SLOPE is Adaptive to Unknown Sparsity and Asymptotically Minimax

Weijie Su* Emmanuel Candès*,†

* Department of Statistics, Stanford University, Stanford, CA 94305, USA
† Department of Mathematics, Stanford University, Stanford, CA 94305, USA

March 2015

Abstract

We consider high-dimensional sparse regression problems in which we observe $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}$, where $\boldsymbol{X}$ is an $n \times p$ design matrix and $\boldsymbol{z}$ is an n-dimensional vector of independent Gaussian errors, each with variance σ^2. Our focus is on the recently introduced SLOPE estimator [15], which regularizes the least-squares estimates with the rank-dependent penalty $\sum_{1\le i\le p} \lambda_i |\widehat{\beta}|_{(i)}$, where $|\widehat{\beta}|_{(i)}$ is the ith largest magnitude of the fitted coefficients. Under Gaussian designs, where the entries of $\boldsymbol{X}$ are i.i.d. $\mathcal{N}(0, 1/n)$, we show that SLOPE, with weights λ_i just about equal to $\sigma \cdot \Phi^{-1}(1 - iq/(2p))$ ($\Phi^{-1}(\alpha)$ is the αth quantile of a standard normal and q is a fixed number in $(0, 1)$) achieves a squared error of estimation obeying

$$\sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}} - \boldsymbol{\beta}\|^2 > (1+\epsilon)\, 2\sigma^2 k \log(p/k)\right) \longrightarrow 0$$

as the dimension p increases to ∞, and where $\epsilon > 0$ is an arbitrary small constant. This holds under weak assumptions on the sparsity level k and is sharp in the sense that this is the best possible error *any* estimator can achieve. A remarkable feature is that SLOPE does not require any knowledge of the degree of sparsity, and yet automatically adapts to yield optimal total squared errors over a wide range of sparsity classes. We are not aware of any other estimator with this property.

Keywords. SLOPE, Lasso, sparse regression, adaptivity, false discovery rate (FDR), Benjamini-Hochberg procedure, FDR thresholding.

1 Introduction

Twenty years ago, Benjamini and Hochberg proposed the false discovery rate (FDR) as a new measure of type-I error for multiple testing, along with a procedure for controlling the FDR in the case of statistically independent tests [8]. In words, the FDR is the expected value of the ratio between the number of false rejections and the total number of rejections, with the convention that this ratio vanishes in case no rejection is made. To describe the Benjamini-Hochberg procedure, henceforth referred to as the BHq procedure, imagine we observe a p-dimensional vector $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_p)$ of independent statistics $\{y_i\}$, and wish to test which means β_i are nonzero. Begin

by ordering the observations as $|y|_{(1)} \geq |y|_{(2)} \geq \cdots \geq |y|_{(p)}$—that is, from the most to the least significant—and compute a data-dependent threshold given by

$$\widehat{t}_{\text{FDR}} = |y|_{(R)},$$

where R is the last time $|y|_{(i)}/\sigma$ exceeds a critical curve λ_i^{BH}: formally,

$$R \triangleq \max\left\{i : |y|_{(i)}/\sigma \geq \lambda_i^{\text{BH}}\right\} \text{ with } \lambda_i^{\text{BH}} = \Phi^{-1}\left(1 - iq/(2p)\right); \tag{1.1}$$

throughout, $0 < q < 1$ is a target FDR level and Φ is the cumulative distribution function of a standard normal random variable. (The chance that a null statistic $z \sim \mathcal{N}(0,1)$ exceeds λ_i^{BH} is $\mathbb{P}(|z| \geq \lambda_i^{\text{BH}}) = q \cdot i/p$.) Then BHq rejects all those hypotheses with $|y_i| \geq \widehat{t}_{\text{FDR}}$ and makes no rejection in the case where all the observations fall below the critical curve, i.e. when the set $\{i : |y|_{(i)}/\sigma \geq \lambda_i^{\text{BH}}\}$ is empty. In short, the hypotheses corresponding to the R most significant statistics are rejected. Letting V be the number of false rejections, Benjamini and Hochberg proved that this procedure controls the FDR in the sense that

$$\text{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \frac{qp_0}{p} \leq q,$$

where $p_0 = |\{i : \beta_i = 0\}|$ is the total number of nulls. Unlike a Bonferroni procedure—see e.g. [14]—where the threshold for significance is fixed in advance, a very appealing feature of the BHq procedure is that the threshold is adaptive as it depends upon the data $\boldsymbol{y}$. Roughly speaking, this threshold is high when there are few discoveries to be made and low when there are many.

Interestingly, the acceptance of the FDR as a valid error measure has been slow coming, and we have learned that the FDR criterion initially met much resistance. Among other things, researchers questioned whether the FDR is the right quantity to control as opposed to more traditional measures such as the familywise error rate (FWER), and even if it were, they asked whether among all FDR controlling procedures, the BHq procedure is powerful enough. Today, we do not need to argue that this step-up procedure is a useful tool for addressing multiple comparison problems, as both the FDR concept and this method have gained enormous popularity in certain fields of science; for instance, they have influenced the practice of genomic research in a very concrete fashion. The point we wish to make is, however, different: as we discuss next, if we look at the multiple testing problem from a different point of view, namely, from that of estimation, then FDR becomes in some sense the right notion to control, and naturally appears as a valid error measure.

Consider estimating $\boldsymbol{\beta}$ from the same data $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_p)$ and suppose we have reasons to believe that the vector of means is sparse in the sense that most of the coordinates of $\boldsymbol{\beta}$ may be zero or close to zero, but have otherwise no idea about the number of 'significant' means. It is well known that under sparsity constraints, thresholding rules can far outperform the maximum likelihood estimate (MLE). A key issue is thus how one should determine an appropriate threshold. Inspired by the adaptivity of BHq, Abramovich and Benjamini [1] suggested estimating the mean sequence by the following *testimation* procedure:[1] use BHq to select which coordinates are worth estimating via the MLE and which do not and can be set to zero. Formally, set $0 < q < 1$ and define the FDR estimate as

$$\widehat{\beta}_i = \begin{cases} y_i, & |y_i| \geq \widehat{t}_{\text{FDR}}, \\ 0, & \text{otherwise.} \end{cases} \tag{1.2}$$

[1] See [4] for the use of this word.

The idea behind the FDR-thresholding procedure is to automatically adapt to the unknown sparsity level of the sequence of means under study. Now a remarkably insightful article [2] published ten years ago rigorously established that this way of thinking is fundamentally correct in the following sense: if one chooses a constant $q \in (0, 1/2]$, then the FDR estimate is asymptotically minimax over the class of k-sparse signals as long as k is neither too small nor too large. More precisely, take any $\boldsymbol{\beta} \in \mathbb{R}^p$ with a number k of nonzero coordinates obeying $\log^5 p \le k \le p^{1-\delta}$ for any constant $\delta > 0$. Then as $p \to \infty$, it holds that

$$\mathrm{MSE} = \mathbb{E}\,\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \le (1 + o(1))\, 2\sigma^2 k \log(p/k). \tag{1.3}$$

It can be shown that the right-hand side is the asymptotic minimax risk over the class of k-sparse signals ([2] provides other asymptotic minimax results for ℓ_p balls) and, therefore, there is a sense in which the FDR estimate asymptotically achieves the best possible mean-square error (MSE). This is remarkable because the FDR estimate is not given any information about the sparsity level k and no matter this value in the stated range, the estimate will be of high quality. To a certain extent, the FDR criterion strikes the perfect balance between bias and variance. Pick a higher threshold/or a more conservative testing procedure and the bias will increase resulting in a loss of minimaxity. Pick a lower threshold/or use a more liberal procedure and the variance will increase causing a similar outcome. Thus we see that the FDR criterion provides a fundamentally correct answer to an estimation problem with squared loss, which is admittedly far from being a pure multiple testing problem.

For the sake of completeness, we emphasize that the FDR thresholding estimate happens to be very close to penalized estimation procedures proposed earlier in the literature, which seek to regularize the maximum likelihood by adding a penalty term of the form

$$\operatorname*{argmin}_{\boldsymbol{b}} \ \|\boldsymbol{y} - \boldsymbol{b}\|_2^2 + \sigma^2\, \mathrm{Pen}(\|\boldsymbol{b}\|_0), \tag{1.4}$$

where $\mathrm{Pen}(k) = 2k \log(p/k)$ see [36] and [13, 51] for related ideas. In fact, [2] begins by considering the penalized MLE with

$$\mathrm{Pen}(k) = \sum_{i \le k} (\lambda_i^{\mathrm{BH}})^2 = (1 + o(1))\, 2k \log(p/k),$$

which is different from the FDR thresholding estimate, and shown to enjoy asymptotic minimaxity under the restrictions on the sparsity levels listed above. In a second step, [2] argues that the FDR thresholding estimate is sufficiently close to this penalized MLE so that the estimation properties carry over.

1.1 SLOPE

Our aim in this paper is to extend the link between estimation and testing by showing that a procedure originally aimed at controlling the FDR in variable selection problems enjoys optimal estimation properties. We work with a linear model, which is far more general than the orthogonal sequence model discussed up until this point; here, we observe an n-dimensional response vector obeying

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}, \tag{1.5}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of regression coefficients and $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ is an error term.

On the testing side, finding finite sample procedures that would test the p hypotheses $H_j : \beta_j = 0$ while controlling the FDR—or other measures of type-I errors—remains a challenging topic. When $p \le n$ and the design $\boldsymbol{X}$ has full column rank, this is equivalent to testing a vector of means under arbitrary correlations since the model is equivalent to $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$ ($\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ is the least-squares estimate). Applying BHq procedure to the least-squares estimate (1) is not known to control the FDR (the positive regression dependency [9] does not hold here), and (2) suffers from high variability in false discovery proportions due to correlations [15]. Having said this, we are aware of recent significant progress on this problem including the development of the knockoff filter [6], which is a powerful FDR controlling method working when $p \le n$, and other innovative ideas [41, 42] relying on assumptions, which may not always hold.

On the estimation side, there are many procedures available for fitting sparse regression models and the most widely used is the Lasso [50]. When the design is orthogonal, the Lasso simply applies the same soft-thresholding rule to all the coordinates of the least-squares estimates. This is equivalent to comparing all the p-values to a *fixed* threshold. In the spirit of the adaptive BHq procedure, [15] proposed a new fitting strategy called SLOPE, a short-hand for Sorted L-One Penalized Estimation: fix a nonincreasing sequence $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$ not all vanishing; then SLOPE is the solution to

$$\underset{\boldsymbol{b}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda_1|b|_{(1)} + \lambda_2|b|_{(2)} + \cdots + \lambda_p|b|_{(p)}, \tag{1.6}$$

where $|b|_{(1)} \ge |b|_{(2)} \ge \cdots \ge |b|_{(p)}$ are the order statistics of $|b_1|, |b_2|, \ldots, |b_p|$. The regularization is a *sorted* ℓ_1 *norm*, which penalizes coefficients whose estimate is larger more heavily than those whose estimate is smaller. This reminds us of the fact that in multiple testing procedures, larger values of the test statistics are compared with higher thresholds. In particular, recall that BHq compares $|y|_{(i)}/\sigma$ with $\lambda_i^{\mathrm{BH}} = \Phi^{-1}(1 - iq/2p)$—the $(1 - iq/2p)$th quantile of a standard normal (for information, the sequence $\boldsymbol{\lambda}^{\mathrm{BH}}$ shall play a crucial role in the rest of this paper). SLOPE is a convex program and [15] demonstrates an efficient solution algorithm (the computational cost of solving a SLOPE problem is roughly the same as that of solving the Lasso).

To gain some insights about SLOPE, it is helpful to consider the orthogonal case, which we can take to be the identity without loss of generality. When $\boldsymbol{X} = \boldsymbol{I}_p$, the SLOPE estimate is the solution to

$$\mathrm{prox}_{\boldsymbol{\lambda}}(\boldsymbol{y}) \triangleq \underset{\boldsymbol{b}}{\mathrm{argmin}} \ \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{b}\|^2 + \lambda_1|b|_{(1)} + \cdots + \lambda_p|b|_{(p)}; \tag{1.7}$$

in the literature on optimization, this solution is called the prox to the sorted ℓ_1 norm evaluated at $\boldsymbol{y}$, hence the notation in the left-hand side. (In the case of a general orthogonal design in which $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$, the SLOPE solution is $\mathrm{prox}_{\boldsymbol{\lambda}}(\boldsymbol{X}'\boldsymbol{y})$.) Suppose the observations are nonnegative and already ordered, i.e. $y_1 \ge y_2 \ge \cdots \ge y_p \ge 0$.[2] Then by [15, Proposition 2.2] SLOPE can be recast as the solution to

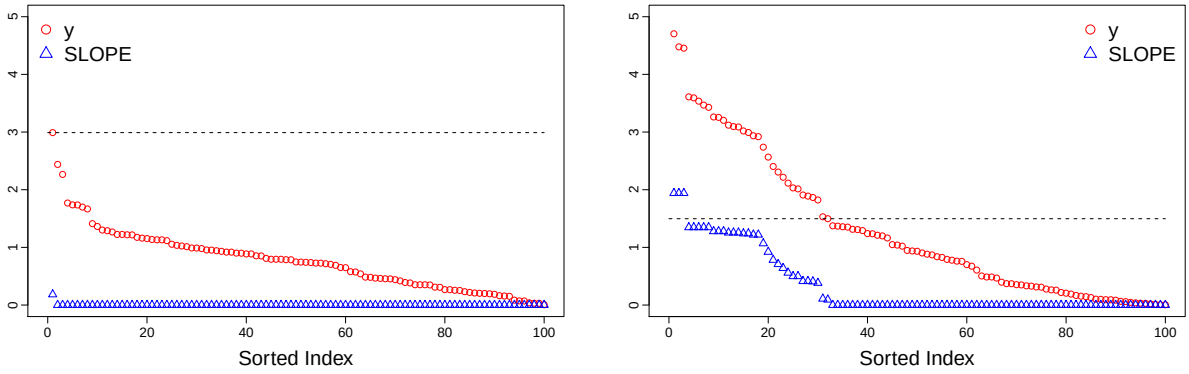
$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\lambda} - \boldsymbol{b}\|^2 = \frac{1}{2}\sum_i (y_i - \lambda_i - b_i)^2 \\ \text{subject to} & b_1 \ge b_2 \ge \cdots \ge b_p \ge 0. \end{array} \tag{1.8}$$

[2]For arbitrary data, the solution can be obtained as follows: let $\boldsymbol{P}$ be a permutation that sorts the magnitudes $|\boldsymbol{y}|$ in a non-increasing fashion. Then $\mathrm{prox}_{\boldsymbol{\lambda}}(\boldsymbol{y}) = \mathrm{sgn}(\boldsymbol{y}) \odot \boldsymbol{P}^{-1}\,\mathrm{prox}_{\boldsymbol{\lambda}}(\boldsymbol{P}|\boldsymbol{y}|)$, where $\odot$ is componentwise multiplication. In words, we can replace the observations by their sorted magnitudes, solve the problem and, finally, undo the ordering and restore the signs.

Two observations are in order: the first is that the fitted values have the same signs and ranks as the original observations; for any pair (i, j), $y_i \geq y_j$ implies that $\widehat{\beta}_i \geq \widehat{\beta}_j$. The second is that the fitted values are as close as possible to the shrunken observations $y_i - \lambda_i$ under the ordering constraint. Hence, SLOPE is a sort of soft-thresholding estimate in which the amount of thresholding is data dependent and such that the original ordering is preserved.

To emphasize the similarities with the BHq procedure, assume that we work with $\lambda_i = \sigma \cdot \lambda_i^{\mathrm{BH}}$ and that we use SLOPE as a multiple testing procedure rejecting $H_i : \beta_i = 0$ if and only if $\widehat{\beta}_i \neq 0$. Then this procedure rejects all the hypotheses the BHq step-down procedure would reject, and accepts all those the step-up procedure would accept. Under independence, i.e. $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_p)$, SLOPE controls the FDR [15], namely, $\mathrm{FDR}(\mathrm{SLOPE}) \leq q p_0/p$, where again p_0 is the number of nulls, i.e. of vanishing means.

Figure 1 displays SLOPE estimates for two distinct data sets, with one set containing many more stronger signals than the other. We see that SLOPE sets a lower threshold of significance when there is a larger number of strong signals. We can also see that SLOPE tends to shrink less as observations decrease in magnitude. In summary, SLOPE encourages sparsity just as the Lasso, but unlike the Lasso its degree of penalization is adaptive to the unknown sparsity level.



(a) Weak signals. **(b)** Strong signals.

Figure 1: Illustrative examples of original observations and SLOPE estimates with the identity design. All observations below the threshold indicated by the dotted line are set to zero; this threshold is data dependent.

1.2 Orthogonal designs

We now turn to estimation properties of SLOPE and begin by considering orthogonal designs. Multiplying both sides of (1.5) by $\boldsymbol{X}'$ gives the statistically equivalent Gaussian sequence model,

$$\boldsymbol{y} = \boldsymbol{\beta} + \boldsymbol{z},$$

where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_p)$. Estimating a sparse mean vector from Gaussian data is a well-studied problem with a long line of contributions, see [10, 26, 35, 13, 21, 40] for example. Among other things, we have already mentioned that the asymptotic risk over sparse signals is known: consider a sequence of problems in which $p \to \infty$ and $k/p \to 0$, then

$$R_p(k) \triangleq \inf_{\widehat{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \leq k} \mathbb{E} \, \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = (1 + o(1))\, 2\sigma^2 k \log(p/k),$$

where the infimum is taken over all measurable estimators, see [27] and [40]. Furthermore, both soft or hard-thresholding at the level of $\sigma\sqrt{2\log(p/k)}$ are asymptotically minimax. Such estimates require knowledge of the sparsity level ahead of time, which is not realistic. Our first result is that SLOPE also achieves asymptotic minimaxity *without* this knowledge.

Theorem 1.1. *Let $\boldsymbol{X}$ be orthogonal and assume that $p \to \infty$ with $k/p \to 0$. Fix $0 < q < 1$. Then SLOPE with $\lambda_i = \sigma \cdot \Phi^{-1}(1 - iq/2p) = \sigma \cdot \lambda_i^{\mathrm{BH}}$ obeys*

$$\sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{E}\,\|\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}} - \boldsymbol{\beta}\|^2 = (1 + o(1))\, 2\sigma^2 k \log(p/k). \tag{1.9}$$

Hence, no matter how we select the parameter q controlling the FDR level in the range $(0, 1)$, we get asymptotic minimaxity (in practice we would probably stick to values of q in the range $[0.05, 0.30]$). There are notable differences with the result from [2] we discussed earlier. First, recall that to achieve minimaxity in that work, the nominal FDR level needs to obey $q \le 1/2$ (the MSE is larger otherwise) and the sparsity level is required to obey $\log^5 p \le k \le p^{1-\delta}$ for a constant $\delta > 0$, i.e. the signal cannot be too sparse nor too dense. The lower bound on sparsity has been improved to $\log^{4.5} p$ [56]. In contrast, there are no restrictions of this nature in Theorem 1.1; this has to do with the fact that SLOPE is a continuous procedure whereas FDR thresholding is highly discontinuous; small perturbations in the data can cause the FDR thresholding estimates to jump. Second, SLOPE effortlessly extends to linear models while it is not clear how one would extend FDR thresholding ideas in a computationally tractable fashion.

One can ask which vectors $\boldsymbol{\beta}$ achieve the equality in (1.9), and it is not very hard to see that equality holds if the k nonzero entries of $\boldsymbol{\beta}$ are very large. Suppose for simplicity that $\beta_1 \gg \beta_2 \gg \cdots \gg \beta_k \gg 1$ and that $\beta_{k+1} = \cdots = \beta_p = 0$. Spacing the nonzero coefficients sufficiently far apart will insure that $y_j - \lambda_j$, $1 \le j \le k$, is nonincreasing with high probability so that the SLOPE estimate is obtained by rank-dependent soft-thresholding:

$$\widehat{\beta}_{\mathrm{SLOPE},j} = y_j - \sigma\lambda_j^{\mathrm{BH}}.$$

Informally, since the mean-square error is the sum of the squared bias and variance, this gives

$$\mathbb{E}(\widehat{\beta}_{\mathrm{SLOPE},j} - \beta_j)^2 \approx \sigma^2 \cdot ((\lambda_j^{\mathrm{BH}})^2 + 1).$$

Since $\sum_{1\le j\le k}(\lambda_j^{\mathrm{BH}})^2 = (1 + o(1))\, 2k\log(p/k)$,[3] summing this approximation over the first k coordinates gives

$$\mathbb{E}\sum_{1\le j\le k}(\widehat{\beta}_{\mathrm{SLOPE},j} - \beta_j)^2 \approx \sigma^2 \cdot \Big(k + \sum_{1\le j\le k}(\lambda_j^{\mathrm{BH}})^2\Big) = (1 + o(1))\, 2\sigma^2 k\log(p/k),$$

where the last inequality follows from the condition $k/p \to 0$. Theorem 1.1 states that in comparison, the $p - k$ vanishing means contribute a negligible MSE.

