big) \le 8\sigma^2\lambda,$$

and (17) follows from that.

For the bound in (18), define $\mathcal{J} = \{J : \|\boldsymbol{\xi}_J\| > \sigma\sqrt{10 s_\star \lambda}\}$. We have

$$\begin{aligned} \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{aew}} - \boldsymbol{X}\boldsymbol{\beta}_\star\| &\le \sum_J \|\boldsymbol{\xi}_J\|\Pi(J) \\ &\le \sigma\sqrt{10\lambda s_\star}\sum_{J\notin\mathcal{J}}\Pi(J) + \sum_{J\in\mathcal{J}}\|\boldsymbol{\xi}_J\|\frac{\Pi(J)}{\Pi(J_\star)}. \end{aligned} \tag{26}$$

By (21) and (25), we have

$$\begin{aligned} \|\boldsymbol{\xi}_J\|\,\frac{\Pi(J)}{\Pi(J_\star)} &\le \|\boldsymbol{\xi}_J\|\,\frac{\binom{p}{s_\star}}{\binom{p}{s}}\exp\left(\lambda(s_\star - s) + 3\zeta_J^2 - \frac{1}{4\sigma^2}\|\boldsymbol{\xi}_J\|^2\right) \\ &\le \frac{\sqrt{10}\sigma}{\binom{p}{s}}\exp\left(\lambda(s_\star - s) + s_\star\log p + 3\zeta_J^2 - \frac{1}{5\sigma^2}\|\boldsymbol{\xi}_J\|^2\right), \end{aligned}$$

where we used the fact that $xe^{-x^2} \le 1/\sqrt{2}$ for all x, and $\binom{p}{s_\star} \le p^{s_\star}$. Hence, since $\lambda \ge (62+4c)\log p$, we have

$$\begin{aligned} \sum_{J\in\mathcal{J}}\|\boldsymbol{\xi}_J\|\frac{\Pi(J)}{\Pi(J_\star)} &\le \sum_{s=0}^{\overline{s}}\sum_{J:|J|=s}\frac{\sqrt{10}\sigma}{\binom{p}{s}}\exp\left(\lambda(s_\star - s) + s_\star\log p + 3\zeta_J^2 - 2\lambda s_\star\right) \\ &\le \sqrt{10}\sigma\sum_{s=0}^{\overline{s}}\exp\left(-(s_\star + s)(\lambda - (61+12c)\log p)\right) \\ &= \sqrt{10}\sigma\cdot 2\exp\left(-s_\star(\lambda - (61+12c)\log p)\right) \\ &\le 2\sqrt{10}\sigma p^{-s_\star}. \end{aligned} \tag{27}$$

The result now follows from

$$\sigma\sqrt{10\lambda s_\star} + 2\sqrt{10}\sigma p^{-s_\star} \le \sqrt{10}\sigma(\sqrt{\lambda s_\star} + 1) \le \sigma\sqrt{12\lambda s_\star},$$

since $p \ge 2$ and $s_\star \ge 1$, as well as $\lambda \ge 25$.

5.2. Proof of Proposition 1

Remember (21). We reformulate (22) in the following way

$$\begin{aligned}\|\boldsymbol{P}_{J_\star}^\perp(\boldsymbol{z})\|^2 - \|\boldsymbol{P}_J^\perp(\boldsymbol{y})\|^2 &= \boldsymbol{y}^\top(\boldsymbol{P}_J - \boldsymbol{P}_{J_\star})\boldsymbol{y} \\ &= -\|\boldsymbol{P}_J^\perp \boldsymbol{X}\boldsymbol{\beta}_\star\|^2 - 2\langle \boldsymbol{P}_J^\perp \boldsymbol{X}\boldsymbol{\beta}_\star, \boldsymbol{z}\rangle + \boldsymbol{z}^\top(\boldsymbol{P}_J - \boldsymbol{P}_{J_\star})\boldsymbol{z}. \end{aligned} \tag{28}$$

The natural idea is then to divide the possible subsets J into the following classes $\mathcal{J}_{s,t} = \{J \subset [p] : |J| = s, |J \cap J_\star| = t, J \neq J_\star\}$ and study the behaviour of the above difference on each of these classes (Note that a similar strategy was carried out in [5] in a model selection framework to derive denoising/prediction oracle inequalities for various models). We first bound the inner product in (28).

Lemma 3. *For any $c > 0$, with probability at least $1 - p^{-c}$,*

$$\frac{\langle \boldsymbol{P}_J^\perp \boldsymbol{X}\boldsymbol{\beta}_\star, \boldsymbol{z}\rangle^2}{\|\boldsymbol{P}_J^\perp \boldsymbol{X}\boldsymbol{\beta}_\star\|^2} \le (10 + 2c)\sigma^2(s \vee s_\star - t)\log p, \tag{29}$$

for all $J \in \mathcal{J}_{s,t}$ with $t \le s \wedge s_\star$.