We pause here to observe that if one hopes SLOPE with weights λ_j to be minimax, then they will need to satisfy

$$\sum_{j=1}^{k}\lambda_j^2 = (1 + o(1))\, 2k\log(p/k)$$

[3] This relation follows from $\Phi^{-1}(1-c) = (1+o(1))\sqrt{2\log(1/c)}$ when $c \searrow 0$ and applying Stirling's approximation.

for all k in the stated range. This somehow implies that λ_j^2 is roughly the derivative of $f(x) = 2x\log(p/x)$ at $x = j$ yielding $\lambda_j^2 \approx 2\log p - 2\log j - 2$, or

$$\lambda_j \approx \sqrt{2\log(p/j)} \approx \Phi^{-1}(1 - jq/2p),$$

where we recall the second order approximation of normal quantiles

$$\Phi^{-1}(1 - jq/2p) \sim \sqrt{2\log\frac{p}{jq\sqrt{\log(p/jq)}}}.$$

As a remark, all our results—e.g. Theorems 1.1 and 1.2—continue to hold if we replace $\lambda_j^{\mathrm{BH}}(q)$ with $\sqrt{2\log(p/j)}$.

Just as Theorem 1.2, Theorem 1.1 extends to other loss functions. For instance, for $r \geq 1$, we have

$$\sup_{\|\boldsymbol{\beta}\|_0 \leq k} \mathbb{E}\,\|\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}} - \boldsymbol{\beta}\|_r^r = (1 + o(1)) \cdot k \cdot \left(2\sigma^2 \log(p/k)\right)^{r/2}.$$

The proof is a little more complicated than that for the squared ℓ_2 loss and is omitted. Furthermore, examining the proof of Theorem 1.1 reveals that for all k not necessarily obeying $k/p \to 0$ (e.g. $k = p/2$),

$$\frac{\sup_{\|\boldsymbol{\beta}\|_0 \leq k} \mathbb{E}\,\|\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}} - \boldsymbol{\beta}\|^2}{R_p(k)} \leq C(q),$$

where $C(q)$ is a positive numerical constant that only depends on q.

1.3 Random designs

We are interested in getting results for sparse regression that would be just as sharp and precise as those presented in the orthogonal case. In order to achieve this, we assume a tractable model in which $\boldsymbol{X}$ is a Gaussian random design with X_{ij} i.i.d. $\mathcal{N}(0, 1/n)$ so that the columns of $\boldsymbol{X}$ have just about unit norm. Random designs allow to analyze fine structures of the models of interest with tools from random matrix theory and large deviation theory, and are very popular for analyzing regression methods in the statistics literature. An incomplete list of works working with Gaussian designs would include [19, 5, 12, 20, 55, 7, 30]. On the one hand, Gaussian designs are amenable to analysis while on the other, they capture some of the features one would encounter in real applications.

To avoid any ambiguity, the theorem below considers a sequence of problems indexed by (k_j, n_j, p_j), where the number of variables $p_j \to \infty$, $k_j/p_j \to 0$ and $(k_j \log p_j)/n_j \to 0$. From now on, we shall omit the subscript.

Theorem 1.2. *Fix $0 < q < 1$ and set $\boldsymbol{\lambda} = \sigma(1+\epsilon)\boldsymbol{\lambda}^{\mathrm{BH}}(q)$ for some arbitrary constant $0 < \epsilon < 1$. Suppose $k/p \to 0$ and $(k\log p)/n \to 0$. Then*

$$\sup_{\|\boldsymbol{\beta}\|_0 \leq k} \mathbb{P}\left(\frac{\|\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}} - \boldsymbol{\beta}\|^2}{2\sigma^2 k\log(p/k)} > 1 + 3\epsilon\right) \longrightarrow 0. \tag{1.10}$$

For information, it is known that under some regularity conditions on the design [46, 54], the minimax risk is on the order of $O(\sigma^2 k \log(p/k))$, without a tight matching in the lower and upper bounds. Against this, our main result states that SLOPE, which does not use any information about the sparsity level, achieves a squared loss bounded by $(1 + o(1))\, 2\sigma^2 k \log(p/k)$ with large probability. This is the best any procedure can do as we show next.

Theorem 1.3. *Under the assumptions of Theorem 1.2, for any $\epsilon > 0$, we have*

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P}\left(\frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2\sigma^2 k \log(p/k)} > 1 - \epsilon \right) \longrightarrow 1.$$

Taken together these two results demonstrate that in a probabilistic sense $2\sigma^2 k \log(p/k)$ is the fundamental limit for the squared loss and that SLOPE achieves it. It is also likely that our methods would yield corresponding bounds for the expected squared loss but this would involve technical issues having to do with the bounding of the loss on rare events. This being said, Theorem 1.2 provides a more accurate description of the squared error than a result in expectation since it asserts that the error is at most $2\sigma^2 k \log(p/k)$ with high probability. The proof of this fact presents several novel elements not found in the literature.

The condition $(k \log p)/n \to 0$ is natural and cannot be fundamentally sharpened. To start with, our results imply that SLOPE perfectly recovers $\boldsymbol{\beta}$ in the limit of vanishing noise. In the high-dimensional setting where $p > n$, this connects with the literature on compressed sensing, which shows that in the noiseless case, $n \ge 2(1 + o(1))k \log(p/k)$ Gaussian samples are necessary for perfect recovery by ℓ_1 methods in the regime of interest [31, 32]. Our condition is a bit more stringent but naturally so since we are dealing with noisy data.

We hope that it is clear that results for orthogonal designs do not imply results for Gaussian designs because of (1) correlations between the columns of the design and (2) the high dimensionality. Under an orthogonal design, when there is no noise, one can recover $\boldsymbol{\beta}$ by just computing $\boldsymbol{X}'\boldsymbol{y}$. However, as discussed above it is far less clear how one should do this in the high-dimensional regime when $p \gg n$. As an aside, with noise it would be foolish to find $\widehat{\boldsymbol{\beta}}$ via $\mathrm{prox}_{\boldsymbol{\lambda}}(\boldsymbol{X}'\boldsymbol{y})$; that is, by applying $\boldsymbol{X}'$ and then pretending that we are dealing with an orthogonal design. Such estimates turn out to have unbounded risks.

We remark that a preprint [34] considers statistical properties of a generalization of OSCAR [17] that coincides with SLOPE. The findings and results are very different from those presented here; for instance, the selection of optimal weights λ_i is not discussed.

Finally, to see our main results under a slightly different light, suppose we get a new sample $(\boldsymbol{x}^*, y^*)$, independent from the 'training set' $(\boldsymbol{X}, \boldsymbol{y})$, obeying the linear model $y^* = \langle \boldsymbol{x}^*, \boldsymbol{\beta} \rangle + \sigma z^*$ with $\boldsymbol{x} \sim \mathcal{N}(0, n^{-1}\boldsymbol{I}_p)$ and $z^* \sim \mathcal{N}(0, \sigma^2)$. Then for any estimate $\widehat{\boldsymbol{\beta}}$, the prediction $\widehat{y} = \langle \boldsymbol{x}^*, \widehat{\boldsymbol{\beta}} \rangle$ obeys

$$\mathbb{E}(y^* - \widehat{y})^2 = n^{-1}\, \mathbb{E}\, \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2 + \sigma^2.$$

so that, in some sense, SLOPE with BH weights actually yields the best possible prediction.

1.4 Back to multiple testing

Although our emphasis is on estimation, we would nevertheless like to briefly return to the multiple testing viewpoint. In [15, 16], a series of experiments demonstrated empirical FDR control whenever $\boldsymbol{\beta}$ is sufficiently sparse. While this paper does not go as far as proving that SLOPE controls the

FDR in our Gaussian setting, the ideas underlying the proof of Theorem 1.2 have some implications for FDR control. Our discussion in this section is less formal.

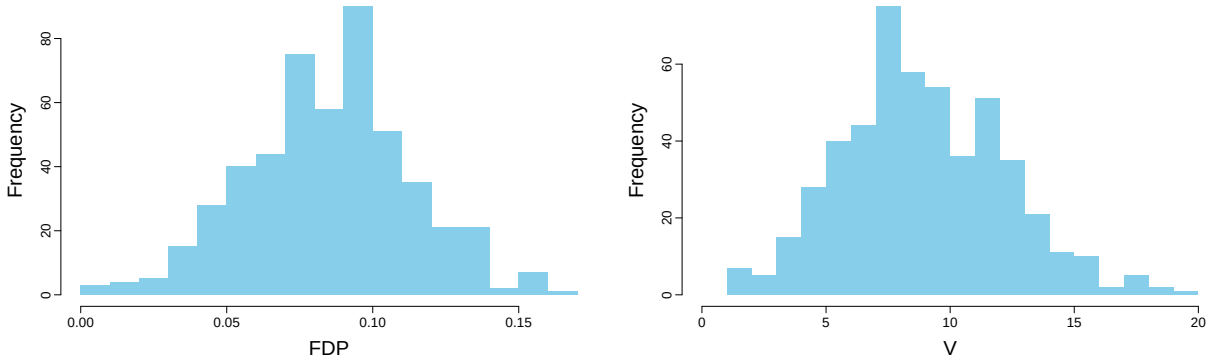
Suppose we wish to keep the false discovery proportion (FDP) $\mathrm{FDP} = V/(R \vee 1) \le q$. Since the number of true discoveries $R - V$ is at most k, the false discovery number $V = \{i : \beta_i = 0 \text{ and } \widehat{\beta}_{\mathrm{SLOPE},i} \neq 0\}$ *must* obey

$$V \le \frac{q}{1-q}\, k. \tag{1.11}$$

Interestingly, an intermediate result of the proof of Theorem 1.2 implies that (1.11) is satisfied with probability tending to one if k is sufficiently large and q is replaced by $(1 + o(1))q$. This is shown in Lemma 4.4. Another consequence of our analysis is that if the nonzero regression coefficients are larger than $1.1\,\sigma\lambda_1^{\mathrm{BH}}(q)$ (technically, we can replace 1.1 with any fixed number greater than one), then the true positive proportion (the ratio between the number of true discoveries and k) approaches one in probability. In this setup, we thus have FDR control in the sense that

$$\mathrm{FDR}_{\mathrm{SLOPE}} \le (1 + o(1))q.$$

Figure 2 demonstrates empirical FDR control at the target level $q = 0.1$. Over 500 replicates, the averaged FDR is 0.09, and the averaged false discovery number V is 9.4, as compared with 11.1, the upper bound in (1.11). We emphasize that [15, 16] also provide strong evidence that FDR is also controlled for moderate signals.



(a) Histogram of FDP. **(b)** Histogram of false discovery number V.

Figure 2: Gaussian design with $(n, p) = (8,000, 10,000)$ and $\sigma = 1$. There are $k = 100$ nonzero coefficients with amplitudes $10\sqrt{2\log p}$. Here, the nominal level is $q = 0.1$ and $\boldsymbol{\lambda} = 1.1\boldsymbol{\lambda}^{\mathrm{BH}}(0.1)$.

Since our paper proves that SLOPE does not make a large number of false discoveries, the support of $\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}}$ is of small size, and thus we see that $\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}} - \boldsymbol{\beta})\|^2$ is very nearly equal to $\|\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}} - \boldsymbol{\beta}\|^2$ since tall Gaussian matrices are near isometries. Therefore, we can carry our results over to the estimation of the mean vector $\boldsymbol{X}\boldsymbol{\beta}$.

Corollary 1.4. *Under the assumptions of Theorem 1.2,*

$$\sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P}\left(\frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2 k \log(p/k)} > 1 + 3\epsilon \right) \longrightarrow 0.$$

As before, there are matching lower bounds: for these, it suffices to restrict attention to estimates of the form $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ since projecting any estimator $\widehat{\boldsymbol{\mu}}$ onto the column space of $\boldsymbol{X}$ never increases the loss.

Corollary 1.5. *Assume $k/p \to 0$ and $p = O(n)$. Then*

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P}\left(\frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2 k \log(p/k)} > 1 - \epsilon \right) \longrightarrow 1$$

Again, SLOPE is optimal for estimating the mean response, and achieves an estimation error which is the same as that holding for the regression coefficients themselves.

1.5 Organization and notations

In the rest of the paper, we briefly explore possible alternatives to SLOPE in Section 2. Section 3 concerns the estimation properties of SLOPE under orthogonal designs and proves Theorem 1.1. We then turn to study SLOPE under Gaussian random designs in Section 4, where both Theorem 1.2 and Corollary 1.4 are proved. Last, we prove corresponding lower bounds in Section 5, including Theorem 1.3. Corollary 1.5 and auxiliary results are proved in the Appendix.

Recall that p, n, k are positive integers with $p \to \infty$, but not necessarily so for k. We use $\overline{S}$ for the complement of S. For any vector $\boldsymbol{a}$, define the support of $\boldsymbol{a}$ as $\mathrm{supp}(\boldsymbol{a}) \triangleq \{i : a_i \neq 0\}$. A bold-faced $\boldsymbol{\lambda}$ denotes a general vector obeying $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$, with at least one strict inequality. For any integer $0 < m < p$, $\boldsymbol{\lambda}^{[m]} \triangleq (\lambda_1, \ldots, \lambda_m)$ and $\boldsymbol{\lambda}^{-[m]} \triangleq (\lambda_{m+1}, \ldots, \lambda_p)$. We write $\boldsymbol{\lambda}_\epsilon$ (the superscript is omitted to save space) for the ϵ-inflated BHq critical values,

$$\lambda_{\epsilon,i} = (1+\epsilon)\lambda_i^{\mathrm{BH}} = (1+\epsilon)\Phi^{-1}\left(1 - iq/(2p)\right).$$

Last and for simplicity, $\widehat{\boldsymbol{\beta}}$ is the SLOPE estimate, unless specified otherwise.

2 Alternatives to SLOPE?

It is natural to wonder whether there are other estimators, which can potentially match the theoretical performance of SLOPE for sparse regression. Although getting an answer is beyond the scope of this paper, we pause to consider a few alternatives.

2.1 Other ℓ_1 penalized methods

The Lasso,

$$\underset{\boldsymbol{b}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda\|\boldsymbol{b}\|_1,$$

serves as a building block for a lot of sparse estimation procedures. If λ is chosen non adaptively, then a value equal to $(1-c)\cdot\sigma\sqrt{2\log p}$ for $0 < c < 1$ would cause a large number of false discoveries and, consequently, the risk when estimating sparse signals would be very high. This phenomenon can already be seen in the orthogonal case [35, 40]. This means that if we choose λ in a non-adaptive fashion then we would need to select $\lambda \ge \sigma\sqrt{2\log p}$. Under the assumptions of Theorem 1.2 and setting $\lambda = (1+c)\cdot\sigma\sqrt{2\log p}$ for an arbitrary positive constant c gives

$$\sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P}\left(\frac{\|\widehat{\boldsymbol{\beta}}_{\text{Lasso}} - \boldsymbol{\beta}\|^2}{2\sigma^2 k \log p} > 1 \right) \to 1. \tag{2.1}$$

The proof is in Appendix A.1. Hence the risk inflation does not decreases as the sparsity level k increases, whereas it does for SLOPE. Note that when $p = n$ and $k = p^{1-\delta}$,

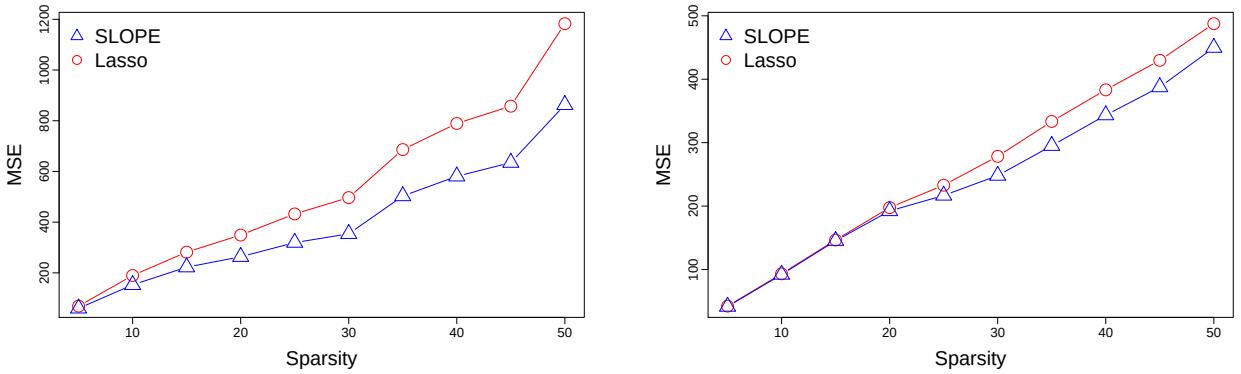
$$\frac{2\sigma^2 k \log p}{2\sigma^2 k \log(p/k)} \to \frac{1}{\delta}.$$

The reason why the Lasso is suboptimal is that the bias is too large (the fitted coefficients are shrunk too much towards zero). All in all, by our earlier considerations and by letting $\delta \to 0$ above, we conclude that no matter how we pick λ non-adaptively,

$$\frac{\text{max risk of Lasso}}{\text{max risk of SLOPE}} \to \infty$$

as $p \to \infty$.

Figure 3 compares SLOPE with Lasso estimates for both strong and moderate signals. SLOPE is more accurate than the Lasso in both cases, and the comparative advantage increases as k gets larger. This is consistent with the reasoning that SLOPE has a lower bias when k gets larger.



(a) Strong signals. **(b)** Moderate signals.

Figure 3: Gaussian design with $(n, p) = (500, 1000)$ and $\sigma = 1$. The risk $\mathbb{E}\,\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ is averaged over 100 replicates. SLOPE uses $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\mathrm{BH}}(q)$ and Lasso uses $\lambda = \lambda_1^{\mathrm{BH}}(q)$ with level $q = 0.05$. In (a), the components have magnitude $10\lambda_1^{\mathrm{BH}}$; in (b), the magnitudes are set to $0.8\lambda_1^{\mathrm{BH}}$.

Of course, one might want to select λ in a data-dependent manner, perhaps by cross-validation (see next section), or by attempting to control a type-I error such as the FDR. For instance, we could travel on the Lasso path and stop 'at some point'. Some recent procedures such as [42] make very strong assumptions about the order in which variables enter the path and are likely not to yield sharp estimation bounds such as (1.10)—provided that they can be analyzed. Others such as [38] are likely to be far too conservative.

2.2 Data-driven procedures

While finding tuning parameters adaptively is an entirely new issue, a data-driven procedure where the regularization parameter of the Lasso is chosen in an adaptive fashion would presumably boost performance. Cross-validation comes to mind whenever applicable, which is not always the case as when $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_p)$. Cross-validation techniques are also subject to variance effects and may

tend to select over-parameterized models. To make the selection of the tuning parameter as easy and accurate as possible, we work in the orthogonal setting where we have available a remarkable unbiased estimate of the risk.

SURE thresholding [28] for estimating a vector of means from $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_p)$ is a cross-validation type procedure in the sense that the thresholding parameter is selected to minimize Stein's unbiased estimate of risk (SURE) [48]. For soft-thresholding at λ, SURE reads

$$\mathrm{SURE}(\lambda) = p\sigma^2 + \sum_{i=1}^{p} y_i^2 \wedge \lambda^2 - 2\sigma^2 \#\{i : |y_i| \le \lambda\}.$$

One then applies the soft-thresholding rule at the minimizer $\widehat{\lambda}$ of $\mathrm{SURE}(\lambda)$. It has been observed [28, 21] that SURE thresholding loses performance in cases of sparse signals $\boldsymbol{\beta}$, an empirical phenomenon which can be made theoretically precise: to be sure, a forthcoming paper [49] establishes that for any fixed sparsity k, SURE thresholding obeys

$$\frac{\sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{E}\, \|\widehat{\boldsymbol{\beta}}_{\mathrm{SURE}} - \boldsymbol{\beta}\|^2}{\sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{E}\, \|\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}} - \boldsymbol{\beta}\|^2} \to \frac{k+1}{k} > 1,$$

where k is allowed to take the value zero and $(k+1)/k = \infty$ if $k = 0$. In particular, SURE has a risk that is infinitely larger than SLOPE under the global null $\boldsymbol{\beta} = \boldsymbol{0}$.

Figure 4 compares SLOPE with SURE in estimation error. In Figures 4a and 4b, we see that SURE thresholding exhibits a squared error, which is consistently larger in mean (risk) and variability. This difference is more pronounced, the sparser the signal. Figures 4c and 4d display the error distribution for $k = 1$; we see that the error of SURE thresholding is distributed over a longer range.