We now bound the quadratic term in (28).

Lemma 4. *For any $c > 0$, with probability at least $1 - p^{-c}$,*

$$\boldsymbol{z}^\top(\boldsymbol{P}_J - \boldsymbol{P}_{J_\star})\boldsymbol{z} \le (20 + 4c)\sigma^2(s \vee s_\star - t)\log p, \tag{30}$$

for all $J \in \mathcal{J}_{s,t}$ with $t \le s \wedge s_\star$.

For a subset $J \subset [p]$, set

$$\gamma_J = \|\boldsymbol{P}_J^\perp \boldsymbol{X}\boldsymbol{\beta}_\star\|. \tag{31}$$

Assume that both (29) and (30) hold, which is true with probability at least $1 - 2p^{-c}$. Then, we have that, for all $J \in \mathcal{J}_{s,t}$:

$$\begin{aligned}\boldsymbol{y}^\top(\boldsymbol{P}_J - \boldsymbol{P}_{J_\star})\boldsymbol{y} &\le -\gamma_J^2 + 2\gamma_J\sigma\sqrt{(10 + 2c)(s \vee s_\star - t)\log p} \\ &\qquad + (20 + 4c)\sigma^2(s \vee s_\star - t)\log p \\ &\le (40 + 8c)\sigma^2(s \vee s_\star - t)\log p - \frac{1}{2}\gamma_J^2 \qquad (32) \\ &\le (40 + 8c)\sigma^2(s \vee s_\star - t)\log p. \qquad (33)\end{aligned}$$

The first inequality comes from (28), (29) and (30). The identity $2ab \le a^2 + b^2$, with $a = \gamma_J/\sqrt{2}$ and $b = \sigma\sqrt{(20 + 4c)(s \vee s_\star - t)\log p}$, justifies the second inequality.

Combining (21) and (33), we get

$$\begin{aligned}
&\sum_{J\,:\,|J|>[(1+\varepsilon)s_\star]}^{\overline{s}} \frac{\Pi(J)}{\Pi(J_\star)} \\
&= \sum_{s=[(1+\varepsilon)s_\star]}^{\overline{s}} \sum_{t=0}^{s_\star} \sum_{J\in\mathcal{J}_{s,t}} \frac{\binom{p}{s_\star}}{\binom{p}{s}} \exp\left(\lambda(s_\star - s) + \frac{1}{2\sigma^2}\boldsymbol{y}^\top(\boldsymbol{P}_J - \boldsymbol{P}_{J_\star})\boldsymbol{y}\right) \qquad (34)\\
&\le \sum_{s=[(1+\varepsilon)s_\star]}^{\overline{s}} \sum_{t=0}^{s_\star} \frac{\binom{s_\star}{t}\binom{p-s_\star}{s-t}\binom{p}{s_\star}}{\binom{p}{s}} \exp\left(\lambda(s_\star - s) + (20+4c)(s-t)\log p\right),
\end{aligned}$$

where we used the fact that $|\mathcal{J}_{s,t}| = \binom{s_\star}{t}\binom{p-s_\star}{s-t}$ in the last inequality.

For the fraction of binomial coefficients, we have

$$\frac{\binom{s_\star}{t}\binom{p-s_\star}{s-t}\binom{p}{s_\star}}{\binom{p}{s}} = \binom{s}{t}\binom{p-s}{s_\star - t}.$$

We then use the standard bound on the binomial coefficient

$$\begin{aligned}
\log\binom{s}{t} + \log\binom{p-s}{s_\star - t} &\le (s-t)\log\left(es/(s-t)\right) \\
&\quad + (s_\star - t)\log\left(e(p-s)/(s_\star - t)\right) \\
&\le 3(s \vee s^* - t)\log p. \qquad (35)
\end{aligned}$$

Hence, we have so far that

$$\sum_{J\,:\,|J|>[(1+\varepsilon)s_\star]}^{\overline{s}} \frac{\Pi(J)}{\Pi(J_\star)} \le \sum_{s=0}^{\overline{s}}\sum_{t=0}^{s_\star} \exp\left(A_{s,t}\right), \qquad (36)$$

where

$$A_{s,t} := \omega(s-t)\log p + \lambda(s_\star - s), \quad \omega := 23 + 4c.$$

Some simple algebra yields

$$\begin{aligned}
\sum_{s\ge[(1+\varepsilon)s_\star]}^{\overline{s}} \sum_{t=0}^{s_\star} \exp\left(A_{s,t}\right) &\le \sum_{s\ge(1+\varepsilon)s_\star} e^{-(\lambda-\omega\log p)(s-s_\star)} \sum_{t=0}^{s_\star} e^{(s_\star - t)\omega\log p} \\
&\le \frac{e^{-(\lambda-\omega\log p)\varepsilon s_\star}}{1 - e^{-\lambda+\omega\log p}} \cdot \frac{e^{(s_\star+1)\omega\log p}}{e^{\omega\log p} - 1} \qquad (37)\\
&\le \frac{p^{-c}}{(1-p^{-\omega})(1-p^{-c})}, \qquad (38)
\end{aligned}$$

where we used the fact that $p^\omega \ge 2$, because $p \ge 2$, and also $-(\lambda - \omega\log p)\varepsilon s_\star + s_\star\omega\log p \le -c\log p$, because of (9). This shows that

$$\begin{aligned}
\Pi(J : |J| > [(1+\varepsilon)s_\star]) &\le \frac{p^{-c}}{(1-p^{-\omega})(1-p^{-c})}\Pi(J_\star) \\
&\le \frac{p^{-c}}{(1-p^{-c})^2}\Pi(J_\star),
\end{aligned}$$

using the fact that $\omega \geq c$. From this, and the fact that $p^{-c} \leq 1/2$, we conclude the proof.

5.3. Proof of Theorem 1

Let $\nu = \nu_{(2+\varepsilon)s_\star}$ for short. The proof of this result is identical to that of Proposition 1 up to (32). We now need a lower bound on γ_J. For this, we use the following irrepresentability result.

Lemma 5. *Let $\boldsymbol{X} = [\boldsymbol{X}_1 \boldsymbol{X}_2]$, with smallest singular value δ, and let $\boldsymbol{P}_2$ denote the orthogonal projection onto $\boldsymbol{X}_2$. Then for any $\boldsymbol{\beta}_1$,*

$$\|(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1\boldsymbol{\beta}_1\| \geq \delta\|\boldsymbol{\beta}_1\|.$$

Note that for any $J \in \mathcal{J}_{s,t}$ with $s - t \leq (1+\varepsilon)s_\star$, the smallest singular value of $[\boldsymbol{X}_{J_\star} \boldsymbol{X}_{J \setminus J_\star}]$ is bounded from below by $\sqrt{n}\nu$; by Lemma 5, this implies that

$$\gamma_J = \|(\boldsymbol{I} - \boldsymbol{P}_J)(\boldsymbol{X}_{J_\star}\boldsymbol{\beta}_\star)\| = \|(\boldsymbol{I} - \boldsymbol{P}_J)(\boldsymbol{X}_{J_\star \setminus J}\boldsymbol{\beta}^*_{J_\star \setminus J})\| \geq \sqrt{n}\nu\|\boldsymbol{\beta}^*_{J_\star \setminus J}\|.$$

Hence,

$$\gamma_J \geq \rho\nu\sqrt{n(s_\star - t)}, \quad \forall J \in \mathcal{J}_{s,t}, \text{ such that } 0 \leq t \leq s_\star \wedge s \text{ and } s \leq t + (1+\varepsilon)s_\star, \tag{39}$$

where we recall that ρ is defined in (12).

In view of (32) and (39) we have, with probability at least $1 - 2p^{-c}$, for all $J \in \mathcal{J}_{s,t}$

$$\begin{aligned} \boldsymbol{y}^\top(\boldsymbol{P}_J - \boldsymbol{P}_{J_\star})\boldsymbol{y} &\leq (40 + 8c)\sigma^2(s \vee s_\star - t)\log p - \frac{1}{2}\gamma_J^2 \\ &\leq (40 + 8c)\sigma^2(s \vee s_\star - t)\log p \\ &\quad - \frac{1}{2}\rho^2\nu^2 n(s_\star - t)\mathbb{1}_{\{s \leq t + (1+\varepsilon)s_\star\}}. \end{aligned} \tag{40}$$

Next, we have

$$\frac{1}{\Pi(J_\star)} = \sum_{J\,:\,|J| > [(1+\varepsilon)s_\star]} \frac{\Pi(J)}{\Pi(J_\star)} + \sum_{J\,:\,|J| \leq [(1+\varepsilon)s_\star]} \frac{\Pi(J)}{\Pi(J_\star)}. \tag{41}$$

The first sum in the right-hand side was already bounded in Proposition 1. We concentrate on the second sum.

Combining (21) and (40), we get

$$\begin{aligned} &\sum_{J\,:\,|J| \leq [(1+\varepsilon)s_\star]} \frac{\Pi(J)}{\Pi(J_\star)} \\ &= \sum_{s=0}^{[(1+\varepsilon)s_\star]} \sum_{t=0}^{s \wedge s_\star} \sum_{J \in \mathcal{J}_{s,t}} \frac{\binom{p}{s_\star}}{\binom{p}{s}} \exp\left(\lambda(s_\star - s) + \frac{1}{2\sigma^2}\boldsymbol{y}^\top(\boldsymbol{P}_J - \boldsymbol{P}_{J_\star})\boldsymbol{y}\right) \end{aligned}$$

$$\leq \sum_{s=0}^{[(1+\varepsilon)s_\star]} \sum_{t=0}^{s\wedge s_\star} \frac{\binom{s_\star}{t}\binom{p-s_\star}{s-t}\binom{p}{s_\star}}{\binom{p}{s}} \exp\left(\lambda(s_\star - s) + (20+4c)(s\vee s_\star - t)\log p - \eta_{s,t}\right),$$

$$\leq \sum_{s=0}^{[(1+\varepsilon)s_\star]} \sum_{t=0}^{s\wedge s_\star} \binom{s}{t}\binom{p-s}{s_\star - t} \exp\left(\lambda(s_\star - s) + (20+4c)(s\vee s_\star - t)\log p - \eta_{s,t}\right),$$

where $\eta_{s,t} := \frac{1}{4\sigma^2}\rho^2\nu^2 n(s_\star - t)\mathbb{1}_{\{s\leq t+[(1+\varepsilon)s_\star]\}}$.