2.3 Variations on FDR thresholding

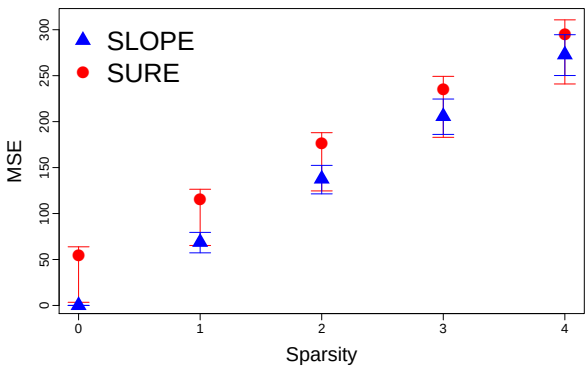
The paper [39] suggests a variation on FDR thresholding, where an adaptive soft-thresholding rule is applied instead of a hard one. This procedure achieves asymptotic minimaxity under the same assumptions as in Theorem 1.1. The issue is that this procedure is still intrinsically limited to sequence models, and cannot be generalized to linear regression. On this subject, consider the *sequential FDR thresholding* rule,

$$\widehat{\boldsymbol{\beta}}_{\mathrm{Seq},i} = \mathrm{sgn}(y_i) \cdot \left(|y_i| - \sigma \lambda^{\mathrm{BH}}_{r(i)}\right)_+$$

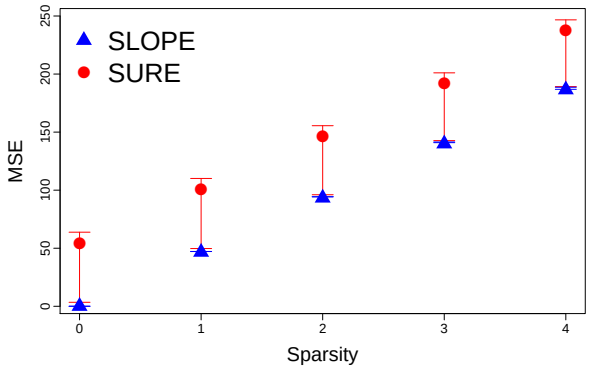
where $r(i)$ is the rank of y_i when sorting the observations by decreasing order of magnitude; that is, we apply soft-thresholding at level $\sigma\lambda_i^{\mathrm{BH}}$ to the ith largest observation (in magnitude). Under the same assumptions as in Theorem 1.1, this estimator also obeys

$$\sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{E}\, \|\widehat{\boldsymbol{\beta}}_{\mathrm{Seq}} - \boldsymbol{\beta}\|^2 = (1 + o(1))\, 2\sigma^2 k \log(p/k). \tag{2.2}$$

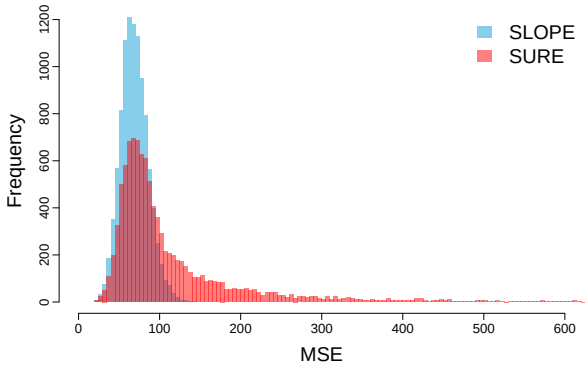
The proof is in Appendix A.1 and resembles that of Theorem 1.1. Even though the worst case performance of this estimate matches that of SLOPE, it is not a desirable procedure for at least two reasons. The first is that it is not monotone; we may have $|y_i| > |y_j|$ and $|\widehat{\beta}_j| > |\widehat{\beta}_i|$, which does not make much sense. A consequence is that it will generally have higher risk. Also note that this estimator is not continuous with respect to $\boldsymbol{y}$, since a small perturbation can change the ordering of magnitudes and, therefore, the amount of shrinkage applied to an individual component. The second reason is that this procedure does not really extend to linear models.



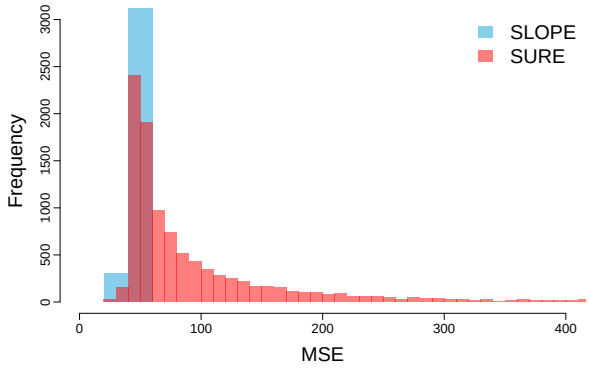
(a) Strong signals.



(b) Moderate signals.



(c) Strong signals with $k = 1$.



(d) Moderate signals with $k = 1$.

Figure 4: Empirical distributions of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ over 10,000 replicates. Strong signals have nonzero β_i set to $100\sqrt{2\log p}$ while this value is $0.8\sqrt{2\log p}$ for moderate signals. In (a) and (b), the bars represent 75% and 25% percentiles.

3 Orthogonal designs

This section proves the optimality of SLOPE under orthogonal designs. As we shall see, the proof is considerably shorter and simpler than that in [2] for FDR thresholding. One reason for this is that SLOPE continuously depends on the observation vector while FDR thresholding does not, a fact which causes serious technical difficulties. The discontinuities of the FDR hard-thresholding procedure also limits the range of its effectiveness (recall the limits on the range of sparsity levels which state that the signal cannot be too sparse or too dense) as false discoveries result in large squared errors.

A reason for separating the proof in the orthogonal case is pedagogical in that the argument is conceptually simple and, yet, some of the ideas and tools will carry over to that of Theorem 1.2. From now on and throughout the paper we set $\sigma = 1$.

3.1 Preliminaries

We collect some preliminary facts, which will prove useful, and begin with a definition used to characterize the solution to SLOPE.

Definition 3.1. *A vector $\boldsymbol{a} \in \mathbb{R}^p$ is said to majorize $\boldsymbol{b} \in \mathbb{R}^p$ if for all $i = 1, \dots, p$,*

$$|a|_{(1)} + \cdots + |a|_{(i)} \geq |b|_{(1)} + \cdots + |b|_{(i)}.$$

This differs from a more standard definition—e.g. see [43]—where the last inequality with $i = p$ is replaced by an equality (and absolute values are omitted). We see that if $\boldsymbol{a}$ majorizes $\boldsymbol{b}$ and $\boldsymbol{c}$ majorizes $\boldsymbol{d}$, then the concatenated vector $(\boldsymbol{a}, \boldsymbol{c})$ majorizes $(\boldsymbol{b}, \boldsymbol{d})$. For convenience, we list below some basic but nontrivial properties of majorization and of the prox to the sorted ℓ_1 norm. All the proofs are deferred to the Appendix.

Fact 3.1. *If $\boldsymbol{a}$ majorizes $\boldsymbol{b}$, then $\|\boldsymbol{a}\| \geq \|\boldsymbol{b}\|$.*

Fact 3.2. *If $\boldsymbol{\lambda}$ majorizes $\boldsymbol{a}$, then $\operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a}) = \boldsymbol{0}$.*

Fact 3.3. *The difference $\boldsymbol{a} - \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})$ is majorized by $\boldsymbol{\lambda}$.*

Fact 3.4. *Let T be a nonempty proper subset of $\{1, \dots, p\}$, and recall that $\boldsymbol{a}_T$ is the restriction of $\boldsymbol{a}$ to T and $\boldsymbol{\lambda}^{-[m]} = (\lambda_{m+1}, \dots, \lambda_p)$. Then*

$$\left\| [\operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})]_{\overline{T}} \right\| \leq \left\| \operatorname{prox}_{\boldsymbol{\lambda}^{-[|T|]}}\left(\boldsymbol{a}_{\overline{T}}\right) \right\|.$$

Lemma 3.1. *For any $\boldsymbol{a}$, it holds that*

$$\| \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a}) \| \leq \| (|\boldsymbol{a}| - \boldsymbol{\lambda})_+ \|,$$

where $|\boldsymbol{a}|$ is the vector of magnitudes $(|a_1|, \dots, |a_p|)$.

Proof of Lemma 3.1. The firm nonexpansiveness (e.g. see pp.131 of [45]) of the prox reads

$$\|\operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a}) - \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{b})\|^2 \leq (\boldsymbol{a} - \boldsymbol{b})' (\operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a}) - \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{b}))$$

for all $\boldsymbol{a}, \boldsymbol{b}$. Taking $\boldsymbol{b} = \operatorname{sgn}(\boldsymbol{a}) \odot \boldsymbol{\lambda}$, where $\odot$ is componentwise multiplication, and observing that $\operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{b}) = \mathbf{0}$ (Fact 3.2) give

$$\begin{aligned}\|\operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})\|^2 &\le \langle \operatorname{sgn}(\boldsymbol{a}) \odot (|\boldsymbol{a}| - \boldsymbol{\lambda}), \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a}) \rangle \\ &\le \langle (|\boldsymbol{a}| - \boldsymbol{\lambda})_+, \operatorname{sgn}(\boldsymbol{a}) \odot \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a}) \rangle \\ &\le \|(|\boldsymbol{a}| - \boldsymbol{\lambda})_+\| \cdot \|\operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})\|,\end{aligned}$$

where we use the nonnegativity of $\operatorname{sgn}(\boldsymbol{a}) \odot \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})$ and the Cauchy-Schwarz inequality. This yields the lemma. □

3.2 Proof of Theorem 1.1

Let S be the support of the vector $\boldsymbol{\beta}$, $S = \operatorname{supp}(\boldsymbol{\beta})$, and decompose the total mean-square error as

$$\mathbb{E}\,\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = \mathbb{E}\,\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S\|^2 + \mathbb{E}\,\|\widehat{\boldsymbol{\beta}}_{\overline{S}} - \boldsymbol{\beta}_{\overline{S}}\|^2,$$

i.e. as a the sum of the contributions on and off support (in case $\|\boldsymbol{\beta}\|_0 < k$, augment S to have size k). Theorem 1.1 follows from the following two lemmas.

Lemma 3.2. *Under the assumptions of Theorem 1.1, for all k-sparse vectors $\boldsymbol{\beta}$,*

$$\mathbb{E}\,\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S\|^2 \le (1 + o(1))\, 2k \log(p/k).$$

Proof. We know from Fact 3.3 that $\boldsymbol{y} - \widehat{\boldsymbol{\beta}}$ is majorized by $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\mathrm{BH}}$, which implies that $\boldsymbol{y}_S - \widehat{\boldsymbol{\beta}}_S = \boldsymbol{\beta}_S + \boldsymbol{z}_S - \widehat{\boldsymbol{\beta}}_S$ is majorized by $\boldsymbol{\lambda}^{[k]}$. The triangle inequality together with Fact 3.1 give

$$\|\boldsymbol{\beta}_S - \widehat{\boldsymbol{\beta}}_S\| = \|\boldsymbol{\beta}_S + \boldsymbol{z}_S - \widehat{\boldsymbol{\beta}}_S - \boldsymbol{z}_S\| \le \|\boldsymbol{y}_S - \widehat{\boldsymbol{\beta}}_S\| + \|\boldsymbol{z}_S\| \le \|\boldsymbol{\lambda}^{[k]}\| + \|\boldsymbol{z}_S\|.$$

This gives

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{\beta}_S - \widehat{\boldsymbol{\beta}}_S\|^2 &\le \sum_{i=1}^{k} (\lambda_i^{\mathrm{BH}})^2 + \mathbb{E}\,\|\boldsymbol{z}_S\|^2 + 2\sqrt{\sum_{1\le i\le k} (\lambda_i^{\mathrm{BH}})^2}\,\mathbb{E}\,\|\boldsymbol{z}_S\| \\ &\le \sum_{i=1}^{k} (\lambda_i^{\mathrm{BH}})^2 + \mathbb{E}\,\|\boldsymbol{z}_S\|^2 + 2\sqrt{\sum_{1\le i\le k} (\lambda_i^{\mathrm{BH}})^2\,\mathbb{E}\,\|\boldsymbol{z}_S\|^2} \\ &\le \sum_{i=1}^{k} (\lambda_i^{\mathrm{BH}})^2 + k + 2\sqrt{k \sum_{1\le i\le k} (\lambda_i^{\mathrm{BH}})^2} \\ &= (1 + o(1))\, 2k \log(p/k),\end{aligned}$$

where the last step makes use of $\sum_{1\le i\le k} (\lambda_i^{\mathrm{BH}})^2 = (1 + o(1)) 2k \log(p/k)$ and $\log(p/k) \to \infty$. □

Lemma 3.3. *Under the assumptions of Theorem 1.1, for all k-sparse vectors $\boldsymbol{\beta}$,*

$$\mathbb{E}\,\|\widehat{\boldsymbol{\beta}}_{\overline{S}} - \boldsymbol{\beta}_{\overline{S}}\|^2 = o(1)\, 2k \log(p/k). \tag{3.1}$$

Proof. It follows from Fact 3.4 that

$$\|\widehat{\boldsymbol{\beta}}_{\overline{S}}\|^2 = \left\|[\mathrm{prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_{\overline{S}}\right\|^2 \le \left\|\mathrm{prox}_{\boldsymbol{\lambda}^{-[k]}}\left(\boldsymbol{z}_{\overline{S}}\right)\right\|^2.$$

We proceed by showing that for $\boldsymbol{\zeta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{p-k})$, $\mathbb{E}\,\|\,\mathrm{prox}_{\boldsymbol{\lambda}^{-[k]}}(\boldsymbol{\zeta})\,\|^2 = o(1)\,2k\log(p/k)$. To do this, pick $A > 0$ sufficiently large such that $q(1+1/A) < 1$ in Lemmas A.3 and A.4, which then give

$$\sum_{i=1}^{p-k} \mathbb{E}\left(|\zeta|_{(i)} - \lambda^{\mathrm{BH}}_{k+i}\right)_+^2 = o(1)\,2k\log(p/k).$$

The conclusion follows from Lemma 3.1 since

$$\mathbb{E}\,\|\mathrm{prox}_{\boldsymbol{\lambda}^{-[k]}}(\boldsymbol{\zeta})\|^2 \le \sum_{i=1}^{p-k} \mathbb{E}\left(|\zeta|_{(i)} - \lambda^{\mathrm{BH}}_{k+i}\right)_+^2 = o(1)\,2k\log(p/k).$$

□

We conclude this section with a probabilistic bound on the squared loss. The proposition below, whose argument is nearly identical to that of Theorem 1.1, shall be used as a step in the proof of Theorem 1.2.

Proposition 3.4. *Fix $0 < q < 1$ and set $\boldsymbol{\lambda} = (1+\epsilon)\boldsymbol{\lambda}^{\mathrm{BH}}(q)$ for some arbitrary $0 < \epsilon < 1$. Suppose $k/p \to 0$, then for each $\delta > 0$ and all k-sparse $\boldsymbol{\beta}$,*

$$\mathbb{P}\left(\frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2(1+\epsilon)^2 k\log(p/k)} < 1+\delta\right) \to 1.$$

Here, the convergence is uniform over ϵ.

Proof. We only sketch the proof. As in the proof of Lemma 3.2, we have

$$\|\boldsymbol{\beta}_S - \widehat{\boldsymbol{\beta}}_S\| \le \|\boldsymbol{\lambda}_\epsilon^{[k]}\| + \|\boldsymbol{z}_S\|.$$

Since $\|\boldsymbol{\lambda}_\epsilon^{[k]}\| = (1+o(1))\cdot(1+\epsilon)\sqrt{2k\log(p/k)}$ and $\|z_S\| = o_{\mathbb{P}}(\sqrt{2k\log(p/k)})$, we have that for each $\delta > 0$,

$$\mathbb{P}\left(\frac{\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S\|^2}{2(1+\epsilon)^2 k\log(p/k)} < 1+\delta/2\right) \to 1.$$

Since $\boldsymbol{\lambda}$ has increased, it is only natural that the off-support error remains under control. In fact, (3.1) still holds, and the Markov inequality then gives

$$\mathbb{P}\left(\frac{\|\widehat{\boldsymbol{\beta}}_{\overline{S}} - \boldsymbol{\beta}_{\overline{S}}\|^2}{2k\log(p/k)} < \frac{\delta}{2}\right) \to 1.$$

This concludes the proof. □

4 Gaussian random designs

When moving from an orthogonal to a non-orthogonal design, the correlations between the columns of $\boldsymbol{X}$ and the high dimensionality create much difficulty. This is already apparent when scanning the literature on penalized sparse estimation procedures such as the Lasso, SCAD [33], the Dantzig selector [24] and MC+ [58], see e.g. [37, 23, 59, 22, 11, 52, 55, 44, 57, 7, 29] for a highly incomplete list of references. For example, a statistical analysis of the Lasso often relies on several ingredients: first, the Karush-Kuhn-Tucker (KKT) optimality conditions; second, appropriate assumptions about the designs such as the Gaussian model we use here, which guarantee a form of local orthogonality (known under the name of restricted isometries or restricted eigenvalue conditions); third, the selection of a penalty λ several times the size of the universal threshold $\sigma\sqrt{2\log p}$, which while introducing a large bias yielding MSEs that cannot possibly approach the precise bounds we develop in this paper, facilitates the analysis since it effectively sets many coordinates to zero.

Our approach must be different for at least two reasons. To begin with, the KKT conditions for SLOPE are not easy to manipulate. Leaving out this technical matter, a more substantial difference is that the SLOPE regularization is far weaker than that of a Lasso model with a large value of the regularization parameter λ. To appreciate this distinction, consider the *orthogonal design* setting. In such a simple situation, it is straightforward to obtain error estimates about a hard thresholding rule set at—or several times—the Bonferroni level. Getting sharp estimates for FDR thresholding is entirely a different matter, compare the 46-page proof in [2]!

4.1 Architecture of the proof

Our aim in this section is to provide a general overview of the proof, explaining the key novel ideas and intermediate results. At a high level, the general structure is fairly simple and is as follows:

1. Exhibit an ideal estimator $\widetilde{\boldsymbol{\beta}}$, which is easy to analyze and achieves the optimal squared error loss with high probability.

2. Prove that the SLOPE estimate $\widehat{\boldsymbol{\beta}}$ is close to this ideal estimate.

We discuss these in turn and recall that throughout, $\boldsymbol{\lambda} = (1+\epsilon)\boldsymbol{\lambda}^{\mathrm{BH}}(q)$.

A solution algorithm for SLOPE is the proximal gradient method, which operates as follows: starting from an initial guess $\boldsymbol{b}^{(0)} \in \mathbb{R}^p$, inductively define

$$\boldsymbol{b}^{(m+1)} = \mathrm{prox}_{t_m \boldsymbol{\lambda}}\left(\boldsymbol{b}^{(m)} - t_m \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{b}^{(m)} - \boldsymbol{y})\right),$$

where $\{t_m\}$ is an appropriate sequence for step sizes. It is empirically observed that under sparsity constraints, the proximal gradient algorithm for SLOPE (and Lasso) converges quickly provided we start from a good initial point. Here, we propose approximating the SLOPE solution by starting from the ground truth and applying just one iteration; that is, with $t_0 = 1$, define

$$\widetilde{\boldsymbol{\beta}} := \mathrm{prox}_{\boldsymbol{\lambda}}\left(\boldsymbol{\beta} + \boldsymbol{X}'\boldsymbol{z}\right). \tag{4.1}$$

This oracle estimator $\widetilde{\boldsymbol{\beta}}$ approximates the SLOPE estimator $\widehat{\boldsymbol{\beta}}$ well—they are equal when the design is orthogonal—and has statistical properties far easier to understand. The lemma below is the subject of Section 4.2.

Lemma 4.1. *Under the assumptions of Theorem 1.2, for all k-sparse $\boldsymbol{\beta}$, we have*

$$\mathbb{P}\left(\frac{\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|^2}{(1+\epsilon)^2\,2k\log(p/k)}<1+\delta\right)\to 1,$$

where $\delta>0$ is an arbitrary constant.

Since we know that $\widetilde{\boldsymbol{\beta}}$ is asymptotically optimal, it suffices to show that the squared distance between $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ is negligible in comparison to that between $\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. This captured by the result below, whose proof is the subject of Section 4.3.

Lemma 4.2. *Let $T\subset\{1,\dots,p\}$ be a subset of columns assumed to contain the supports of $\widehat{\boldsymbol{\beta}}$, $\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$; i.e. $T\supset \operatorname{supp}(\widehat{\boldsymbol{\beta}})\cup\operatorname{supp}(\widetilde{\boldsymbol{\beta}})\cup\operatorname{supp}(\boldsymbol{\beta})$. Suppose all the eigenvalues of $\boldsymbol{X}_T'\boldsymbol{X}_T$ lie in $[1-\delta,1+\delta]$ for some $\delta<1/2$. Then*

$$\|\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\|^2\le\frac{3\delta}{1-2\delta}\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|^2.$$

In particular, $\widehat{\boldsymbol{\beta}}=\widetilde{\boldsymbol{\beta}}$ under orthogonal designs.

We thus see that everything now comes down to showing that there is a set of small cardinality containing the supports of $\widehat{\boldsymbol{\beta}}$, $\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. While it is easy to show that $\operatorname{supp}(\widetilde{\boldsymbol{\beta}})\cup\operatorname{supp}(\boldsymbol{\beta})$ is of small cardinality, it is delicate to show that this property still holds with the addition of the support of the SLOPE estimate. Below, we introduce the *resolvent set*, which will prove to contain $\operatorname{supp}(\widehat{\boldsymbol{\beta}})\cup\operatorname{supp}(\widetilde{\boldsymbol{\beta}})\cup\operatorname{supp}(\boldsymbol{\beta})$ with high probability.