Next, we use again (35) to get

$$\sum_{J\,:\,|J|\leq[(1+\varepsilon)s_\star]} \frac{\Pi(J)}{\Pi(J_\star)} \leq \sum_{s=0}^{[(1+\varepsilon)s_\star]} \sum_{t=0}^{s\wedge s_\star} \exp\left(A_{s,t}\right), \tag{42}$$

where

$$A_{s,t} := \omega(s\vee s_\star - t)\log p + \lambda(s_\star - s) - \eta_{s,t}, \quad \omega := 23 + 4c.$$

Let $\alpha = \frac{\nu^2 n\rho^2}{4\sigma^2} - \omega\log p$, and note that $\alpha \geq 2\lambda \geq \lambda + c\log p$ by (9) and (12). When $s \leq s_\star$, we have $A_{s,t} = -\alpha(s-t) - (\alpha - \lambda)(s_\star - s)$, so that

$$\begin{aligned}\sum_{s=0}^{s_\star}\sum_{t=0}^{s} \exp\left(A_{s,t}\right) &\leq \sum_{s=1}^{s_\star} e^{-(s_\star - s)c\log p} \sum_{t=0}^{s} e^{-\alpha(s-t)} \\ &\leq \frac{1}{(1-e^{-\alpha})(1-p^{-c})}. \end{aligned}\tag{43}$$

When $s_\star < s \leq (1+\varepsilon)s_\star$, we have $A_{s,t} = -\alpha(s_\star - t) - (\lambda - \omega\log p)(s - s_\star)$, with $\lambda \geq \omega\log p + c\log p$, leading to

$$\begin{aligned}\sum_{s=s_\star+1}^{[(1+\varepsilon)s_\star]}\sum_{t=0}^{s_\star} \exp\left(A_{s,t}\right) &\leq \sum_{s=s_\star+1}^{\infty} e^{-(s-s_\star)c\log p} \sum_{t=0}^{s_\star} e^{-\alpha(s_\star - t)} \\ &\leq \frac{p^{-c}}{(1-e^{-\alpha})(1-p^{-c})}. \end{aligned}\tag{44}$$

Combining (38) with (41)-(44), we conclude that

$$\begin{aligned}\frac{1}{\Pi(J_\star)} &\leq \frac{1}{(1-e^{-\alpha})(1-p^{-c})} + \frac{p^{-c}}{(1-e^{-\alpha})(1-p^{-c})} + \frac{p^{-c}}{(1-p^{-\omega})(1-p^{-c})} \\ &\leq \frac{1+2p^{-c}}{(1-p^{-c})^2},\end{aligned}$$

using the fact that $\alpha \geq \omega \geq c$. From this, we get

$$\Pi(J_\star) \geq (1-p^{-c})^2(1-2p^{-c}) \geq (1-2p^{-c})^2 \geq 1 - 4p^{-c}.$$

This concludes the proof of Theorem 1. We note that the proof of (13) is virtually identical.

5.4. Proof of Theorem 2

When (9) is satisfied with $\varepsilon \le 1/2$, then λ satisfies both the conditions of Proposition 1 and Theorem 5. Hence, with probability at least $1-2p^{-c}-p^{-c} = 1-3p^{-c}$, we have both that $|\widehat{J}_{\mathrm{map}}| \le (1+\varepsilon)s_\star$ and (17). Hence, the support of $\widehat{\boldsymbol{\beta}}_{\mathrm{map}} - \boldsymbol{\beta}_\star$ is of size at most $(1+\varepsilon)s_\star + s_\star = (2+\varepsilon)s_\star$, and we have

$$\|\widehat{\boldsymbol{\beta}}_{\mathrm{map}} - \boldsymbol{\beta}_\star\| \le \frac{1}{\nu_{(2+\varepsilon)s_\star}} \|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{\mathrm{map}} - \boldsymbol{\beta}_\star)\|,$$

with

$$\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{\mathrm{map}} - \boldsymbol{\beta}_\star)\| = \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{map}} - \boldsymbol{X}\boldsymbol{\beta}_\star\| \le \sigma\sqrt{8s_\star\lambda},$$

and the result follows.