Definition 4.1 (Resolvent set)**.** *Fix $S=\operatorname{supp}(\boldsymbol{\beta})$ of cardinality at most k, and an integer $k^\star$ obeying $k<k^\star<p$. The set $S^\star=S^\star(S,k^\star)$ is said to be a resolvent set if it is the union of S and the $k^\star-k$ indices with the largest values of $|\boldsymbol{X}_i'\boldsymbol{z}|$ among all $i\in\{1,\dots,p\}\setminus S$.*

Under the assumptions of Theorem 1.2, we shall see in Section 4.4 that we can choose $k^\star$ in such a way that on the one hand $k^\star$ is sufficiently small compared to p and $n/\log p$, and on the other, the resolvent set $S^\star$ is still expected to contain $\operatorname{supp}(\widetilde{\boldsymbol{\beta}})$ (easier) and $\operatorname{supp}(\widehat{\boldsymbol{\beta}})$ (more difficult). Formally, Lemma 4.4 below shows that

$$\inf_{\|\boldsymbol{\beta}\|_0\le k}\mathbb{P}\left(\operatorname{supp}(\boldsymbol{\beta})\cup\operatorname{supp}(\widehat{\boldsymbol{\beta}})\cup\operatorname{supp}(\widetilde{\boldsymbol{\beta}})\subset S^\star\right)\to 1. \tag{4.2}$$

One can view the resolvent solution as a sophisticated type of a dual certificate method, better known as primal-dual witness method [55, 22, 47] in the statistics literature. A significant gradation in the difficulty of detecting the support of the SLOPE solution a priori comes from the false discoveries we commit because we happen to live on the edge, i.e. work with a procedure as liberal as can be.

With (4.2) in place, Theorem 1.2 merely follows from Lemma 4.2 and the accuracy of $\widetilde{\boldsymbol{\beta}}$ explained by Lemma 4.1; all the bookkeeping is in Section 4.5. Furthermore, Corollary 1.4 is just one stone throw away, please also see Section 4.5 for all the necessary details.

4.2 One-step approximation

The proof of Lemma 4.1 is an immediate consequence from Proposition 3.4. In brief, Borell's inequality—see Lemma A.5—provides a well-known deviation bound about chi-square random variables, namely,

$$\mathbb{P}\left(\|\boldsymbol{z}\| \le (1+\epsilon)\sqrt{n}\right) \ge 1 - \mathrm{e}^{-\epsilon^2 n/2} \to 1$$

since $\epsilon^2 n \to \infty$. Hence, to prove our claim, it suffices to establish that

$$\mathbb{P}\left(\frac{\|\operatorname{prox}_{\boldsymbol{\lambda}_\epsilon}(\boldsymbol{\beta} + \boldsymbol{X}'\boldsymbol{z}) - \boldsymbol{\beta}\|^2}{(1+\epsilon)^2 2k\log(p/k)} < 1+\delta \;\Big|\; \|\boldsymbol{z}\| \le (1+\epsilon)\sqrt{n}\right) \to 1. \tag{4.3}$$

Conditional on $\|\boldsymbol{z}\| = c\sqrt{n}$ for some $0 < c \le 1+\epsilon$, $\boldsymbol{X}'\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, c^2\boldsymbol{I}_p)$ and, therefore, conditionally,

$$\begin{aligned}\|\operatorname{prox}_{\boldsymbol{\lambda}_\epsilon}\left(\boldsymbol{\beta} + \boldsymbol{X}'\boldsymbol{z}\right) - \boldsymbol{\beta}\| &\stackrel{d}{=} \|\operatorname{prox}_{\boldsymbol{\lambda}_\epsilon}\left(\boldsymbol{\beta} + c\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)\right) - \boldsymbol{\beta}\| \\ &= c\|\operatorname{prox}_{\boldsymbol{\lambda}_{\epsilon'}}\left(\boldsymbol{\beta}/c + \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)\right) - \boldsymbol{\beta}/c\|\end{aligned}$$

for $\epsilon' = (1+\epsilon)/c - 1 \ge 0$. Hence, Proposition 3.4 gives

$$\mathbb{P}\left(\frac{\|\operatorname{prox}_{\boldsymbol{\lambda}_{\epsilon'}}\left(\boldsymbol{\beta}/c + \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)\right) - \boldsymbol{\beta}/c\|^2}{(1+\epsilon')^2\, 2k\log(p/k)} < 1+\delta\right) \to 1.$$

Since $(1+\epsilon)^2/c^2 = (1+\epsilon')^2$, this is equivalent to

$$\mathbb{P}\left(\frac{c^2\|\operatorname{prox}_{\boldsymbol{\lambda}_{\epsilon'}}\left(\boldsymbol{\beta}/c + \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)\right) - \boldsymbol{\beta}/c\|^2}{(1+\epsilon)^2\, 2k\log(p/k)} < 1+\delta\right) \to 1.$$

This completes the proof since we can deduce (4.3) by averaging over $\|\boldsymbol{z}\|$.

4.3 $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ are close when $\boldsymbol{X}$ is nearly orthogonal

We prove Lemma 4.2 in the case where $T = \{1, \dots, p\}$, first. Set $J_{\boldsymbol{\lambda}}(\boldsymbol{b}) = \sum_{1\le i\le p}\lambda_i|b|_{(i)}$, by definition $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ respectively minimize

$$\begin{aligned}L_1(\boldsymbol{b}) &:= \tfrac{1}{2}\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{b})\|^2 + \boldsymbol{z}'\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{b}) + J_{\boldsymbol{\lambda}}(\boldsymbol{b}) \\ L_2(\boldsymbol{b}) &:= \tfrac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{b}\|^2 + \boldsymbol{z}'\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{b}) + J_{\boldsymbol{\lambda}}(\boldsymbol{b}).\end{aligned}$$

Next the assumptions about the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ implies that these two functions are related,

$$\begin{aligned}L_2(\widetilde{\boldsymbol{\beta}}) - \frac{\delta}{2}\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|^2 \le L_1(\widetilde{\boldsymbol{\beta}}) \le L_2(\widetilde{\boldsymbol{\beta}}) + \frac{\delta}{2}\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|^2, \\ L_2(\widehat{\boldsymbol{\beta}}) - \frac{\delta}{2}\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2 \le L_1(\widehat{\boldsymbol{\beta}}) \le L_2(\widehat{\boldsymbol{\beta}}) + \frac{\delta}{2}\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2.\end{aligned}$$

Chaining these inequalities gives

$$L_2(\widetilde{\boldsymbol{\beta}}) + \frac{\delta\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|^2}{2} \ge L_1(\widetilde{\boldsymbol{\beta}}) \ge L_1(\widehat{\boldsymbol{\beta}}) \ge L_2(\widehat{\boldsymbol{\beta}}) - \frac{\delta\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2}{2}. \tag{4.4}$$

Now the strong convexity of L_2 also gives

$$L_2(\widehat{\boldsymbol{\beta}}) \geq L_2(\widetilde{\boldsymbol{\beta}}) + \frac{\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|^2}{2},$$

and plugging this in the right-hand side of (4.4) yields

$$\frac{\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|^2}{2} - \frac{\delta\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2}{2} \leq \frac{\delta\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|^2}{2}. \tag{4.5}$$

Since $\delta\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2/2 \leq \delta\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|^2 + \delta\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|^2$ (this is essentially the basic inequality $(a+b)^2 \leq 2a^2 + 2b^2$), the conclusion follows.

We now consider the general case. Let m be the cardinality of T and for $\boldsymbol{b} \in \mathbb{R}^m$, set $J_{\boldsymbol{\lambda}^{[m]}}(\boldsymbol{b}) = \sum_{1\leq i\leq m} \lambda_i |b|_{(i)}$, and observe that by assumption, $\widehat{\boldsymbol{\beta}}_T$ and $\widetilde{\boldsymbol{\beta}}_T$ are solutions to the reduced problems

$$\operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^{|T|}} \quad \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}_T\boldsymbol{b}\|^2 + J_{\boldsymbol{\lambda}^{[m]}}(\boldsymbol{b}) \tag{4.6}$$

and

$$\operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^{|T|}} \quad \tfrac{1}{2}\|\boldsymbol{\beta}_T + \boldsymbol{X}_T'\boldsymbol{z} - \boldsymbol{b}\|^2 + J_{\boldsymbol{\lambda}^{[m]}}(\boldsymbol{b}).$$

Using the fact that $\boldsymbol{X}\beta = \boldsymbol{X}_T\beta_T$, we see that $\widehat{\boldsymbol{\beta}}_T$ and $\widetilde{\boldsymbol{\beta}}_T$ respectively minimize

$$\begin{aligned} L_1(\boldsymbol{b}) &:= \tfrac{1}{2}\|\boldsymbol{X}_T(\boldsymbol{\beta}_T - \boldsymbol{b})\|^2 + \boldsymbol{z}'\boldsymbol{X}_T(\boldsymbol{\beta}_T - \boldsymbol{b}) + J_{\boldsymbol{\lambda}^{[m]}}(\boldsymbol{b}) \\ L_2(\boldsymbol{b}) &:= \tfrac{1}{2}\|\boldsymbol{\beta}_T - \boldsymbol{b}\|^2 + \boldsymbol{z}'\boldsymbol{X}_T(\boldsymbol{\beta}_T - \boldsymbol{b}) + J_{\boldsymbol{\lambda}^{[m]}}(\boldsymbol{b}). \end{aligned}$$

From now on, the proof is just as before.

4.4 Support localization

Below we write $\boldsymbol{a} \preceq \boldsymbol{b}$ as a short-hand for $\boldsymbol{b}$ majorizes $\boldsymbol{a}$ and

$$S^\diamond = \operatorname{supp}(\boldsymbol{\beta}) \cup \operatorname{supp}(\widehat{\boldsymbol{\beta}}) \cup \operatorname{supp}(\widetilde{\boldsymbol{\beta}}). \tag{4.7}$$

Lemma 4.3 (Reduced SLOPE)**.** *Let $\widehat{\boldsymbol{b}}_T$ be the solution to the reduced SLOPE problem* (4.6)*, which only fits regression coefficients with indices in T. If*

$$\boldsymbol{X}_{\overline{T}}'(\boldsymbol{y} - \boldsymbol{X}_T\widehat{\boldsymbol{b}}_T) \preceq \boldsymbol{\lambda}^{-[|T|]}, \tag{4.8}$$

then it is the solution to the full SLOPE problem in the sense that $\widehat{\boldsymbol{\beta}}$ defined as $\widehat{\boldsymbol{\beta}}_T = \widehat{\boldsymbol{b}}_T$ and $\widehat{\boldsymbol{\beta}}_{\overline{T}} = \boldsymbol{0}$ is solution.

The inequality (4.8), which implies localization of the solution, reminds us of a similar condition for the Lasso. In particular, if $\lambda_1 = \lambda_2 = \cdots = \lambda_p$, then SLOPE is the Lasso and (4.8) is equivalent to $\|\boldsymbol{X}_{\overline{T}}'(\boldsymbol{y} - \boldsymbol{X}_T\widehat{\boldsymbol{b}}_T)\|_\infty \leq \lambda$. In this case, it is well known that this implies that a solution to the Lasso is supported on T, see e.g. [55, 22, 47].

The main result of this section is this:

Lemma 4.4. *Suppose*

$$k^\star \geq \max\left\{\frac{1+c}{1-q}\, k, k+d\right\}$$

for an arbitrary small constant $c > 0$, where d is a deterministic sequence diverging to infinity[4] *in such a way that $k^\star/p \to 0$ and $(k^\star \log p)/n \to 0$. Then*

$$\inf_{\|\boldsymbol{\beta}\|_0 \leq k} \mathbb{P}\left(S^\diamond \subset S^\star\right) \to 1.$$

Proof of Lemma 4.4. By construction, $\mathrm{supp}(\boldsymbol{\beta}) \subset S^\star$ so we only need to show (i) $\mathrm{supp}(\widehat{\boldsymbol{\beta}}) \subset S^\star$ and (ii) $\mathrm{supp}(\widetilde{\boldsymbol{\beta}}) \subset S^\star$. We begin with (i). By Lemma 4.3, $\mathrm{supp}(\widehat{\boldsymbol{\beta}})$ is contained in $S^\star$ if

$$\boldsymbol{X}'_{\overline{S^\star}}(\boldsymbol{y} - \boldsymbol{X}_{S^\star}\widehat{\boldsymbol{\beta}}_{S^\star}) \preceq \boldsymbol{\lambda}_\epsilon^{-[k^\star]},$$

which would follow from

$$\boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{X}_{S^\star}(\boldsymbol{\beta}_{S^\star} - \widehat{\boldsymbol{\beta}}_{S^\star}) \preceq \frac{\epsilon}{2}\left(\lambda^{\mathrm{BH}}_{k^\star+1}, \ldots, \lambda^{\mathrm{BH}}_p\right) \tag{4.9}$$

and

$$\boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{z} \preceq (1+\epsilon/2)\left(\lambda^{\mathrm{BH}}_{k^\star+1}, \ldots, \lambda^{\mathrm{BH}}_p\right). \tag{4.10}$$

Lemma A.12 in Appendix concludes that (4.9) holds with probability tending to one, since, by assumption, $\epsilon > 0$ is constant and $\sqrt{(k^\star \log p)/n} \to 0$. To show that (4.10) also holds with probability approaching one, we resort to Lemma A.9. Conditional on $\boldsymbol{z}$, $\boldsymbol{X}'_{\overline{S}}\boldsymbol{z} \sim \mathcal{N}(0, \|\boldsymbol{z}\|^2/n \cdot \boldsymbol{I}_{p-k})$. By definition, $\boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{z}$ is formed from $\boldsymbol{X}'_{\overline{S}}\boldsymbol{z}$ by removing its $k^\star - k$ largest entries in absolute value. Denoting by $\zeta_1, \ldots, \zeta_{p-k}$ i.i.d. standard Gaussian random variables, (4.10) thus boils down to

$$\left(|\zeta|_{(k^\star-k+1)}, |\zeta|_{(k^\star-k+2)}, \ldots, |\zeta|_{(p-k)}\right) \preceq \frac{(1+\epsilon/2)\sqrt{n}}{\|\boldsymbol{z}\|}\left(\lambda^{\mathrm{BH}}_{k^\star+1}, \ldots, \lambda^{\mathrm{BH}}_p\right). \tag{4.11}$$

Borell's inequality (Lemma A.5) gives

$$\mathbb{P}\left((1+\epsilon/2)\sqrt{n}/\|\boldsymbol{z}\| < 1\right) = \mathbb{P}\left(\|\boldsymbol{z}\| - \sqrt{n} > \epsilon\sqrt{n}/2\right) \leq \mathrm{e}^{-n\epsilon^2/8} \to 0.$$

The conclusion follows from Lemma A.9.

We turn to (ii) and note that

$$(\boldsymbol{\beta} + \boldsymbol{X}'\boldsymbol{z})_{\overline{S^\star}} = \boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{z}.$$

Now our previous analysis implies $\boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{z} \preceq \boldsymbol{\lambda}_\epsilon^{-[k^\star]}$ with probability tending to one. However, it follows from Facts 3.4 and 3.2 that

$$\|\widetilde{\boldsymbol{\beta}}_{\overline{S^\star}}\| = \|\operatorname{prox}_{\boldsymbol{\lambda}_\epsilon}\left(\boldsymbol{\beta} + \boldsymbol{X}'\boldsymbol{z}\right)_{\overline{S^\star}}\| \leq \|\operatorname{prox}_{\boldsymbol{\lambda}_\epsilon^{-[k^\star]}}\left(\boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{z}\right)\| = \boldsymbol{0}.$$

In summary, $\boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{z} \preceq \boldsymbol{\lambda}_\epsilon^{-[k^\star]} \implies \mathrm{supp}(\widetilde{\boldsymbol{\beta}}) \subset S^\star$. This concludes the proof. □

[4]Recall that we are considering a sequence of problems with (k_j, n_j, p_j) so that this is saying that $k_j^\star \geq \max(2(1-q)^{-1}k_j, k_j + d_j)$ with $d_j \to \infty$.

4.5 Proof of Theorem 1.2 and Corollary 1.4

Put

$$\delta = \frac{1+3\epsilon}{(1+\epsilon)^2} - 1 = \frac{\epsilon - \epsilon^2}{(1+\epsilon)^2} > 0,$$

and choose any $\delta' > 0$ such that

$$(1+\delta')\left(\sqrt{3\delta'/(1-2\delta')}+1\right)^2(1+\delta/2) < (1+\delta).$$

Let $\mathscr{A}_1$ be the event $S^\diamond \subset S^\star$, $\mathscr{A}_2$ that all the singular values of $\boldsymbol{X}_{S^\star}$ lie in $\left[\sqrt{1-\delta'}, \sqrt{1+\delta'}\right]$, and $\mathscr{A}_3$ that

$$\frac{\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{(1+\epsilon)^2\, 2k\log(p/k)} < 1 + \frac{\delta}{2}.$$

We prove that each event happens with probability tending to one. For $\mathscr{A}_1$, use Lemma 4.4, and set

$$d = \min\left\{\left\lfloor \sqrt{kn/\log p}\right\rfloor, \lfloor\sqrt{p}\rfloor\right\},$$

which diverges to ∞, and

$$k^\star = \max\left\{\lceil 2k/(1-q)\rceil, k+d\right\}.$$

It is easy to see that $k^\star$ satisfies the assumptions of Lemma 4.4, which asserts that $\mathbb{P}(\mathscr{A}_1) \to 1$ uniformly over all k-sparse $\boldsymbol{\beta}$. For $\mathscr{A}_2$, since $(k^\star \log p)/n \to 0$ implies that $k^\star \log(p/k^\star)/n \to 0$, then taking t sufficiently small in Lemma A.11 gives $\mathbb{P}(\mathscr{A}_2) \to 1$ uniformly over all k-sparse $\boldsymbol{\beta}$. Finally, $\mathbb{P}(\mathscr{A}_3) \to 1$ also uniformly over all k-sparse $\boldsymbol{\beta}$ by Lemma 4.1 since $\epsilon^2 n \to \infty$.

Hence, $\mathbb{P}(\mathscr{A}_1 \cap \mathscr{A}_2 \cap \mathscr{A}_3) \to 1$ uniformly over all $\boldsymbol{\beta}$ with sparsity at most k. Consequently, it suffices to show that on this intersection,

$$\frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2k\log(p/k)} < 1+3\epsilon, \quad \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2k\log(p/k)} < 1+3\epsilon.$$

On $\mathscr{A}_2 \cap \mathscr{A}_3$, all the eigenvalues values of $\boldsymbol{X}'_{S^\diamond}\boldsymbol{X}_{S^\diamond}$ are between $1-\delta'$ and $1+\delta'$. By definition, all the coordinates of $\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ vanish outside of $S^\diamond$. Thus, Lemma 4.2 gives

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| \le \|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\| + \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\| &\le \left(\sqrt{\frac{3\delta'}{1-2\delta'}} + 1\right)\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\| \\ &\le \left(\frac{1+\delta}{(1+\delta/2)(1+\delta')}\right)^{1/2}\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|.\end{aligned}$$

Hence, on $\mathscr{A}_1 \cap \mathscr{A}_2 \cap \mathscr{A}_3$, we have

$$\frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2k\log(p/k)} \le \frac{1+\delta}{(1+\delta/2)(1+\delta')} \cdot \frac{\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2k\log(p/k)} < \frac{(1+\delta)(1+\epsilon)^2}{1+\delta'} < 1+3\epsilon,$$

and similarly,

$$\frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2k\log(p/k)} \le (1+\delta')\frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2k\log(p/k)} < (1+\delta')\frac{(1+\delta)(1+\epsilon)^2}{1+\delta'} = 1+3\epsilon.$$

This finishes the proof.

5 Lower bounds

We here prove Theorem 1.3, the lower matching bound for Theorem 1.2, and leave the proof of Corollary 1.5 to Appendix A.4. Once again, we warm up with the orthogonal design and develop tools that can be readily applied to the regression case.

5.1 Orthogonal designs

Suppose $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{I}_p)$. The first result states that in this model, the squared loss for estimating 1-sparse vectors cannot be lower than $2\log p$. The proof is in Appendix A.4.

Lemma 5.1. *Let $\tau_p = (1+o(1))\sqrt{2\log p}$ be a sequence obeying $\sqrt{2\log p} - \tau_p \to \infty$. Consider the prior $\boldsymbol{\pi}$ for $\boldsymbol{\beta}$, which selects a coordinate i uniformly at random in $\{1,\dots,p\}$, and sets $\beta_i = \tau_p$ and $\beta_j = 0$ for $j \neq i$. For each $\epsilon > 0$,*

$$\inf_{\widehat{\boldsymbol{\beta}}} \mathbb{P}_{\boldsymbol{\pi}} \left(\frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2\log p} > 1-\epsilon \right) \to 1.$$

Next, we state a counterpart to Theorem 1.1, whose proof constructs k independent 1-sparse recovery problems.

Proposition 5.2. *Suppose $k/p \to 0$. Then for any $\epsilon > 0$, we have*

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P} \left(\frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2k\log(p/k)} > 1-\epsilon \right) \to 1.$$

Proof. The fundamental duality between 'min max' and 'max min' gives

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P} \left(\frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2k\log(p/k)} > 1-\epsilon \right) \ge \sup_{\|\widetilde{\boldsymbol{\pi}}\|_0 \le k} \inf_{\widehat{\boldsymbol{\beta}}} \mathbb{P}_{\widetilde{\boldsymbol{\pi}}} \left(\frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2k\log(p/k)} > 1-\epsilon \right).$$

Above, $\widetilde{\boldsymbol{\pi}}$ denotes any distribution on $\mathbb{R}^p$ such that any realization $\boldsymbol{\beta}$ obeys $\|\boldsymbol{\beta}\|_0 \le k$, and $\mathbb{P}_{\widetilde{\boldsymbol{\pi}}}(\cdot)$ emphasizes that $\boldsymbol{\beta}$ follows the prior $\widetilde{\boldsymbol{\pi}}$, as earlier in Lemma 5.1. It is therefore sufficient to construct a prior $\widetilde{\boldsymbol{\pi}}$ with a right-hand side approaching one.

Assume p is a multiple of k (otherwise, replace p with $p_0 = k\lfloor p/k \rfloor$ and let $\boldsymbol{\pi}$ be supported on $\{1,\dots,p_0\}$). Partition $\{1,\dots,p\}$ into k consecutive blocks $\{1,\dots,p/k\}$, $\{p/k+1,\dots,2p/k\}$ and so on. Our prior is a product prior, where on each block, we select a coordinate uniformly at random and sets its amplitude to $\tau = (1+o(1))\sqrt{\log(p/k)}$ and $\sqrt{2\log(p/k)} - \tau \to \infty$. Next, let $\widehat{\boldsymbol{\beta}}$ be any estimator and write the loss $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = L_1 + \cdots + L_k$, where L_j is the contribution from the jth block. The lemma is reduced to proving

$$\inf_{\widehat{\boldsymbol{\beta}}} \mathbb{P}_{\boldsymbol{\pi}} \left(\frac{L_1 + \cdots + L_k}{2k\log(p/k)} > 1-\epsilon \right) \to 1. \tag{5.1}$$

For any constant $\epsilon' > 0$, since $p/k \to \infty$, Lemma 5.1 claims that

$$\inf_{\widehat{\boldsymbol{\beta}}} \mathbb{P}_{\boldsymbol{\pi}} \left(\frac{L_j}{2\log(p/k)} > 1-\epsilon' \right) \to 1 \tag{5.2}$$

uniformly over $j = 1, \dots, k$ since distinct blocks are stochastically independent. Set

$$\bar{L}_j = \min\{L_j, 2\log(p/k)\} \le L_j.$$

On one hand,

$$\frac{\mathbb{E}\left(\bar{L}_1 + \cdots + \bar{L}_k\right)}{2k\log(p/k)} \le (1-\epsilon)\cdot \mathbb{P}_{\boldsymbol{\pi}}\left(\frac{\bar{L}_1 + \cdots + \bar{L}_k}{2k\log(p/k)} \le 1-\epsilon\right) + \mathbb{P}_{\boldsymbol{\pi}}\left(\frac{\bar{L}_1 + \cdots + \bar{L}_k}{2k\log(p/k)} > 1-\epsilon\right).$$

On the other,

$$\frac{\mathbb{E}\left(\bar{L}_1 + \cdots + \bar{L}_k\right)}{2k\log(p/k)} \ge \frac{1-\epsilon'}{k}\sum_{j=1}^{k} \mathbb{P}_{\boldsymbol{\pi}}\left(\frac{\bar{L}_j}{2\log(p/k)} > 1-\epsilon'\right).$$

All in all, this gives

$$\sup_{\widehat{\boldsymbol{\beta}}} \mathbb{P}_{\boldsymbol{\pi}}\left(\frac{\bar{L}_1 + \cdots + \bar{L}_k}{2k\log(p/k)} \le 1-\epsilon\right) \le \frac{1}{\epsilon}\cdot\left(1 - (1-\epsilon')\inf_{\widehat{\boldsymbol{\beta}},j} \mathbb{P}_{\boldsymbol{\pi}}\left(\frac{\bar{L}_j}{2\log(p/k)} > 1-\epsilon'\right)\right).$$

Finally, take the limit $p \to \infty$ in the above inequality. Since $\bar{L}_j/(2\log(p/k)) > 1-\epsilon'$ if and only if $L_j/(2\log(p/k)) > 1-\epsilon'$, it follows from (5.2) that

$$\limsup_{p\to\infty} \sup_{\widehat{\boldsymbol{\beta}}} \mathbb{P}_{\boldsymbol{\pi}}\left(\frac{\bar{L}_1 + \cdots + \bar{L}_k}{2k\log(p/k)} \le 1-\epsilon\right) \le \frac{\epsilon'}{\epsilon}.$$

We conclude by taking $\epsilon' \to 0$. □

5.2 Random designs

We return to the regression setup $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_p)$, where $\boldsymbol{X}$ is our Gaussian design.

Lemma 5.3. *Fix $\alpha \le 1$ and*

$$\tau_{p,n} = \left(\sqrt{2\log p} - \log\sqrt{2\log p}\right)\left(1 - 2\sqrt{(\log p)/n}\right).$$

Let $\boldsymbol{\pi}$ be the prior from Lemma 5.1 with amplitude set to $\alpha\cdot\tau_{n,p}$. Assume $(\log p)/n \to 0$. Then for any $\epsilon > 0$,

$$\inf_{\widehat{\boldsymbol{\beta}}} \mathbb{P}_{\boldsymbol{\pi}}\left(\frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{\alpha^2\cdot 2\log p} > 1-\epsilon\right) \to 1.$$

With this, we are ready to prove a stronger version of Theorem 1.3.

Theorem 5.4. *[Stronger version of Theorem 1.3] Consider $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_p)$, where $\boldsymbol{X}$ is our Gaussian design, $k/p \to 0$ and $\log(p/k)/n \to 0$. Then for each $\epsilon > 0$,*

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P}\left(\frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{\sigma^2\cdot 2k\log(p/k)} > 1-\epsilon\right) \to 1.$$

Proof. The proof follows that of Proposition 5.2. As earlier, assume that $\sigma = 1$ without loss of generality. The block prior $\boldsymbol{\pi}$ and the decomposition of the loss L are exactly the same as before except that we work with

$$\tau = \left(\sqrt{2\log(p/k)} - \log\sqrt{2\log(p/k)}\right)\left(1 - 2\sqrt{\log(p/k)/n}\right).$$

Hence, it suffices to prove (5.2) in the current setting, which does not directly follow from Lemma 5.3 because of correlations between the columns of $\boldsymbol{X}$. Thus, write the linear model as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z} = \boldsymbol{X}^{(1)}\boldsymbol{\beta}^{(1)} + \boldsymbol{X}^{-(1)}\boldsymbol{\beta}^{-(1)} + \boldsymbol{z},$$

where $\boldsymbol{X}^{(1)}$ (resp. $\boldsymbol{\beta}^{(1)}$) are the first p/k columns of $\boldsymbol{X}$ (resp. coordinates of $\boldsymbol{\beta}$) and $\boldsymbol{X}^{(-1)}$ all the others. Then

$$\widetilde{\boldsymbol{z}} := \boldsymbol{X}^{-(1)}\boldsymbol{\beta}^{-(1)} + \boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{0}, (\tau^2(k-1)/n + 1)\boldsymbol{I}_n\right),$$

and is independent of $\boldsymbol{X}^{(1)}$ and $\boldsymbol{\beta}^{(1)}$. Since $\tau^2(k-1)/n+1 \geq 1$ and $n/\log(p/k) \to \infty$, we can apply Lemma 5.3 to

$$\boldsymbol{y} = \boldsymbol{X}^{(1)}\boldsymbol{\beta}^{(1)} + \widetilde{\boldsymbol{z}}.$$

This establishes (5.2). □

6 Discussion

Regardless of the design, SLOPE is a concrete and rapidly computable estimator, which also has intuitive statistical appeal. For Gaussian designs, taking Benjamini-Hochberg weights achieves asymptotic minimaxity over large sparsity classes. Furthermore, it is likely that our novel methods would allow us to extend our optimality results to designs with i.i.d. sub-Gaussian entries; for example, designs with independent Bernoulli entries. Since SLOPE runs without any knowledge of the unknown degree of sparsity, we hope that taken together, adaptivity and minimaxity would confirm the appeal of this procedure.

It would of course be of great interest to extend our results to a broader class of designs. In particular, we would like to know what types of results are available when the variables are correlated. In such settings, is there a good way to select the sequence of weights $\{\lambda_i\}$ when the rows of the design are independently sampled from a multivariate Gaussian distribution with zero mean and covariance $\boldsymbol{\Sigma}$, say? How should we tune this sequence for fixed designs? This paper does not address such important questions, and we leave these open for future research.

Finally, returning to the issue of FDR control it would be interesting to establish rigorously whether or not SLOPE controls the FDR in sparse settings.

Acknowledgements

W. S. is partially supported by a General Wang Yaowu Stanford Graduate Fellowship. E. C. is partially supported by NSF under grant CCF-0963835 and by the Math + X Award from the Simons Foundation. W. S. would like to thank Małgorzata Bogdan and Iain Johnstone for helpful discussions. We thank Yuxin Chen, Rina Foygel Barber and Chiara Sabatti for their helpful comments about an early version of the manuscript.

References

[1] F. Abramovich and Y. Benjamini. Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis*, 22:351–361, 1996.

[2] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653, 2006.

[3] M. Abramowitz and I. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Corporation, 2012.

[4] S. R. Adke, V. B. Waikar, and F.J. Schurmann. A two stage shrinkage testimator for the mean of an exponential distribution. *Communications in Statistics-Theory and Methods*, 16(6):1821–1834, 1987.

[5] Y. Baraud. Model selection for regression on a random design. *ESAIM: Probability and Statistics*, 6:127–146, 2002.

[6] R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *arXiv preprint arXiv:1404.5609*, 2014.

[7] M. Bayati and A. Montanari. The Lasso risk for Gaussian matrices. *IEEE Trans. Inform. Theory*, 58(4):1997–2017, 2012.

[8] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[9] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[10] P. J. Bickel. Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *The Annals of Statistics*, 9(6):1301–1309, 1981.

[11] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[12] L. Birgé. Model selection for Gaussian regression with random design. *Bernoulli*, 10(6):1039–1051, 2004.

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

[14] J. M. Bland and D. G. Altman. Multiple significance tests: the Bonferroni method. *the British Medical Journal*, 310(6973):170, 1995.

[15] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès. SLOPE–adaptive variable selection via convex optimization. *arXiv preprint arXiv:1407.3824*, 2014.

[16] M. Bogdan, E. van den Berg, W. Su, and E. J. Candès. Statistical estimation and testing via the sorted ℓ_1 norm. *arXiv preprint arXiv:1310.1969*, 2013.

[17] H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1):115–123, 2008.

[18] S. Boucheron and M. Thomas. Concentration inequalities for order statistics. *Electronic Communications in Probability*, 17(51):1–12, 2012.

[19] L. D. Brown, T. T. Cai, M. G. Low, and C. H. Zhang. Asymptotic equivalence theory for nonparametric regression with random design. *The Annals of statistics*, pages 688–707, 2002.

[20] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

[21] T. T. Cai and H. H. Zhou. A data-driven block thresholding approach to wavelet estimation. *The Annals of Statistics*, pages 569–595, 2009.

[22] E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(6):2145–2177, 2009.

[23] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[24] E. J. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

[25] L. de Haan and A. Ferreira. *Extreme value theory: An introduction*. Springer Science & Business Media, 2007.

[26] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrica*, 81:424–455, 1994.

[27] D. L. Donoho and I. M. Johnstone. Minimax risk over ℓ_p-balls for ℓ_q-error. *Probability Theory and Related Fields*, 99(2):277–303, 1994.

[28] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

[29] D. L. Donoho, I. M. Johnstone, A. Maleki, and A. Montanari. Compressed sensing over ℓ_p-balls: Minimax mean square error. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 129–133. IEEE, 2011.

[30] D. L. Donoho and A. Montanari. High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *arXiv preprint arXiv:1310.7320*, 2013.

[31] D. L. Donoho and J. Tanner. Counting faces of randomly projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22(1):1–53, 2009.

[32] D. L. Donoho and J. Tanner. Exponential bounds implying construction of compressed sensing matrices, error-correcting codes, and neighborly polytopes by random sampling. *IEEE Trans. Inform. Theory*, 56(4):2002–2016, 2010.

[33] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[34] M. Figueiredo and R. Nowak. Sparse estimation with strongly correlated variables using ordered weighted ℓ_1 regularization. *arXiv preprint arXiv:1409.4005*, 2014.

[35] D. P. Foster and E. I. George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975, 1994.

[36] D. P. Foster and R. A. Stine. Local asymptotic coding and the minimum description length. *IEEE Trans. Inform. Theory*, 45(4):1289–1293, 1999.

[37] E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.

[38] M. G'Sell, S. Wager, A. Chouldechova, and R. Tibshirani. Sequential selection procedures and false discovery rate control. *arXiv preprint arXiv:1309.5352*, 2013.

[39] W. Jiang and C.-H. Zhang. Adaptive threshold estimation by FDR. *arXiv preprint arXiv:1312.7840*, 2013.

[40] I. M. Johnstone. Gaussian estimation: Sequence and wavelet models. http://statweb.stanford.edu/ imj/GE06-11-13.pdf, 2013.

[41] W. Liu. Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978, 2013.

[42] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the Lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

[43] A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of majorization and its applications*. Springer Science & Business Media, New York, 2010.

[44] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009.

[45] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):123–231, 2013.

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

[47] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[48] C. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151, 1981.

[49] W. Su and E. J. Candès. Approximating Stein's unbiased risk estimate by drifted Brownian motion. In preparation, 2015.

[50] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, February 1996.

[51] R. Tibshirani and K. Knight. The covariance inflation criterion for adaptive model selection. *J. Roy. Statist. Soc. B*, 55:757–796, 1999.

[52] S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[53] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing, Theory and Applications*, 2012.

[54] N. Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

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

[56] Z. Wu and H. H. Zhou. Model selection and sharp asymptotic minimaxity. *Probability Theory and Related Fields*, 156(1-2):165–191, 2013.

[57] F. Ye and C. H. Zhang. Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *The Journal of Machine Learning Research*, 11:3519–3540, 2010.

[58] C. H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, pages 894–942, 2010.

[59] H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

A Proofs of technical results

As is standard, we write $a_n \asymp b_n$ for two positive sequences a_n and b_n if there exist two constants C_1 and C_2 (possibly depending on q) such that $C_1 a_n \le b_n \le C_2 a_n$ for all n. Also, we write $a_n \sim b_n$ if $a_n/b_n \to 1$.

A.1 Proofs for Section 2

We remind the reader that the proofs in this subsection rely on some lemmas to be stated later in the Appendix.

Proof of (2.1). For simplicity, denote by $\widehat{\boldsymbol{\beta}}$ the (full) Lasso solution $\widehat{\boldsymbol{\beta}}_{\text{Lasso}}$, and $\widehat{\boldsymbol{b}}_S$ the solution to the reduced Lasso problem

$$\underset{\boldsymbol{b}\in\mathbb{R}^k}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}_S\boldsymbol{b}\|^2+\lambda\|\boldsymbol{b}\|_1,$$

where S is the support of the ground truth $\boldsymbol{\beta}$. We show that (i)

$$\left\|\boldsymbol{X}_{\overline{S}}'\boldsymbol{z}\right\|_\infty \le (1+c/2)\sqrt{2\log p} \tag{A.1}$$

and (ii)

$$\left\|\boldsymbol{X}_{\overline{S}}'\boldsymbol{X}_S(\boldsymbol{\beta}_S-\widehat{\boldsymbol{b}}_S)\right\|_\infty < C\sqrt{(k\log^2 p)/n} \tag{A.2}$$

for some constant C, both happen with probability tending to one. Now observe that $\boldsymbol{X}_{\overline{S}}'(\boldsymbol{y}-\boldsymbol{X}_S\widehat{\boldsymbol{b}}_S)=\boldsymbol{X}_{\overline{S}}'\boldsymbol{z}+\boldsymbol{X}_{\overline{S}}'\boldsymbol{X}_S(\boldsymbol{\beta}_S-\widehat{\boldsymbol{b}}_S)$. Hence, combining (A.1) and (A.2) and using the fact that $(k\log p)/n\to 0$ give

$$\begin{aligned}\left\|\boldsymbol{X}_{\overline{S}}'(\boldsymbol{y}-\boldsymbol{X}_S\widehat{\boldsymbol{b}}_S)\right\|_\infty &\le \left\|\boldsymbol{X}_{\overline{S}}'\boldsymbol{X}_S(\boldsymbol{\beta}_S-\widehat{\boldsymbol{b}}_S)\right\|_\infty + \left\|\boldsymbol{X}_{\overline{S}}'\boldsymbol{z}\right\|_\infty\\ &\le C\sqrt{(k\log^2 p)/n}+(1+c/2)\sqrt{2\log p}\\ &= o(\sqrt{2\log p})+(1+c/2)\sqrt{2\log p}\\ &< (1+c)\sqrt{2\log p}\end{aligned}$$

with probability approaching one. This last inequality together with the fact that $\widehat{\boldsymbol{b}}_S$ obeys the KKT conditions for the reduced Lasso problem imply that padding $\widehat{\boldsymbol{b}}_S$ with zeros on $\overline{S}$ obeys the KKT conditions for the full Lasso problem and is, therefore, solution.

We need to justify (A.1) and (A.2). First, Lemmas A.6 and A.5 imply (A.1). Next, to show (A.2), we rewrite the left-hand side in (A.2) as

$$\boldsymbol{X}_{\overline{S}}'\boldsymbol{X}_S(\boldsymbol{\beta}_S-\widehat{\boldsymbol{b}}_S)=\boldsymbol{X}_{\overline{S}}'\boldsymbol{X}_S(\boldsymbol{X}_S'\boldsymbol{X}_S)^{-1}(\boldsymbol{X}_S'(\boldsymbol{y}-\boldsymbol{X}_S\widehat{\boldsymbol{b}}_S)-\boldsymbol{X}_S'\boldsymbol{z}).$$

By Lemma A.7, we have that

$$\left\|\boldsymbol{X}_S'(\boldsymbol{y}-\boldsymbol{X}_S\widehat{\boldsymbol{b}}_S)-\boldsymbol{X}_S'\boldsymbol{z}\right\| \le \sqrt{k}\lambda+\left\|\boldsymbol{X}_S'\boldsymbol{z}\right\| \le \sqrt{k}\lambda+\sqrt{32k\log(p/k)} \le C'\sqrt{k\log p}$$

holds with probability at least $1-\mathrm{e}^{-n/2}-(\sqrt{2}\mathrm{e}k/p)^k\to 1$. In addition, Lemma A.11 with $t=1/2$ gives

$$\left\|\boldsymbol{X}_S(\boldsymbol{X}_S'\boldsymbol{X}_S)^{-1}\right\| \le \frac{1}{\sqrt{1-1/n}-\sqrt{k^\star/n}-1/2} < 3$$

with probability at least $1-\mathrm{e}^{-n/8}\to 1$. Hence, from the last two inequalities it follows that

$$\left\|\boldsymbol{X}_S(\boldsymbol{X}_S'\boldsymbol{X}_S)^{-1}(\boldsymbol{X}_S'(\boldsymbol{y}-\boldsymbol{X}_S\widehat{\boldsymbol{b}}_S)-\boldsymbol{X}_S'\boldsymbol{z})\right\| \le C''\sqrt{k\log p} \tag{A.3}$$

with probability at least $1-\mathrm{e}^{-n/2}-(\sqrt{2}\mathrm{e}k/p)^k-\mathrm{e}^{-n/8}\to 1$. Since $\boldsymbol{X}_{\overline{S}}'$ is independent of $\boldsymbol{X}_S(\boldsymbol{X}_S'\boldsymbol{X}_S)^{-1}(\boldsymbol{X}_S'(\boldsymbol{y}-\boldsymbol{X}_S\widehat{\boldsymbol{b}}_S)-\boldsymbol{X}_S'\boldsymbol{z})$, Lemma A.6 gives

$$\begin{aligned}&\left\|\boldsymbol{X}_{\overline{S}}'\boldsymbol{X}_S(\boldsymbol{X}_S'\boldsymbol{X}_S)^{-1}(\boldsymbol{X}_S'(\boldsymbol{y}-\boldsymbol{X}_S\widehat{\boldsymbol{b}}_S)-\boldsymbol{X}_S'\boldsymbol{z})\right\|_\infty\\ &\qquad\le \sqrt{\frac{2\log p}{n}}\left\|\boldsymbol{X}_S(\boldsymbol{X}_S'\boldsymbol{X}_S)^{-1}(\boldsymbol{X}_S'(\boldsymbol{y}-\boldsymbol{X}_S\widehat{\boldsymbol{b}}_S)-\boldsymbol{X}_S'\boldsymbol{z})\right\|\end{aligned}$$

with probability approaching one. Combining this with (A.3) gives (A.2).