5.5. Proof of Theorem 3

We prove the result for $\widehat{\boldsymbol{\beta}}_{\mathrm{map}}$. The proof for $\widehat{\boldsymbol{\beta}}_{\mathrm{rew}}$ is the same up to some trivial modifications. For $r > 0$, we have

$$\begin{aligned}\mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}_{\mathrm{map}} - \boldsymbol{\beta}_\star\|_\infty > r\right) &\le \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}_{J_\star} - \boldsymbol{\beta}_\star\|_\infty > r, \widehat{J}_{\mathrm{map}} = J_\star\right) \\ &\quad + \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}_{\mathrm{map}} - \boldsymbol{\beta}_\star\|_\infty > r, \widehat{J}_{\mathrm{map}} \ne J_\star\right) \\ &\le \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}_{J_\star} - \boldsymbol{\beta}_\star\|_\infty > r\right) + \mathbb{P}\left(\widehat{J}_{\mathrm{map}} \ne J_\star\right).\end{aligned}$$

By Theorem 1, $\widehat{J}_{\mathrm{map}} = J_\star$ with probability at least $1-2p^{-c}$, so that the second term on the RHS is bounded by $2p^{-c}$.

Next, we know that $\widehat{\boldsymbol{\beta}}_{J_\star} \sim N(\boldsymbol{\beta}_\star, \sigma^2 \frac{1}{n}\Psi_\star^{-1})$ with $\Psi_\star := \frac{1}{n}\boldsymbol{X}_{J_\star}^\top \boldsymbol{X}_{J_\star}$, and in particular, $\widehat{\beta}_{J_\star,j} - \beta_{\star,j} \sim \mathcal{N}(0, \sigma^2\tau_j^2/n)$, where τ_j^2 is the jth diagonal entry of $\Psi_\star^{-1}$. This matrix being positive semi-definite, its diagonal terms are all bounded from above by its largest eigenvalue, which is the inverse of the smallest eigenvalue of $\Psi_\star$, which in turn is larger than $\nu_{s_\star}^2$. Hence, $\mathrm{Var}(\widehat{\beta}_{J_\star,j}) \le \sigma^2/(n\nu_{s_\star}^2)$ for all $j \in J_\star$, so that a standard tail bound on the normal distribution and the union bound give

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}_{J_\star} - \boldsymbol{\beta}_\star\|_\infty > r\right) \le s_\star \exp\left(-\frac{n\nu_{s_\star}^2 r^2}{2\sigma^2}\right). \tag{45}$$

Taking $r = \sigma\sqrt{2(c+1)\log(p)/(n\nu_{s_\star}^2)}$ bounds this by p^{-c}, and the desired result follows.

5.6. Proof of Theorem 4

Again, we only prove the result for $\widehat{\boldsymbol{\beta}}_{\mathrm{map}}$. The proof for $\widehat{\boldsymbol{\beta}}_{\mathrm{rew}}$ is also almost identical up to some trivial modifications.

We have

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}_{\mathrm{map}}-\boldsymbol{\beta}_\star\|_\infty &\le \sum_J \|\widehat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}_\star\|_\infty \Pi(J)\\ &\le \|\widehat{\boldsymbol{\beta}}_{J_\star}-\boldsymbol{\beta}_\star\|_\infty \Pi(J_\star)+\sum_{J\neq J_\star}\|\widehat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}_\star\|_\infty\Pi(J)\\ &\le \|\widehat{\boldsymbol{\beta}}_{J_\star}-\boldsymbol{\beta}_\star\|_\infty+\sum_{J\neq J_\star}\|\widehat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}_\star\|_\infty\Pi(J). \qquad (46)\end{aligned}$$

For any $c>0$, we have with probability at least $1-p^{-c}$, for any $J\subset[p]$ with $\nu_J>0$, that

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}_J\|_\infty &\le \sqrt{|J|}\|\widehat{\boldsymbol{\beta}}_J\|\\ &\le \frac{\sqrt{|J|}}{\sqrt{n}\nu_J}\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_J\|\\ &\le \frac{\sqrt{|J|}}{\sqrt{n}\nu_J}\left[\|\boldsymbol{P}_J(\boldsymbol{z})\|+\|\boldsymbol{P}_J^\perp(\boldsymbol{X}\boldsymbol{\beta}_\star)\|\right]\\ &\le \frac{\sqrt{|J|}}{\sqrt{n}\nu_J}\left[\sigma\sqrt{(20+4c)|J|\log p}+\|\boldsymbol{X}\boldsymbol{\beta}_\star\|\right],\end{aligned}$$

where we have used Cauchy-Schwarz's inequality in the first line and (24) in the last line.