Let $\widetilde{\boldsymbol{b}}_S$ be the solution to

$$\underset{\boldsymbol{b}\in\mathbb{R}^k}{\text{minimize}}\ \frac{1}{2}\left\|\boldsymbol{\beta}_S+\boldsymbol{X}_S'\boldsymbol{z}-\boldsymbol{b}\right\|^2+\lambda\|\boldsymbol{b}\|_1.$$

To complete the proof of (2.1), it suffices to establish (i) that for any constant $\delta > 0$,

$$\sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P}\left(\frac{\|\widetilde{\boldsymbol{b}}_S - \boldsymbol{\beta}_S\|^2}{2(1+c)^2 k \log p} > 1 - \delta \right) \to 1, \tag{A.4}$$

and (ii)

$$\|\widetilde{\boldsymbol{b}}_S - \widehat{\boldsymbol{b}}_S\| = o_{\mathbb{P}}\left(\|\widetilde{\boldsymbol{b}}_S - \boldsymbol{\beta}_S\| \right) \tag{A.5}$$

since (A.4) and (A.5) give

$$\sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P}\left(\frac{\|\widehat{\boldsymbol{b}}_S - \boldsymbol{\beta}_S\|^2}{2(1+c)^2 k \log p} > 1 - \delta \right) \to 1 \tag{A.6}$$

for each $\delta > 0$. Note that taking $\delta = 1 - 1/(1+c)^2$ in (A.6) and using the fact that $\widehat{\boldsymbol{b}}_S$ is solution to Lasso with probability approaching one finish the proof

Proof of (A.4). Let $\beta_i = \infty$ if $i \in S$ and otherwise zero (treat ∞ as a sufficiently large positive constant). For each $i \in S$, $\widetilde{b}_{S,i} = \beta_i + \boldsymbol{X}_i'\boldsymbol{z} - \lambda$, and

$$|\widetilde{b}_{S,i} - \beta_i| = |\boldsymbol{X}_i'\boldsymbol{z} - \lambda| \ge \lambda - |\boldsymbol{X}_i'\boldsymbol{z}|.$$

On the event $\{\max_{i \in S} |\boldsymbol{X}_i'\boldsymbol{z}| \le \lambda\}$, which happens with probability tending to one, this inequality gives

$$\begin{aligned} \|\widetilde{\boldsymbol{b}}_S - \boldsymbol{\beta}_S\|^2 \ge \sum_{i \in S} (\lambda - |\boldsymbol{X}_i'\boldsymbol{z}|)^2 &= k\lambda^2 - 2\lambda \sum_{i \in S} |\boldsymbol{X}_i'\boldsymbol{z}| + \sum_{i \in S} (\boldsymbol{X}_i'\boldsymbol{z})^2 \\ &= (1 + o_{\mathbb{P}}(1)) 2(1+c)^2 k \log p, \end{aligned}$$

where we have used that both $\sum_{i \in S} (\boldsymbol{X}_i'\boldsymbol{z})^2$ and $\sum_{i \in S} |\boldsymbol{X}_i'\boldsymbol{z}|$ are $O_{\mathbb{P}}(k)$. This proves the claim.

Proof of (A.5). Apply Lemma 4.2 with T replaced by S (here each of $\widehat{\boldsymbol{b}}_S, \widetilde{\boldsymbol{b}}_S$ and $\boldsymbol{\beta}$ is supported on S). Since $k/p \to 0$, for any constant $\delta' > 0$, all the singular values of $\boldsymbol{X}_S$ lie in $(1 - \delta', 1 + \delta')$ with overwhelming probability (see, for example, [53]). Consequently, Lemma 4.2 ensures (A.5). □

Proof of (2.2). We assume $\sigma = 1$ and put $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\mathrm{BH}}$. As in the proof of Theorem 1.1, we decompose the total loss as

$$|\widehat{\boldsymbol{\beta}}_{\mathrm{Seq}} - \boldsymbol{\beta}\|^2 = \|\widehat{\boldsymbol{\beta}}_{\mathrm{Seq},S} - \boldsymbol{\beta}_S\|^2 + \|\widehat{\boldsymbol{\beta}}_{\mathrm{Seq},\overline{S}} - \boldsymbol{\beta}_{\overline{S}}\|^2 = \|\widehat{\boldsymbol{\beta}}_{\mathrm{Seq},S} - \boldsymbol{\beta}_S\|^2 + \|\widehat{\boldsymbol{\beta}}_{\mathrm{Seq},\overline{S}}\|^2.$$

The largest possible value of the loss off support is achieved when $\boldsymbol{y}_{\overline{S}}$ is sequentially soft-thresholded by $\boldsymbol{\lambda}^{-[k]}$. Hence, by the proof of Lemma 3.3, we obtain

$$\mathbb{E}\,\|\widehat{\boldsymbol{\beta}}_{\mathrm{Seq},\overline{S}}\|^2 = o\left(2k \log(p/k)\right)$$

for all k-sparse $\boldsymbol{\beta}$.

Now, we turn to consider the loss on support. For any $i \in S$, the loss is at most

$$\left(|z_i| + \lambda_{r(i)}\right)^2 = \lambda_{r(i)}^2 + z_i^2 + 2|z_i|\lambda_{r(i)}.$$

Summing the above equalities over all $i \in S$ gives

$$\mathbb{E}\,\|\widehat{\boldsymbol{\beta}}_{\mathrm{Seq},S} - \boldsymbol{\beta}_S\|^2 \le \sum_{i=1}^{k} \lambda_i^2 + \sum_{i\in S} z_i^2 + 2\sum_{i\in S} |z_i|\lambda_{r(i)}.$$

Note that the first term $\sum_{i=1}^{k}\lambda_i^2 = (1+o(1))\,2k\log(p/k)$, and the second term has expectation $\mathbb{E}\sum_{i\in S} z_i^2 = k = o(2k\log(p/k))$, so that it suffices to show that

$$\mathbb{E}\left[2\sum_{i\in S}|z_i|\lambda_{r(i)}\right] = o\left(2k\log(p/k)\right). \tag{A.7}$$

We emphasize that both z_i and $r(i)$ are random so that $\{\lambda_{r(i)}\}_{i\in S}$ and $\{z_i\}_{i\in S}$ may not be independent. Without loss of generality, assume $S = \{1,\dots,k\}$ and for $1 \le i \le k$, let $r'(i)$ be the rank of the ith observation among the first k. Since $\boldsymbol{\lambda}$ is nonincreasing and $r'(i) \le r(i)$, we have

$$\sum_{1\le i\le k} |z_i|\lambda_{r(i)} \le \sum_{1\le i\le k} |z_i|\lambda_{r'(i)} \le \sum_{1\le i\le k} |z|_{(i)}\lambda_i,$$

where $|z|_{(1)} \ge \cdots \ge |z|_{(k)}$ are the order statistics of $z_1,\dots,z_k$. The second inequality follows from the fact that for any nonnegative sequences $\{a_i\}$ and $\{b_i\}$, $\sum_i a_i b_i \le \sum_i a_{(i)}b_{(i)}$. Therefore, letting $\zeta_1,\dots,\zeta_k$ be i.i.d. $\mathcal{N}(0,1)$, (A.7) follows from the estimate

$$\sum_{i=1}^{k} \lambda_i\,\mathbb{E}\,|\zeta|_{(i)} = o\left(2k\log(p/k)\right). \tag{A.8}$$

To argue about (A.8), we work with the approximations $\lambda_i \sim \sqrt{2\log(p/i)}$ and $\mathbb{E}\,|\zeta|_{(i)} = O\left(\sqrt{2\log(2k/i)}\right)$ (see e.g. (A.15)), so that the claim is a consequence of

$$\sum_{i=1}^{k}\sqrt{\log\frac{p}{i}\log\frac{2k}{i}} = o\left(2k\log(p/k)\right),$$

which is justified as follows:

$$\begin{aligned}
\sum_{i=1}^{k}\sqrt{\log\frac{p}{i}\log\frac{2k}{i}} &\le k\int_0^1 \sqrt{\log\frac{p/k}{x}\log\frac{2}{x}}\mathrm{d}x \\
&\le k\int_0^1 \sqrt{\log\frac{p}{k}}\sqrt{\log\frac{2}{x}} + \frac{\log\frac{1}{x}\sqrt{\log\frac{2}{x}}}{2\sqrt{\log(p/k)}}\mathrm{d}x \\
&= C_1 k\sqrt{\log\frac{p}{k}} + \frac{C_2 k}{\sqrt{\log(p/k)}}
\end{aligned}$$

for some absolute constants C_1, C_2. Since $\log(p/k)\to\infty$, it is clear that the right-hand side of the above display is of $o(2k\log(p/k))$. □

A.2 Proofs for Section 3

To begin with, we derive a dual formulation of the SLOPE program (1.6), which provides a nice geometrical interpretation. This dual formulation will also be used in the proof of Lemma 4.3. Our exposition largely borrows from [15].

Rewrite (1.6) as

$$\underset{\boldsymbol{b},\boldsymbol{r}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{r}\|^2 + \sum_i \lambda_i |b|_{(i)} \quad \text{subject to} \quad \boldsymbol{X}\boldsymbol{b} + \boldsymbol{r} = \boldsymbol{y}, \tag{A.9}$$

whose Lagrangian is

$$\mathcal{L}(\boldsymbol{b}, \boldsymbol{r}, \boldsymbol{\nu}) := \frac{1}{2}\|\boldsymbol{r}\|^2 + \sum_i \lambda_i |b|_{(i)} - \boldsymbol{\nu}'(\boldsymbol{X}\boldsymbol{b} + \boldsymbol{r} - \boldsymbol{y}).$$

Hence, the dual objective is given by

$$\begin{aligned}\inf_{\boldsymbol{b},\boldsymbol{r}} \ \mathcal{L}(\boldsymbol{b}, \boldsymbol{r}, \boldsymbol{\nu}) &= \boldsymbol{\nu}'\boldsymbol{y} - \sup_{\boldsymbol{r}} \left\{ \boldsymbol{\nu}'\boldsymbol{r} - \frac{1}{2}\|\boldsymbol{r}\|^2 \right\} - \sup_{\boldsymbol{b}} \left\{ (\boldsymbol{X}'\boldsymbol{\nu})'\boldsymbol{b} - \sum_i \lambda_i |b|_{(i)} \right\} \\ &= \boldsymbol{\nu}'\boldsymbol{y} - \frac{1}{2}\|\boldsymbol{\nu}\|^2 - \begin{cases} 0 & \boldsymbol{\nu} \in C_{\boldsymbol{\lambda},\boldsymbol{X}} \\ +\infty & \text{otherwise,} \end{cases}\end{aligned}$$

where $C_{\boldsymbol{\lambda},\boldsymbol{X}} := \{\boldsymbol{\nu} : \boldsymbol{X}'\boldsymbol{\nu} \text{ is majorized by } \boldsymbol{\lambda}\}$ is a (convex) polytope. It thus follows that the dual reads

$$\underset{\boldsymbol{\nu}}{\text{maximize}} \quad \boldsymbol{\nu}'\boldsymbol{y} - \frac{1}{2}\|\boldsymbol{\nu}\|^2 \quad \text{subject to } \boldsymbol{\nu} \in C_{\boldsymbol{\lambda},\boldsymbol{X}}. \tag{A.10}$$

The equality $\boldsymbol{\nu}'\boldsymbol{y} - \|\boldsymbol{\nu}\|^2/2 = -\|\boldsymbol{y} - \boldsymbol{\nu}\|^2/2 + \|\boldsymbol{y}\|^2/2$ reveals that the dual solution $\widehat{\boldsymbol{\nu}}$ is indeed the projection of $\boldsymbol{y}$ onto $C_{\boldsymbol{\lambda},\boldsymbol{X}}$. The minimization of the Lagrangian over $\boldsymbol{r}$ is attained at $\boldsymbol{r} = \boldsymbol{\nu}$. This implies that the primal solution $\widehat{\boldsymbol{\beta}}$ and the dual solution $\widehat{\boldsymbol{\nu}}$ obey

$$\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\nu}}. \tag{A.11}$$

We turn to proving the facts.

Proof of Fact 3.1. Without loss of generality, suppose both $\boldsymbol{a}$ and $\boldsymbol{b}$ are nonnegative and arranged in nonincreasing order. Denote by $T_k^{\boldsymbol{a}}$ the sum of the first k terms of $\boldsymbol{a}$ with $T_0^{\boldsymbol{a}} \triangleq 0$, and similarly for $\boldsymbol{b}$. We have

$$\|\boldsymbol{a}\|^2 = \sum_{k=1}^{p} a_k (T_k^{\boldsymbol{a}} - T_{k-1}^{\boldsymbol{a}}) = \sum_{k=1}^{p-1} T_k^{\boldsymbol{a}}(a_k - a_{k+1}) + a_p T_p^{\boldsymbol{a}} \ge \sum_{k=1}^{p-1} T_k^{\boldsymbol{b}}(a_k - a_{k+1}) + a_p T_p^{\boldsymbol{b}} = \sum_{k=1}^{p} a_k b_k.$$

Similarly,

$$\|\boldsymbol{b}\|^2 = \sum_{k=1}^{p-1} T_k^{\boldsymbol{b}}(b_k - b_{k+1}) + b_p T_p^{\boldsymbol{b}} \le \sum_{k=1}^{p-1} T_k^{\boldsymbol{a}}(b_k - b_{k+1}) + b_p T_p^{\boldsymbol{a}} = \sum_{k=1}^{p} b_k (T_k^{\boldsymbol{a}} - T_{k-1}^{\boldsymbol{a}}) = \sum_{k=1}^{p} a_k b_k,$$

which proves the claim. □

Proofs of Facts 3.2 and 3.3. Taking $\boldsymbol{X} = \boldsymbol{I}_p$ in the dual formulation, (A.11) immediately implies that $\boldsymbol{a} - \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})$ is the projection of $\boldsymbol{a}$ onto the polytope $C_{\boldsymbol{\lambda},\boldsymbol{I}_p}$. By definition, $C_{\boldsymbol{\lambda},\boldsymbol{I}_p}$ consists of all vectors majorized by $\boldsymbol{\lambda}$. Hence, $\boldsymbol{a} - \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})$ is always majorized by $\boldsymbol{\lambda}$. In particular, if $\boldsymbol{a}$ is majorized by $\boldsymbol{\lambda}$, then the projection $\boldsymbol{a} - \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})$ of $\boldsymbol{a}$ is identical to $\boldsymbol{a}$ itself. This gives $\operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a}) = \boldsymbol{0}$. □

Proof of Fact 3.4. Assume $\boldsymbol{a}$ is nonnegative without loss of generality. It is intuitively obvious that

$$\boldsymbol{b} \geq \boldsymbol{a} \quad \Longrightarrow \quad \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{b}) \geq \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a}),$$

where as usual $\boldsymbol{b} \geq \boldsymbol{a}$ means that $\boldsymbol{b} - \boldsymbol{a} \in \mathbb{R}^p_+$. In other words, if the observations increase, the fitted values do not decrease. To save time, we directly verify this claim by using Algorithm 3 (FastProxSL1) from [15]. By the averaging step of that algorithm, we can see that for each $1 \leq i, j \leq p$,

$$\frac{\partial \left[\operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})\right]_i}{\partial a_j} = \begin{cases} \frac{1}{\#\{1 \leq k \leq p:\ [\operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})]_k = [\operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})]_j\}}, & \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})_j = \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})_i > 0, \\ 0, & \text{otherwise.} \end{cases}$$

This holds for all $\boldsymbol{a} \in \mathbb{R}^p$ except for a set of measure zero. The nonnegativity of $\partial \left[\operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})\right]_i / \partial a_j$ along with the Lipschitz continuity of the prox imply the monotonicity property. A consequence is that $\| \left[\operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})\right]_{\overline{T}} \|$ does not decrease as we let $a_i \to \infty$ for all $i \in T$. In the limit, $\| \left[\operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})\right]_{\overline{T}} \|$ monotonically converges to $\| \operatorname{prox}_{\boldsymbol{\lambda}_{-|T|}}(\boldsymbol{a}_{\overline{T}}) \|$. This gives the desired inequality.

□

As a remark, we point out that the proofs of Facts 3.2 and 3.3 suggest a very simple proof of Lemma 3.1. Since $\boldsymbol{a} - \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})$ is the projection of $\boldsymbol{a}$ onto $C_{\boldsymbol{\lambda},\boldsymbol{I}_p}$, $\| \operatorname{prox}_{\boldsymbol{\lambda}}(\boldsymbol{a}) \|$ is thus the distance between $\boldsymbol{a}$ and the polytope $C_{\boldsymbol{\lambda},\boldsymbol{I}_p}$. Hence, it suffices to find a point in the polytope at a distance of $\|(|\boldsymbol{a}| - \boldsymbol{\lambda})_+\|$ away from $\boldsymbol{a}$. The point $\boldsymbol{b}$ defined as $b_i = \min\{|a_i|, \lambda_i\}$ does the job.

Now, we proceed to prove the preparatory lemmas for Theorem 1.1, namely, Lemmas A.3 and A.4. The first two lemmas below can be found in [3].

Lemma A.1. *Let U be a* $\mathrm{Beta}(a, b)$ *random variable. Then*

$$\mathbb{E} \log U = (\log \Gamma(a))' - (\log \Gamma(a+b))',$$

where Γ denotes the Gamma function and $(\log \Gamma(x))'$ is the derivative with respect to x.

Lemma A.2. *For any integer $m \geq 1$,*

$$(\log \Gamma(m))' = -\gamma + \sum_{j=1}^{m-1} \frac{1}{j} = \log m + O\Big(\frac{1}{m}\Big),$$

where $\gamma = 0.577215\cdots$ is the Euler constant.

Lemma A.3. *Let $\boldsymbol{\zeta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{p-k})$. Under the assumptions of Theorem 1.1, for any constant $A > 0$,*

$$\frac{1}{2k \log(p/k)} \sum_{i=1}^{\lfloor Ak \rfloor} \mathbb{E}\left(|\zeta|_{(i)} - \lambda^{\mathrm{BH}}_{k+i}\right)^2_+ \to 0.$$

Proof of Lemma A.3. Write $\lambda_i = \lambda_i^{\mathrm{BH}}$ for simplicity. It is sufficient to prove a stronger version in which the order statistics $|\zeta|_{(i)}$ come from p i.i.d. $\mathcal{N}(0,1)$. The reason is that the order statistics will be stochastically larger, thus enlarging $\mathbb{E}\left(|\zeta|_{(i)} - \lambda_{k+i}^{\mathrm{BH}}\right)_+^2$, since $(\zeta-\lambda)_+^2$ is nondecreasing in ζ. Applying the bias-variance decomposition, we get

$$\mathbb{E}\left(|\zeta|_{(i)} - \lambda_{k+i}\right)_+^2 \le \mathbb{E}\left(|\zeta|_{(i)} - \lambda_{k+i}\right)^2 = \mathrm{Var}(|\zeta_{(i)}|) + \left(\mathbb{E}\,|\zeta_{(i)}| - \lambda_{k+i}\right)^2. \tag{A.12}$$

We proceed to control each term separately.

For the variance, a direct application of Proposition 4.2 in [18] gives

$$\mathrm{Var}(|\zeta_{(i)}|) = O\Big(\frac{1}{i\log(p/i)}\Big) \tag{A.13}$$

for all $i \le p/2$. Hence,

$$\sum_{i=1}^{\lfloor Ak\rfloor} \mathrm{Var}(|\zeta_{(i)}|) = O\left(\sum_{i=1}^{\lfloor Ak\rfloor} \frac{1}{i\log(p/i)}\right) = o(2k\log(p/k)),$$

where the last step makes use of $\log(p/k)\to\infty$. It remains to show that

$$\sum_{i=1}^{\lfloor Ak\rfloor} \left(\mathbb{E}\,|\zeta_{(i)}| - \lambda_{k+i}\right)^2 = o(2k\log(p/k)). \tag{A.14}$$

Let $U_1,\dots,U_p$ be i.i.d. uniform random variables on $(0,1)$ and $U_{(i)}$ be the i^{th} smallest—please note that for a change, the U_i's are sorted in increasing order. We know that $U_{(i)}$ is distributed as $\mathrm{Beta}(i, p+1-i)$ and that $|\zeta|_{(i)}$ has the same distribution as $\Phi^{-1}(1-U_{(i)}/2)$. Making use of Lemmas A.1 and A.2 then gives

$$\mathbb{E}\,|\zeta|_{(i)}^2 = \mathbb{E}\left[\Phi^{-1}(1-U_{(i)}/2)^2\right] \sim \mathbb{E}\left[2\log(2/U_{(i)})\right] = 2\log 2 + 2\sum_{j=i}^{p}\frac{1}{j} = (1+o(1))2\log(p/i),$$

where the second step follows from $(1+o_{\mathbb{P}}(1))2\log(2/U_{(i)}) \le \Phi^{-1}(1-U_{(i)}/2)^2 \le 2\log(2/U_{(i)})$ for $i = o(p)$. As a result,

$$\begin{aligned}
&\mathbb{E}\,|\zeta_{(i)}| \le \sqrt{\mathbb{E}\,|\zeta|_{(i)}^2} = (1+o(1))\sqrt{2\log(p/i)} \\
&\mathbb{E}\,|\zeta_{(i)}| = \sqrt{\mathbb{E}\,|\zeta|_{(i)}^2 - \mathrm{Var}(|\zeta|_{(i)})} = (1+o(1))\sqrt{2\log(p/i)}.
\end{aligned} \tag{A.15}$$

Similarly, since $k+i = o(p)$ and q is constant, we have the approximation

$$\lambda_{k+i} = (1+o(1))\sqrt{2\log(p/(k+i))},$$

which together with (A.15) reveals that

$$\left(\mathbb{E}\,|\zeta_{(i)}| - \lambda_{k+i}\right)^2 \le (1+o(1))\,2\left[\sqrt{\log(p/i)} - \sqrt{\log(p/(k+i))}\right]^2 + o(1)\log(p/i). \tag{A.16}$$

The second term in the right-hand side contributes at most $o(1)\,Ak\log(p/(Ak)) = o(1)\,2k\log(p/k)$ in the sum (A.14). For the first term, we get

$$\left[\sqrt{\log(p/i)} - \sqrt{\log(p/(k+i))}\right]^2 = \frac{\log^2(1+k/i)}{\left[\sqrt{\log(p/i)} + \sqrt{\log(p/(k+i))}\right]^2} = o(1)\log^2(1+k/i).$$

Hence, it contributes at most

$$\begin{aligned} o(1)\sum_{i=1}^{\lfloor Ak\rfloor}\log^2(1+k/i) &\le o(1)\sum_{i=1}^{\lfloor Ak\rfloor} k\int_{\frac{i-1}{k}}^{\frac{i}{k}}\log^2(1+1/x)\mathrm{d}x \\ &= o(1)k\int_0^A \log^2(1+1/x)\mathrm{d}x = o(2k\log(p/k)). \end{aligned} \tag{A.17}$$

Combining (A.17), (A.16) and (A.14) concludes the proof. □

Lemma A.4. *Let $\boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{p-k})$ and $A > 0$ be any constant satisfying $q(1+A)/A < 1$. Then, under the assumptions of Theorem 1.1,*

$$\frac{1}{2k\log(p/k)}\sum_{i=\lceil Ak\rceil}^{p-k}\mathbb{E}\left(|\zeta|_{(i)} - \lambda_{k+i}^{\mathrm{BH}}\right)_+^2 \to 0.$$

Proof of Lemma A.4. Again, write $\lambda_i = \lambda_i^{\mathrm{BH}}$ for simplicity. As in the proof of Lemma A.3 we work on a stronger version by assuming $\boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$. Denote by $q' = q(1+A)/A$. For any $u \ge 0$, let $\alpha_u := \mathbb{P}(|\mathcal{N}(0,1)| > \lambda_{k+i} + u) = 2\Phi(-\lambda_{k+i} - u)$. Then $\mathbb{P}(|\zeta|_{(i)} > \lambda_{k+i} + u)$ is just the tail probability of the binomial distribution with p trials and success probability α_u. By the Chernoff bound, this probability is bounded as

$$\mathbb{P}(|\zeta|_{(i)} > \lambda_{k+i} + u) \le \mathrm{e}^{-p\,\mathrm{KL}(i/p\|\alpha_u)}, \tag{A.18}$$

where $\mathrm{KL}(a\|b) := a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}$ is the Kullback-Leibler divergence. Note that

$$\frac{\partial\,\mathrm{KL}(i/p\|b)}{\partial b} = -\frac{i/p}{b} + \frac{1-i/p}{1-b} \le -\frac{i}{pb} + 1 \tag{A.19}$$

for all $0 < b < i/p$. Hence, from (A.19) it follows that

$$\begin{aligned} \mathrm{KL}\left(i/p\|\alpha_u\right) - \mathrm{KL}\left(i/p\|\alpha_0\right) = -\int_{\alpha_u}^{\alpha_0}\frac{\partial\,\mathrm{KL}}{\partial b}\mathrm{d}b &\ge \int_{\alpha_u}^{\alpha_0}\frac{i}{pb} - 1\mathrm{d}b \\ &\ge \int_{\mathrm{e}^{-u\lambda_{k+i}}\alpha_0}^{\alpha_0}\frac{i}{pb} - 1\mathrm{d}b \\ &= \frac{iu\lambda_{k+i}}{p} - \alpha_0\left(1 - \mathrm{e}^{-u\lambda_{k+i}}\right), \end{aligned} \tag{A.20}$$

where the second inequality makes use of $\alpha_u \le \mathrm{e}^{-u\lambda_{k+i}}\alpha_0$. With the proviso that $q(1+A)/A < 1$ and $i \ge Ak$, it follows that

$$\alpha_0 = q(k+i)/p \le q'i/p. \tag{A.21}$$

Hence, substituting (A.21) into (A.20), we see that (A.18) yields

$$\begin{aligned}\mathbb{P}(|\zeta|_{(i)} > \lambda_{k+i} + u) &\le \mathrm{e}^{-p\left(\mathrm{KL}(\frac{i}{p}\|\alpha_u)-\mathrm{KL}(\frac{i}{p}\|\alpha_0)\right)}\mathrm{e}^{-p\,\mathrm{KL}(\frac{i}{p}\|\alpha_0)} \\ &\le \mathrm{e}^{-p\left(\mathrm{KL}(\frac{i}{p}\|\alpha_u)-\mathrm{KL}(\frac{i}{p}\|\alpha_0)\right)} \\ &\le \exp\left(-iu\lambda_{k+i} + q'i\big(1-\exp\big(-u\lambda_{k+i}\big)\big)\right).\end{aligned} \tag{A.22}$$

With this preparation, we conlude the proof of our lemma as follows:

$$\begin{aligned}\mathbb{E}\left(|\zeta|_{(i)} - \lambda_{k+i}\right)_+^2 &= \int_0^\infty \mathbb{P}\left((|\zeta|_{(i)} - \lambda_{k+i})_+^2 > x\right)\mathrm{d}x \\ &= \int_0^\infty \mathbb{P}(|\zeta|_{(i)} > \lambda_{k+i} + \sqrt{x})\mathrm{d}x \\ &= 2\int_0^\infty u\mathbb{P}(|\zeta|_{(i)} > \lambda_{k+i} + u)\mathrm{d}u,\end{aligned}$$

and plugging (A.22) gives

$$\begin{aligned}\mathbb{E}\left(|\zeta|_{(i)} - \lambda_{k+i}\right)_+^2 &\le 2\int_0^\infty u\exp\Big(-iu\lambda_{k+i} + q'i\big(1-\exp\big(-u\lambda_{k+i}\big)\big)\Big)\mathrm{d}u \\ &= \frac{2}{\lambda_{k+i}^2}\int_0^\infty x\mathrm{e}^{-(x-q'(1-\mathrm{e}^{-x}))i}\mathrm{d}x \\ &\le \frac{2}{\lambda_p^2}\int_0^\infty x\mathrm{e}^{-(x-q'(1-\mathrm{e}^{-x}))i}\mathrm{d}x.\end{aligned}$$

This yields the upper bound

$$\begin{aligned}\sum_{i=\lceil Ak\rceil}^{p-k}\mathbb{E}\left(|\zeta|_{(i)} - \lambda_{k+i}\right)_+^2 &\le \frac{2}{\lambda_p^2}\sum_{i=\lceil Ak\rceil}^{p-k}\int_0^\infty x\mathrm{e}^{-(x-q'(1-\mathrm{e}^{-x}))i}\mathrm{d}x \\ &\le \frac{2}{\Phi^{-1}(1-q/2)^2}\sum_{i=1}^{\infty}\int_0^\infty x\mathrm{e}^{-(x-q'(1-\mathrm{e}^{-x}))i}\mathrm{d}x \\ &= \frac{2}{\Phi^{-1}(1-q/2)^2}\int_0^\infty \frac{x\mathrm{e}^{-(x-q'(1-\mathrm{e}^{-x}))}}{1-\mathrm{e}^{-(x-q'(1-\mathrm{e}^{-x}))}}\mathrm{d}x.\end{aligned}$$

Since the integrand obeys

$$\lim_{x\to 0}\frac{x\mathrm{e}^{-(x-q'(1-\mathrm{e}^{-x}))}}{1-\mathrm{e}^{-(x-q'(1-\mathrm{e}^{-x}))}} = \frac{1}{1-q'}$$

and decays exponentially fast as $x \to \infty$, we conclude that $\sum_{i=\lceil Ak\rceil}^{p-k}\mathbb{E}(|\zeta|_{(i)} - \lambda_{k+i})_+^2$ is bounded by a constant. This is a bit more than we need since $2k\log(p/k) \to \infty$. □

A.3 Proofs for Section 4

In this paper, we often use the Borell inequality to show that $\mathbb{P}(\|\mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)\| > \sqrt{n}+t) \le \exp(-t^2/2)$.

Lemma A.5 (Borell's inequality)**.** *Let $\boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ and f be an L-Lipschitz continuous function in $\mathbb{R}^n$. Then*

$$\mathbb{P}\left(f(\boldsymbol{\zeta}) > \mathbb{E}\, f(\boldsymbol{\zeta}) + t\right) \le \mathrm{e}^{-\frac{t^2}{2L^2}}$$

for every $t > 0$.

Lemma A.6. *Let $\zeta_1, \dots, \zeta_p$ be i.i.d. $\mathcal{N}(0,1)$. Then*

$$\max_i |\zeta_i| \le \sqrt{2 \log p}$$

holds with probability approaching one.

The latter classical result can be proved in many different ways. Suffices to say that it follows from a more subtle fact, namely, that

$$\sqrt{2\log p}\left(\max_i \zeta_i - \sqrt{2\log p} + \frac{\log\log p + \log 4\pi}{2\sqrt{2\log p}}\right)$$

converges weakly to a Gumbel distribution [25].

Proof of Lemma 4.3. Let $\widehat{\boldsymbol{\beta}}^{\mathrm{lift}}$ be the lift of $\widehat{\boldsymbol{b}}_T$ in the sense that $\widehat{\boldsymbol{\beta}}^{\mathrm{lift}}_T = \widehat{\boldsymbol{b}}_T$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{lift}}_{\overline{T}} = \mathbf{0}$ and let $|T| = m$. Further, set $\widetilde{\boldsymbol{\nu}} := \boldsymbol{y} - \boldsymbol{X}_T \widehat{\boldsymbol{b}}_T = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\mathrm{lift}}$. Applying (A.10) and (A.11) to the reduced SLOPE program, we get that

$$\boldsymbol{X}_T' \widetilde{\boldsymbol{\nu}} \preceq \boldsymbol{\lambda}^{[m]}.$$

By the assumption, $\boldsymbol{X}_{\overline{T}}' \widetilde{\boldsymbol{\nu}}$ is majorized by $\boldsymbol{\lambda}^{-[m]}$. Hence, $\boldsymbol{X}' \widetilde{\boldsymbol{\nu}}$—the concatenation of $\boldsymbol{X}_T' \widetilde{\boldsymbol{\nu}}$ and $\boldsymbol{X}_{\overline{T}}' \widetilde{\boldsymbol{\nu}}$—is majorized by $\boldsymbol{\lambda} = (\boldsymbol{\lambda}^{[m]}, \boldsymbol{\lambda}^{-[m]})$. This confirms that $\widetilde{\boldsymbol{\nu}}$ is dual feasible with respect to the full SLOPE program. If additionally we show that

$$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\mathrm{lift}}\|^2 + \sum_i \lambda_i |\widehat{\beta}^{\mathrm{lift}}|_{(i)} = \widetilde{\boldsymbol{\nu}}' \boldsymbol{y} - \frac{1}{2}\|\widetilde{\boldsymbol{\nu}}\|^2, \tag{A.23}$$

then the strong duality claims that $\widehat{\boldsymbol{\beta}}^{\mathrm{lift}}$ and $\widetilde{\boldsymbol{\nu}}$ must, respectively, be the optimal solutions to the full primal and dual.

In fact, (A.23) is self-evident. The right-hand side is the optimal value of the reduced dual (i.e., replacing $\boldsymbol{X}$ and $\boldsymbol{\lambda}$ by $\boldsymbol{X}_T$ and $\boldsymbol{\lambda}^{[m]}$ in (A.10)), while the left-hand side agrees with the optimal value of the reduced primal since

$$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\mathrm{lift}}\|^2 = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{b}}_T\|^2 \text{ and } \sum_{i=1}^{p} \lambda_i |\widehat{\beta}^{\mathrm{lift}}|_{(i)} = \sum_{i=1}^{m} \lambda_i |\widehat{b}_T|_{(i)}.$$

Since the reduced primal only has linear equality constraints and is clearly feasible, strong duality holds, and (A.23) follows from this. □

Lemma A.7. *Let $1 \le k^\star < p$ be any (deterministic) integer, then*

$$\sup_{|T|=k^\star} \|\boldsymbol{X}_T' \boldsymbol{z}\| \le \sqrt{32 k^\star \log(p/k^\star)}$$

with probability at least $1 - \mathrm{e}^{-n/2} - (\sqrt{2}\mathrm{e}k^\star/p)^{k^\star}$. Above, the supremum is taken over all the subsets of $\{1, \dots, p\}$ with cardinality $k^\star$.

Proof of Lemma A.7. Conditional on $\boldsymbol{z}$, it is easy to see that $\boldsymbol{X}'\boldsymbol{z}$ is distributed as i.i.d. centered Gaussian random variables with variance $\|\boldsymbol{z}\|^2/n$. This observation enables us to write

$$\boldsymbol{X}'\boldsymbol{z} \overset{d}{=} \frac{\|\boldsymbol{z}\|}{\sqrt{n}}(\zeta_1, \dots, \zeta_p),$$

where $\boldsymbol{\zeta} := (\zeta_1, \dots, \zeta_p)$ consists of i.i.d. $\mathcal{N}(0,1)$ independent of $\|\boldsymbol{z}\|$. Hence, it is sufficient to prove that

$$\|\boldsymbol{z}\| \le 2\sqrt{n}, \quad |\zeta|_{(1)}^2 + \cdots + |\zeta|_{(k^\star)}^2 \le 8k^\star \log(p/k^\star)$$

simultaneously with probability at least $1 - \mathrm{e}^{-n/2} - \left(\sqrt{2}\mathrm{e}k^\star/p\right)^{k^\star}$. From Lemma A.5, we know that $\mathbb{P}(\|\boldsymbol{z}\| > 2\sqrt{n}) \le \mathrm{e}^{-n/2}$ so we just need to establish the other inequality. To this end, observe that

$$\begin{aligned}
\mathbb{P}\left(|\zeta|_{(1)}^2 + \cdots + |\zeta|_{(k^\star)}^2 > 8k^\star \log(p/k^\star)\right) &\le \frac{\mathbb{E}\,\mathrm{e}^{\frac{1}{4}\left(|\zeta|_{(1)}^2 + \cdots + |\zeta|_{(k^\star)}^2\right)}}{\mathrm{e}^{2k^\star \log \frac{p}{k^\star}}} \\
&\le \frac{\sum_{i_1 < \cdots < i_{k^\star}} \mathbb{E}\,\mathrm{e}^{\frac{1}{4}\left(|\zeta|_{i_1}^2 + \cdots + |\zeta|_{i_{k^\star}}^2\right)}}{\mathrm{e}^{2k^\star \log \frac{p}{k^\star}}} \\
&= \frac{\binom{p}{k^\star} 2^{k^\star/2}}{\mathrm{e}^{2k^\star \log \frac{p}{k^\star}}} \\
&\le \left(\frac{\sqrt{2}\mathrm{e}k^\star}{p}\right)^{k^\star}.
\end{aligned}$$

□

We record an elementary result which simply follows from $\Phi^{-1}(1 - c/2) \le \sqrt{2\log 1/c}$ for each $0 < c < 1$.

Lemma A.8. *Fix $0 < q < 1$. Then for all $1 \le k \le p/2$,*

$$\sum_{i=1}^{k} \left(\lambda_i^{\mathrm{BH}}\right)^2 \le C_q \cdot k \log(p/k),$$

for some constant $C_q > 0$.

In the next two lemmas, we use the BHq critical values $\boldsymbol{\lambda}^{\mathrm{BH}}$ to majorize sequences of Gaussian order statistics. Again, $\boldsymbol{a} \preceq \boldsymbol{b}$ means that $\boldsymbol{b}$ majorizes $\boldsymbol{a}$.

Lemma A.9. *Given any constant $c > 1/(1-q)$, suppose $\max\{ck, k+d\} \le k^\star < p$ for any (deterministic) sequence d that diverges to ∞. Let $\zeta_1, \dots, \zeta_{p-k}$ be i.i.d. $\mathcal{N}(0,1)$. Then*

$$\left(|\zeta|_{(k^\star - k + 1)}, |\zeta|_{(k^\star - k + 2)}, \dots, |\zeta|_{(p-k)}\right) \preceq \left(\lambda_{k^\star+1}^{\mathrm{BH}}, \lambda_{k^\star+2}^{\mathrm{BH}}, \dots, \lambda_p^{\mathrm{BH}}\right)$$

with probability approaching one.

Proof of Lemma A.9. It suffices to prove the stronger case where $\boldsymbol{\zeta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. Let $U_1, \dots, U_p$ be i.i.d. uniform random variables on $[0,1]$ and $U_{(1)} \le \cdots \le U_{(p)}$ the corresponding order statistics. Since

$$\left(|\zeta|_{(k^\star - k + 1)}, \dots, |\zeta|_{(p-k)}\right) \overset{d}{=} \left(\Phi^{-1}(1 - U_{(k^\star - k + 1)}/2), \dots, \Phi^{-1}(1 - U_{(p-k)}/2)\right),$$

the conclusion would follow from

$$\mathbb{P}\left(U_{(k^\star-k+j)} \geq q(k^\star+j)/p,\ \forall j \in \{1,\dots,p-k^\star\}\right) \to 1.$$

Let $E_1,\dots,E_{p+1}$ be i.i.d. exponential random variables with mean 1 and denote by $T_i = E_1 + \cdots + E_i$. Then the order statistics $U_{(i)}$ have the same joint distribution with T_i/T_{p+1}. Fixing an arbitrary constant $q' \in (q, 1-1/c)$, we have

$$\mathbb{P}\left(U_{(k^\star-k+j)} \geq q(k^\star+j)/p,\, \forall j\right) \geq \mathbb{P}\left(T_{k^\star-k+j} \geq q'(k^\star+j),\, \forall j\right) - \mathbb{P}\left(T_{p+1} > q'p/q\right).$$

Since $\mathbb{P}\left(T_{p+1} > q'p/q\right) \to 0$ by the law of large numbers, it is sufficient to prove

$$\mathbb{P}\left(T_{k^\star-k+j} \geq q'(k^\star+j),\, \forall j \in \{1,\dots,p-k^\star\}\right) \to 1. \tag{A.24}$$

This event can be rewritten as

$$T_{k^\star-k+j} - T_{k^\star-k} - q'j \geq q'k^\star - T_{k^\star-k}$$

for all $1 \leq j \leq p-k^\star$. Hence, (A.24) is reduced to proving

$$\mathbb{P}\left(\min_{1\leq j\leq p-k^\star} T_{k^\star-k+j} - T_{k^\star-k} - q'j \geq q'k^\star - T_{k^\star-k}\right) \to 1. \tag{A.25}$$

As a random walk, $T_{k^\star-k+j} - T_{k^\star-k} - q'j$ has i.i.d. increments with mean $1-q' > 0$ and variance 1. Thus $\min_{1\leq j\leq p-k^\star} T_{k^\star-k+j} - T_{k^\star-k} - q'j$ converges weakly to a bounded random variable in distribution. Consequently, (A.25) holds if one can demonstrate that $q'k^\star - T_{k^\star-k}$ diverges to $-\infty$ as $p\to\infty$ in probability. To see this, observe that

$$q'k^\star - T_{k^\star-k} = \frac{q'k^\star}{k^\star-k}(k^\star-k) - T_{k^\star-k} \leq \frac{q'c}{c-1}(k^\star-k) - T_{k^\star-k},$$

where we use the fact $k^\star \geq ck$. Under our hypothesis $q'c/(c-1) < 1$, the process $\{q'ct/(c-1) - T_t : t\in\mathbb{N}\}$ is a random walk drifting towards $-\infty$. Recognizing that $k^\star - k \geq d \to \infty$, we see that $q'c(k^\star-k)/(c-1) - T_{k^\star-k}$ (weakly) diverges to $-\infty$ since it corresponds to a position of the preceding random walk at $t\to\infty$. This concludes the proof.

□

Lemma A.10. *Let $\zeta_1,\dots,\zeta_{p-k}$ be i.i.d. $\mathcal{N}(0,1)$. Then there exists a constant C_q only depending on q such that*

$$(\zeta_1,\dots,\zeta_{p-k}) \preceq C_q \cdot \sqrt{\frac{\log p}{\log(p/k)}}\left(\lambda^{\mathrm{BH}}_{k+1},\dots,\lambda^{\mathrm{BH}}_{p}\right)$$

with probability tending to one as $p\to\infty$ and $k/p\to 0$.

Proof of Lemma A.10. Let $U_1,\dots,U_{p-k}$ be i.i.d. uniform random variables on $[0,1]$ and replace ζ_i by $\Phi^{-1}(1-U_i/2)$. Note that

$$\Phi^{-1}\left(1-U_i/2\right) \leq \sqrt{2\log\frac{2}{U_i}}, \quad \lambda^{\mathrm{BH}}_{k+i} \asymp \sqrt{2\log\frac{2p}{k+i}};$$

Hence, it suffices to prove that for some constant κ'_q,

$$\log(2/U_{(i)}) \log(p/k) \le \kappa'_q \cdot \log p \cdot \log(2p/(k+i)) \tag{A.26}$$

holds for all $i = 1, \dots, p-k$ with probability approaching one. Applying the representation given in the proof of Lemma A.9 and noting that $T_{p+1} = (1 + o_{\mathbb{P}}(1))p$, we see that (A.26) is implied by

$$\log(3p/T_i) \log(p/k) \le \kappa'_q \cdot \log p \cdot \log(2p/(k+i)). \tag{A.27}$$

We consider $i \le 4\sqrt{p}$ and $i > 4\sqrt{p}$ separately.

Suppose first that $i \le 4\sqrt{p}$. In this case,

$$\log(2p/(k+i)) = (1 + o(1)) \log(p/k).$$

Thus (A.27) would follow from

$$\log(3p/T_i) = O(\log p)$$

for all such i. This is, however, self-evident since $T_i \ge E_1 \ge 1/p$ with probability $1 - \mathrm{e}^{-1/p} = o(1)$.