We now assume that $\nu_{\overline{s}}>0$, which implies that $\nu_J>0$ for any $J\subset[p]$ with $|J|\le\overline{s}$. Combining the previous display with (45) and (46) and a union bound argument, we get with probability at least $1-2p^{-c}$,

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}_{\mathrm{map}}-\boldsymbol{\beta}_\star\|_\infty &\le \sigma\sqrt{\frac{2(c+1)\log p}{n\nu_{s_\star}}}\\ &\quad+\sum_{J\neq J_\star}\left[\frac{\sigma|J|}{\nu_{\overline{s}}}\sqrt{(20+4c)\log p}+\frac{\sqrt{|J|}}{\sqrt{n}\nu_{\overline{s}}}\|\boldsymbol{X}\boldsymbol{\beta}_\star\|+\|\boldsymbol{\beta}_\star\|_\infty\right]\Pi(J).\end{aligned}$$

Next, we combine the above display with (13) and a union bound argument to get with probability at least $1-4p^{-c}$ that

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}_{\mathrm{map}}-\boldsymbol{\beta}_\star\|_\infty &\le \sigma\sqrt{\frac{2(c+1)\log p}{n\nu_{s_\star}}}\\ &\quad+\frac{4p^{-c}}{\nu_{\overline{s}}}\left[\sigma\sqrt{(20+4c)\frac{\log p}{n}}+\frac{\|\boldsymbol{X}\boldsymbol{\beta}_\star\|}{\sqrt{n}}+\nu_{\overline{s}}\|\boldsymbol{\beta}_\star\|_\infty\right].\end{aligned}$$

Note that the same reasoning applied to $\widetilde{\boldsymbol{\beta}}$ yields the same l_∞-norm estimation bound with $\nu_{\overline{s}}$ replaced by $\nu_{\min}$.

We now assume that $\nu_{s_\star+\overline{s}}>0$. Then, for any $J\subset[p]$ with $|J|\le\overline{s}$, we have

$$\|\widehat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}_\star\|_\infty\le\|\widehat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}_\star\|\le\frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_J-\boldsymbol{X}\boldsymbol{\beta}_\star\|}{\sqrt{n}\nu_{s_\star+\overline{s}}}.$$

Combining this last inequality with (46), we get

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}_{\mathrm{map}} - \boldsymbol{\beta}_\star\|_\infty &\le \|\widehat{\boldsymbol{\beta}}_{J_\star} - \boldsymbol{\beta}_\star\|_\infty + \frac{1}{\sqrt{n}\nu_{s_\star+\overline{s}}} \sum_{J \notin \mathcal{J}, J \neq J_\star} \|\boldsymbol{\xi}_J\| \Pi(J) + \frac{1}{\sqrt{n}\nu_{s_\star+\overline{s}}} \sum_{J \in \mathcal{J}} \|\boldsymbol{\xi}_J\| \Pi(J) \\ &\le \|\widehat{\boldsymbol{\beta}}_{J_\star} - \boldsymbol{\beta}_\star\|_\infty + \frac{\sigma\sqrt{10 s_\star}}{\sqrt{n}\nu_{s_\star+\overline{s}}} \Pi(\mathcal{J}^c \setminus J_\star) + \frac{1}{\sqrt{n}\nu_{s_\star+\overline{s}}} \sum_{J \in \mathcal{J}} \|\boldsymbol{\xi}_J\| \Pi(J),\end{aligned}$$

where we recall that $\boldsymbol{\xi}_J = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_J - \boldsymbol{X}\boldsymbol{\beta}_\star$ and $\mathcal{J} = \big\{J \subset [p] : \|\boldsymbol{\xi}_J\| > \sigma\sqrt{10 s_\star \lambda}\big\}$. In view of Theorem 1, we have with probability at least $1 - 2p^{-c}$ that

$$\Pi(\mathcal{J}^c \setminus J_\star) \le 1 - \Pi(J_\star) \le 4p^{-c};$$

and in view of (27),

$$\sum_{J \in \mathcal{J}} \|\boldsymbol{\xi}_J\| \Pi(J) \le 2\sqrt{10}\sigma p^{-s_\star}.$$

Combining the three last displays with (45), we get the result.

5.7. *Proofs of auxiliary results*

Lemma 2 is a special case of Lemma 4 where $J_\star = \emptyset$, and we prove Lemma 4 below.

5.7.1. *Proof of Lemma 3*

First, note that $u_J := \langle \boldsymbol{P}_J^\perp \boldsymbol{X}\boldsymbol{\beta}_\star, \boldsymbol{z}\rangle \sim \mathcal{N}(0, \sigma^2\gamma_J^2)$, where γ_J is defined in (31), so that $v_J := u_J/(\sigma\gamma_J) \sim \mathcal{N}(0,1)$. By the union bound and a standard tail bound on the normal distribution, for $a > 0$, we have

$$\mathbb{P}\left(\max_{J \in \mathcal{J}_{s,t}} v_J^2 > a^2\right) \le \binom{s_\star}{t}\binom{p - s_\star}{s - t} \exp(-a^2/2).$$

As in (35), we have

$$\begin{aligned}\log\binom{s_\star}{t} + \log\binom{p - s_\star}{s - t} &\le (s_\star - t)\log(e s_\star) + (s - t)\log(ep) \\ &\le 3(s \vee s_\star - t)\log p. \qquad (47)\end{aligned}$$

Hence,

$$\mathbb{P}\left(\max_{J \in \mathcal{J}_{s,t}} v_J^2 > (10{+}2c)(s \vee s_\star - t)\log p\right) \le \exp\big(-(2{+}c)(s \vee s_\star - t)\log p\big) \le p^{-(2+c)},$$

since $s \vee s_\star - t = 0$ would imply $J = J_\star$. We then apply the union bound again,

$$\mathbb{P}\left(\max_{s,t} \max_{J \in \mathcal{J}_{s,t}} \frac{v_J^2}{s \vee s_\star - t} > (10 + 2c)\sigma^2 \log p\right) \le \overline{s}\,(s \wedge s_\star + 1)p^{-(2+c)} \le p^{-c},$$

which the result we wanted.

5.7.2. Proof of Lemma 4

Fix $J \in \mathcal{J}_{s,t}$. First, we notice that

$$\boldsymbol{z}^\top(\boldsymbol{P}_J - \boldsymbol{P}_{J_\star})\boldsymbol{z} = \boldsymbol{z}^\top(\boldsymbol{P}_J - \boldsymbol{P}_{J\cap J_\star})\boldsymbol{z} - \boldsymbol{z}^\top(\boldsymbol{P}_{J_\star} - \boldsymbol{P}_{J\cap J_\star})\boldsymbol{z} \le \boldsymbol{z}^\top(\boldsymbol{P}_J - \boldsymbol{P}_{J\cap J_\star})\boldsymbol{z},$$

since $\boldsymbol{P}_{J_\star} - \boldsymbol{P}_{J\cap J_\star}$ is an orthogonal projection, and therefore positive semidefinite. And $\boldsymbol{Q}_J := \boldsymbol{P}_J - \boldsymbol{P}_{J\cap J_\star}$ is also an orthogonal projection, of rank $s-t$, so that $\|\boldsymbol{Q}_J\boldsymbol{z}\|^2 \sim \sigma^2\chi^2_{s-t}$. Chernoff's Bound applied to the chi-square distribution yields

$$\log \mathbb{P}\left(\chi^2_m > a\right) \le -\frac{m}{2}(a/m - 1 - \log(a/m)) \le -\frac{a}{4}, \quad \forall a \ge 2m.$$

The union bound and (47), and this tail bound, yields

$$\mathbb{P}\left(\max_{J\in\mathcal{J}_{s,t}} \|\boldsymbol{Q}_J\boldsymbol{z}\|^2 > (20{+}4c)\sigma^2(s\vee s_\star - t)\log p\right) \le \exp\left(-(2+c)(s\vee s_\star - t)\log p\right).$$

The rest of the proof is exactly the same as that of Lemma 3.

5.8. An irrepresentability result

We have

$$\begin{aligned}
\|(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1\boldsymbol{\beta}_1\|^2 &= \min_{\boldsymbol{\beta}_2} \|\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2\|^2 \\
&= \min_{\boldsymbol{\beta}_2} \boldsymbol{\beta}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} \\
&\ge \min_{\boldsymbol{\beta}_2} \delta^2\|\boldsymbol{\beta}\|^2 \\
&= \delta^2\|\boldsymbol{\beta}_1\|^2,
\end{aligned}$$

where $\boldsymbol{\beta} := (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, implying $\|\boldsymbol{\beta}\|^2 = \|\boldsymbol{\beta}_1\|^2 + \|\boldsymbol{\beta}_2\|^2$.

References

[1] Abramovich, F. and V. Grinshtein (2010). Map model selection in gaussian regression. *Electron. J. Stat. 4*, 932–949. MR2721039

[2] Alquier, P. and K. Lounici (2011). Pac-bayesian theorems for sparse regression estimation with exponential weights. *Electronic Journal of Statistics 5*, 127–145. Arxiv:1009.2707. MR2786484

[3] Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, New York, NY, USA, pp. 33–40. ACM.

[4] Bickel, P., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics 37*(4), 1705–1732. MR2533469

[5] Birgé, L. and P. Massart (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS) 3*(3), 203–268. MR1848946

[6] Bunea, F. (2008). Consistent selection via the Lasso for high dimensional approximating regression models. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, Volume 3 of *Inst. Math. Stat. Collect.*, pp. 122–137. Beachwood, OH: Inst. Math. Statist. MR2459221

[7] Bunea, F. and A. Nobel (2008). Sequential procedures for aggregating arbitrary estimators of a conditional mean. *IEEE Trans. Inform. Theory 54*(4), 1725–1735. MR2450298

[8] Bunea, F., A. Tsybakov, and M. Wegkamp (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics 1*, 169–194. MR2312149

[9] Cai, T. T. and L. Wang (2011). Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Trans. Inform. Theory 57*(7), 4680–4688. MR2840484

[10] Candès, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist. 35*(6), 2313–2351. MR2382644

[11] Candès, E. J. and M. A. Davenport (2013). How well can we estimate a sparse vector? *Appl. Comput. Harmon. Anal. 34*(2), 317–323. MR3008569

[12] Candès, E. J. and Y. Plan (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist. 37*(5A), 2145–2177. MR2543688

[13] Catoni, O. (2004). *Statistical learning theory and stochastic optimization*, Volume 1851 of *Lecture Notes in Mathematics*. Berlin: Springer-Verlag. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. MR2163920

[14] Chen, J. and Z. Chen (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771. MR2443189

[15] Chipman, H., E. I. George, and R. E. McCulloch (2001). The practical implementation of Bayesian model selection. In *Model selection*, Volume 38 of *IMS Lecture Notes Monogr. Ser.*, pp. 65–134. Beachwood, OH: Inst. Math. Statist. With discussion by M. Clyde, Dean P. Foster, and Robert A. Stine, and a rejoinder by the authors. MR2000752

[16] Dalalyan, A. and J. Salmon (2011). Optimal aggregation of affine estimators. In *Proceedings of the 24th annual conference on Computational Learning Theory*, Budapest (Hungary).

[17] Dalalyan, A. and J. Salmon (2012). Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist. 40*(4), 2327–2355. MR3059085

[18] Dalalyan, A. and A. Tsybakov (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, Volume 4539 of *Lecture Notes in Comput. Sci.*, pp. 97–111. Berlin: Springer. MR2397581

[19] Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911. MR2530322

[20] Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory 57*(8), 5467–5484. MR2849368

[21] Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist. 32*(3), 928–961. MR2065194

[22] Gautier, E. and A. Tsybakov (2011, October). High-dimensional instrumental variables regression and confidence sets. Technical report, Arxiv preprint 1105.2454v3. MR2867623

[23] Giraud, C. (2008). Mixing least-squares estimators when the variance is unknown. *Bernoulli 14*(4), 1089–1107. MR2543587

[24] Ji, P. and J. Jin (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist. 40*(1), 73–103. MR3013180

[25] Jin, J., C. Zhang, and Q. Zhang (2012). Optimality of graphlet screening in high dimensional variable selection. Available online at http://arxiv.org/abs/1204.6452.

[26] Juditsky, A., P. Rigollet, and A. B. Tsybakov (2008). Learning by mirror averaging. *Ann. Statist. 36*(5), 2183–2206. MR2458184

[27] Leung, G. and A. Barron (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory 52*(8), 3396–3410. MR2242356

[28] Lounici, K. (2007). Generalized mirror averaging and D-convex aggregation. *Math. Methods Statist. 16*(3), 246–259. MR2356820

[29] Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics 2*, 90–102. MR2386087

[30] Lounici, K. (2009). *Statistical Estimation in High-Dimension, Sparsity Oracle Inequalities*. Ph. D. thesis, University Paris Diderot - Paris 7.

[31] Lounici, K., M. Pontil, A. Tsybakov, and S. van de Geer (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist. 39*(4), 2164–2204. MR2893865

[32] Meinshausen, N., P. Bühlmann, and E. Zürich (2006). High dimensional graphs and variable selection with the lasso. *Annals of Statistics 34*, 1436–1462. MR2278363

[33] Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist. 37*(1), 246–270. MR2488351

[34] Raskutti, G., M. J. Wainwright, and B. Yu (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory 57*(10), 6976–6994. MR2882274

[35] Rigollet, P. and A. Tsybakov (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist. 39*(2), 731–771. MR2816337

[36] Rigollet, P. and A. B. Tsybakov (2012). Sparse estimation by exponential weighting. *Statist. Sci. 27*(4), 558–575. MR3025134

[37] Rudelson, M. and S. Zhou (2013). Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory 59*(6), 3434–3447. MR3061256

[38] Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica 7*(2), 221–264. With comments and a rejoinder by the author. MR1466682

[39] Stewart, G. W. and J. G. Sun (1990). *Matrix perturbation theory*. Computer Science and Scientific Computing. Boston, MA: Academic Press Inc. MR1061154

[40] Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. Available from http://arxiv.org/abs/1011.3027. MR2963170

[41] Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron. J. Stat. 6*, 38–90. MR2879672

[42] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory 55*(5), 2183–2202. MR2729873

[43] Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli 10*(8), 25–47. MR2044592

[44] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika 92*(4), 937–950. MR2234196

[45] Ye, F. and C.-H. Zhang (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res. 11*, 3519–3540. MR2756192

[46] Zhang, C.-H. (2007). Information-theoretic optimality of variable selection with concave penalty. Technical report, Dept. Statistics, Rutgers Univ.

[47] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist. 38*(2), 894–942. MR2604701

[48] Zhang, C.-H. and T. Zhang (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci. 27*(4), 576–593. MR3025135

[49] Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res. 7*, 2541–2563. MR2274449