Suppose now that $i > 4\sqrt{p}$. In this case, we make use of the fact that $T_i > i/2 - \sqrt{p}$ for all i with probability tending to one as $p \to \infty$. Then we prove a stronger result, namely,

$$\log \frac{3p}{i/2 - \sqrt{p}} \cdot \log \frac{p}{k} \le \kappa'_q \log p \cdot \log \frac{2p}{k+i}.$$

for all $i > 4\sqrt{p}$. This follows from the two observations below:

$$\log \frac{3p}{i/2 - \sqrt{p}} \asymp \log \frac{p}{i}, \;\; \log \frac{2p}{k+i} \ge \min \left\{ \log \frac{p}{i}, \log \frac{p}{k} \right\}.$$

□

In the proofs of the next two lemmas, namely, Lemma A.11 and Lemma A.12, we introduce an orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ that obeys

$$\boldsymbol{Q}\boldsymbol{z} = (\|\boldsymbol{z}\|, 0, \dots, 0)\,.$$

In the proofs, $\boldsymbol{Q}$ is further set to be measurable with respect to $\boldsymbol{z}$. Hence, $\boldsymbol{Q}$ is independent of $\boldsymbol{X}$. There are many options available to construct such a $\boldsymbol{Q}$, including the Householder transformation. Set

$$\boldsymbol{W} = \begin{bmatrix} \widetilde{\boldsymbol{w}} \\ \widetilde{\boldsymbol{W}} \end{bmatrix} := \boldsymbol{Q}\boldsymbol{X},$$

where $\widetilde{\boldsymbol{w}} \in \mathbb{R}^{1 \times p}$ and $\widetilde{\boldsymbol{W}} \in \mathbb{R}^{(n-1) \times p}$. The independence between $\boldsymbol{Q}$ and $\boldsymbol{X}$ suggests that $\boldsymbol{W}$ is still a Gaussian random matrix, consisting of i.i.d. $\mathcal{N}(0, 1/n)$ entries. Note that

$$\boldsymbol{X}_i'\boldsymbol{z} = (\boldsymbol{Q}\boldsymbol{X}_i)'(\boldsymbol{Q}\boldsymbol{z}) = \|\boldsymbol{z}\|(\boldsymbol{Q}\boldsymbol{X}_i)_1 = \|\boldsymbol{z}\|\widetilde{w}_i.$$

This implies that $S^\star$ is constructed as the union of S and the $k^\star - k$ indices in $\{1, \dots, p\} \setminus S$ with the largest $|\widetilde{w}_i|$. Since $\widetilde{\boldsymbol{w}}$ and $\widetilde{\boldsymbol{W}}$ are independent, we see that both $\widetilde{\boldsymbol{W}}_{\overline{S^\star}}$ and $\widetilde{\boldsymbol{W}}_{S^\star}$ are also Gaussian random matrices. These points are crucial in the proof of these two lemmas.

Lemma A.11. *Let $k < k^\star < \min\{n, p\}$ be any (deterministic) integer. Denote by $\sigma_{\min}$ and $\sigma_{\max}$, respectively, the smallest and the largest singular value of $\boldsymbol{X}_{S^\star}$. Then for any $t > 0$,*

$$\sigma_{\min} > \sqrt{1 - 1/n} - \sqrt{k^\star/n} - t$$

holds with probability at least $1 - \mathrm{e}^{-nt^2/2}$. Furthermore,

$$\sigma_{\max} < \sqrt{1 - 1/n} + \sqrt{k^\star/n} + \sqrt{8k^\star \log(p/k^\star)/n} + t$$

holds with probability at least $1 - \mathrm{e}^{-nt^2/2} - (\sqrt{2}\mathrm{e}k^\star/p)^{k^\star}$.

Proof of Lemma A.11. Recall that $\widetilde{\boldsymbol{W}}_{S^\star} \in \mathbb{R}^{(n-1)\times k^\star}$ is a Gaussian design with i.i.d. $\mathcal{N}(0, 1/n)$ entries. Since $\boldsymbol{W}_{S^\star}$ and $\boldsymbol{X}_{S^\star}$ have the the same set of singular values, we consider $\boldsymbol{W}_{S^\star}$.

Classical theory on Wishart matrices (see [53], for example) asserts that (i) all the singular values of $\widetilde{\boldsymbol{W}}_{S^\star}$ are larger than $\sqrt{1 - 1/n} - \sqrt{k^\star/n} - t$ with probability at least $1 - \mathrm{e}^{-nt^2/2}$, and (ii) are all smaller than $\sqrt{1 - 1/n} + \sqrt{k^\star/n} + t$ with probability at least $1 - \mathrm{e}^{-nt^2/2}$. Clearly, all the singular values larger of $\boldsymbol{W}_{S^\star}$ are at least as large as $\sigma_{\min}(\widetilde{\boldsymbol{W}}_{S^\star})$. Thus, (i) yields the first claim. For the other, Lemma A.7 asserts that the event $\|\widetilde{\boldsymbol{w}}_{S^\star}\| \le \sqrt{8k^\star \log(p/k^\star)}$ happens with probability at least $1 - (\sqrt{2}\mathrm{e}k^\star/p)^{k^\star}$. On this event,

$$\|\boldsymbol{W}_{S^\star}\| \le \sqrt{\|\widetilde{\boldsymbol{W}}_{S^\star}\|^2 + 8k^\star \log(p/k^\star)},$$

where $\|\cdot\|$ denotes the spectral norm. Hence, (ii) gives

$$\|\boldsymbol{W}_{S^\star}\| \le \|\widetilde{\boldsymbol{W}}_{S^\star}\| + \sqrt{8k^\star \log(p/k^\star)} \le \sqrt{1 - 1/n} + \sqrt{k^\star/n} + t + \sqrt{8k^\star \log(p/k^\star)}$$

with probability at least $1 - \mathrm{e}^{-nt^2/2} - (\sqrt{2}\mathrm{e}k^\star/p)^{k^\star}$.

□

Lemma A.12. *Denote by $\widehat{\boldsymbol{b}}_{S^\star}$ the solution to the reduced SLOPE problem (4.6) with $T = S^\star$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\boldsymbol{\epsilon}}$. Keep the assumptions from Lemma A.9, and additionally assume $k^\star/\min\{n, p\} \to 0$. Then there exists a constant C_q only depending on q such that*

$$\boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{X}_{S^\star}(\boldsymbol{\beta}_{S^\star} - \widehat{\boldsymbol{b}}_{S^\star}) \preceq C_q \cdot \sqrt{\frac{k^\star \log p}{n}} \left(\lambda^{\mathrm{BH}}_{k^\star+1}, \ldots, \lambda^{\mathrm{BH}}_{p}\right)$$

with probability tending to one.

Proof of Lemma A.12. In this proof, C is a constant that only depends on q and whose value may change at each occurrence. Rearrange the objective term as

$$\begin{aligned}
\boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{X}_{S^\star}(\boldsymbol{\beta}_{S^\star} - \widehat{\boldsymbol{b}}_{S^\star}) &= \boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{X}_{S^\star}(\boldsymbol{X}'_{S^\star}\boldsymbol{X}_{S^\star})^{-1}(\boldsymbol{X}'_{S^\star}(\boldsymbol{y} - \boldsymbol{X}_{S^\star}\widehat{\boldsymbol{b}}_{S^\star}) - \boldsymbol{X}'_{S^\star}\boldsymbol{z}) \\
&= \boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{X}_{S^\star}(\boldsymbol{X}'_{S^\star}\boldsymbol{X}_{S^\star})^{-1}(\boldsymbol{X}'_{S^\star}(\boldsymbol{y} - \boldsymbol{X}_{S^\star}\widehat{\boldsymbol{b}}_{S^\star}) - \boldsymbol{X}'_{S^\star}\boldsymbol{z}) \\
&= \boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{Q}'\boldsymbol{\xi},
\end{aligned}$$

where

$$\boldsymbol{\xi} := \boldsymbol{Q}\boldsymbol{X}_{S^\star}(\boldsymbol{X}'_{S^\star}\boldsymbol{X}_{S^\star})^{-1}\left(\boldsymbol{X}'_{S^\star}(\boldsymbol{y} - \boldsymbol{X}_{S^\star}\widehat{\boldsymbol{b}}_{S^\star}) - \boldsymbol{X}'_{S^\star}\boldsymbol{z}\right).$$

We begin by bounding $\|\boldsymbol{\xi}\|$. It follows from the KKT condition of SLOPE that $\boldsymbol{X}'_{S^\star}(\boldsymbol{y}-\boldsymbol{X}_{S^\star}\widehat{\boldsymbol{b}}_{S^\star})$ is majorized by $\boldsymbol{\lambda}^{[k^\star]}$. Hence, it follows from Fact 3.1 that

$$\left\|\boldsymbol{X}'_{S^\star}(\boldsymbol{y}-\boldsymbol{X}_{S^\star}\widehat{\boldsymbol{b}}_{S^\star})\right\| \le \|\boldsymbol{\lambda}^{[k^\star]}\|. \tag{A.28}$$

Lemma A.11 with $t = 1/2$ gives

$$\left\|\boldsymbol{X}_{S^\star}(\boldsymbol{X}'_{S^\star}\boldsymbol{X}_{S^\star})^{-1}\right\| \le \left(\sqrt{1-1/n}-\sqrt{k^\star/n}-1/2\right)^{-1} < 2.01 \tag{A.29}$$

with probability at least $1-\mathrm{e}^{-n/8}$ for sufficiently large p, where in the last step we have used $k^\star/n \to 0$. Hence, from (A.28) and (A.29) we get

$$\begin{aligned}
\|\boldsymbol{\xi}\| &\le \left\|\boldsymbol{X}_{S^\star}(\boldsymbol{X}'_{S^\star}\boldsymbol{X}_{S^\star})^{-1}\right\| \cdot \left\|\boldsymbol{X}'_{S^\star}(\boldsymbol{y}-\boldsymbol{X}_{S^\star}\widehat{\boldsymbol{b}}_{S^\star}) - \boldsymbol{X}'_{S^\star}\boldsymbol{z}\right\| \\
&\le 2.01\left(\|\boldsymbol{\lambda}^{[k^\star]}\| + 4\sqrt{2k^\star\log(p/k^\star)}\right) \\
&\le 2.01\left((1+\epsilon)\sqrt{C}+4\sqrt{2}\right)\sqrt{k^\star\log(p/k^\star)} \\
&= C\cdot\sqrt{k^\star\log(p/k^\star)}
\end{aligned} \tag{A.30}$$

with probability at least $1-\mathrm{e}^{-n/2}-(\sqrt{2}\mathrm{e}k^\star/p)^{k^\star}-\mathrm{e}^{-n/8} \to 1$; we used Lemma A.7 in the second line and Lemma A.8 in the third. (A.30) will help us in finishing the proof.

Write

$$\boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{X}_{S^\star}(\boldsymbol{\beta}_{S^\star}-\widehat{\boldsymbol{b}}_{S^\star}) = \boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{Q}'\boldsymbol{\xi} = \boldsymbol{W}'_{\overline{S^\star}}\boldsymbol{\xi} = \left(\widetilde{\boldsymbol{w}}'_{\overline{S^\star}},\boldsymbol{0}\right)\boldsymbol{\xi} + \left(\boldsymbol{0},\widetilde{\boldsymbol{W}}'_{\overline{S^\star}}\right)\boldsymbol{\xi}. \tag{A.31}$$

It follows from Lemma A.9 that $\widetilde{\boldsymbol{w}}_{\overline{S^\star}}$ is majorized by $\left(\lambda^{\mathrm{BH}}_{k^\star+1},\lambda^{\mathrm{BH}}_{k^\star+2},\ldots,\lambda^{\mathrm{BH}}_{p}\right)/\sqrt{n}$ in probability. As a result, the first term in the right-hand side obeys

$$\left(\widetilde{\boldsymbol{w}}'_{\overline{S^\star}},\boldsymbol{0}\right)\boldsymbol{\xi} = \xi_1\cdot\widetilde{\boldsymbol{w}}'_{\overline{S^\star}} \preceq \|\boldsymbol{\xi}\|\cdot\widetilde{\boldsymbol{w}}'_{\overline{S^\star}} \preceq C\cdot\sqrt{\frac{k^\star}{n}\log\frac{p}{k^\star}}\left(\lambda^{\mathrm{BH}}_{k^\star+1},\lambda^{\mathrm{BH}}_{k^\star+2},\ldots,\lambda^{\mathrm{BH}}_{p}\right) \tag{A.32}$$

with probability tending to one. For the second term, by exploiting the independence between $\boldsymbol{\xi}$ and $\widetilde{\boldsymbol{W}}_{\overline{S^\star}}$, we have

$$\left(\boldsymbol{0},\widetilde{\boldsymbol{W}}'_{\overline{S^\star}}\right)\boldsymbol{\xi} \overset{d}{=} \sqrt{\frac{\xi_2+\cdots+\xi_n^2}{n}}(\zeta_1,\ldots,\zeta_{p-k^\star}),$$

where $\zeta_1,\ldots,\zeta_{p-k^\star}$ are i.i.d. $\mathcal{N}(0,1/n)$. Since $k^\star/p \to 0$, applying Lemma A.10 gives

$$(\zeta_1,\ldots,\zeta_{p-k^\star}) \preceq C\cdot\sqrt{\frac{\log p}{\log(p/k^\star)}}\left(\lambda^{\mathrm{BH}}_{k^\star+1},\ldots,\lambda^{\mathrm{BH}}_{p}\right)$$

with probability approaching one. Hence, owing to (A.30),

$$\left(\boldsymbol{0},\widetilde{\boldsymbol{W}}'_{\overline{S^\star}}\right)\boldsymbol{\xi} \preceq C\cdot\sqrt{\frac{k^\star\log p}{n}}\left(\lambda^{\mathrm{BH}}_{k^\star+1},\ldots,\lambda^{\mathrm{BH}}_{p}\right) \tag{A.33}$$

holds with probability approaching one. Finally, combining (A.32) and (A.33) gives that

$$\boldsymbol{X}'_{\overline{S^\star}}\boldsymbol{X}_{S^\star}(\boldsymbol{\beta}_{S^\star}-\widehat{\boldsymbol{b}}_{S^\star}) = \left(\widetilde{\boldsymbol{w}}'_{\overline{S^\star}},\boldsymbol{0}\right)\boldsymbol{\xi} + \left(\boldsymbol{0},\widetilde{\boldsymbol{W}}'_{\overline{S^\star}}\right)\boldsymbol{\xi}$$

$$\preceq C \cdot \left(\sqrt{\frac{k^\star}{n} \log \frac{p}{k^\star}} + \sqrt{\frac{k^\star \log p}{n}} \right) \cdot \left(\lambda^{\mathrm{BH}}_{k^\star+1}, \ldots, \lambda^{\mathrm{BH}}_p \right)$$
$$\preceq C \cdot \sqrt{\frac{k^\star \log p}{n}} \left(\lambda^{\mathrm{BH}}_{k^\star+1}, \ldots, \lambda^{\mathrm{BH}}_p \right)$$

holds with probability tending to one. □

A.4 Proofs for Section 5

Lemma A.13. *Keep the assumptions from Lemma 5.1 and let $\zeta_1, \ldots, \zeta_p$ be i.i.d. $\mathcal{N}(0,1)$. Then*

$$\#\left\{2 \le i \le p : \zeta_i > \tau + \zeta_1\right\} \to \infty$$

in probability.

Proof of Lemma A.13. With probability tending to one, $\tau' := \tau + \zeta_1$ also obeys $\tau'/\sqrt{2\log p} \to 1$ and $\sqrt{2\log p} - \tau' \to \infty$. This shows that we only need to prove a simpler version of this lemma, namely, $\#\left\{1 \le i \le p : \zeta_i > \tau\right\} \to \infty$ in probability.

Put $\Delta = \sqrt{2\log p} - \tau = o(\sqrt{2\log p})$ and $a = \mathbb{P}(\xi_1 > \tau)$. Then, $\#\left\{1 \le i \le p : \zeta_i > \tau\right\}$ is a binomial random variable with p trials and success probability a. Hence, it suffices to demonstrate that $ap \to \infty$. To this end, note that

$$a = 1 - \Phi(\tau) \sim \frac{1}{\tau} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\tau^2}{2}} \asymp \frac{1}{\sqrt{2\log p}} \mathrm{e}^{-\log p - \Delta^2/2 + \Delta\sqrt{2\log p}}$$
$$= \frac{1}{p\sqrt{2\log p}} \mathrm{e}^{(1+o(1))\Delta\sqrt{2\log p}},$$

which gives

$$ap \asymp \frac{1}{\sqrt{2\log p}} \mathrm{e}^{(1+o(1))\Delta\sqrt{2\log p}}.$$

Since $\Delta \to \infty$ (in fact, it is sufficient to have Δ bounded away from 0 from below), we have

$$\mathrm{e}^{(1+o(1))\Delta\sqrt{2\log p}} / \sqrt{2\log p} \to \infty,$$

as we wish. □

Proof of Lemma 5.1. For sufficiently large p, $2(1-\epsilon)\log p \le (1-\epsilon/2)\tau^2$. Hence, it is sufficient to show

$$\mathbb{P}_{\boldsymbol{\pi}} \left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \le (1-\epsilon/2)\tau^2 \right) \to 0$$

uniformly for all estimators $\widehat{\boldsymbol{\beta}}$. Letting I be the random coordinate,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = \sum_{j \ne I} \widehat{\beta}_j^2 + (\widehat{\beta}_I - \tau)^2 = \|\widehat{\boldsymbol{\beta}}\|^2 + \tau^2 - 2\tau\widehat{\beta}_I,$$

which is smaller than or equal to $(1-\epsilon/2)\tau^2$ if and only if

$$\widehat{\beta}_I \ge \frac{2\|\widehat{\boldsymbol{\beta}}\|^2 + \epsilon\tau^2}{4\tau}.$$

Denote by $A = A(\boldsymbol{y}; \widehat{\boldsymbol{\beta}})$ the set of all $i \in \{1, \ldots, p\}$ such that $\widehat{\beta}_i \geq (2\|\widehat{\boldsymbol{\beta}}\|^2 + \epsilon\tau^2)/(4\tau)$, and let $\widehat{b}$ be the minimum value of these $\widehat{\beta}_i$. Then

$$\widehat{b} \geq \frac{2\|\widehat{\boldsymbol{\beta}}\|^2 + \epsilon\tau^2}{4\tau} \geq \frac{2|A|\widehat{b}^2 + \epsilon\tau^2}{4\tau} \geq \frac{2\sqrt{2|A|\widehat{b}^2 \cdot \epsilon\tau^2}}{4\tau},$$

which gives

$$|A| \leq 2/\epsilon. \tag{A.34}$$

Recall that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \leq (1 - \epsilon/2)\tau^2$ if and only if I is among these $|A|$ components. Hence,

$$\mathbb{P}_{\boldsymbol{\pi}}\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \leq (1 - \epsilon/2)\tau^2 \Big| \boldsymbol{y}\right) = \mathbb{P}_{\boldsymbol{\pi}}(I \in A|\boldsymbol{y}) = \sum_{i \in A} \mathbb{P}_{\boldsymbol{\pi}}(I = i|\boldsymbol{y}) = \frac{\sum_{i \in A} \mathrm{e}^{\tau y_i}}{\sum_{i=1}^{p} \mathrm{e}^{\tau y_i}}, \tag{A.35}$$

where we use the fact that A is almost surely determined by $\boldsymbol{y}$. Since (A.35) is maximal if A is the set of indices with the largest y_i's, (A.35) and (A.34) together yield

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\pi}}&\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \leq (1 - \epsilon/2)\tau^2\right) \\ &\leq \mathbb{P}_{\boldsymbol{\pi}}\left(y_I = \tau + z_I \text{ is at least the } \lceil 2/\epsilon \rceil^{\text{th}} \text{ largest among } y_1, \ldots, y_p\right) \to 0,\end{aligned}$$

where the last step is provided by Lemma A.13.

□

Proof of Lemma 5.3. To closely follow the proof of Lemma 5.1, denote by $A = A(\boldsymbol{y}, \boldsymbol{X}; \widehat{\boldsymbol{\beta}})$ the set of all $i \in \{1, \ldots, p\}$ such that $\widehat{\beta}_i \geq (2\|\widehat{\boldsymbol{\beta}}\|^2 + \epsilon\alpha^2\tau^2)/(4\alpha\tau)$, and keep the same notation $\widehat{b}$ as before. Then $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \leq (1 - \epsilon/2)\alpha^2\tau^2$ if and only if $I \in A$. Hence,

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\pi}}\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \leq (1 - \epsilon/2)\alpha^2\tau^2 \Big| \boldsymbol{y}, \boldsymbol{X}\right) &= \mathbb{P}_{\boldsymbol{\pi}}(I \in A|\boldsymbol{y}, \boldsymbol{X}) \\ &= \sum_{i \in A} \mathbb{P}_{\boldsymbol{\pi}}(I = i|\boldsymbol{y}, \boldsymbol{X}) \\ &= \frac{\sum_{i \in A} \exp\left(\alpha\tau \boldsymbol{X}_i'\boldsymbol{y} - \alpha^2\tau^2\|\boldsymbol{X}_i\|^2/2\right)}{\sum_{i=1}^{p} \exp\left(\alpha\tau \boldsymbol{X}_i'\boldsymbol{y} - \alpha^2\tau^2\|\boldsymbol{X}_i\|^2/2\right)}\end{aligned} \tag{A.36}$$

and this quantity is maximal if A is the set of indices i with the largest values of $\boldsymbol{X}_i'\boldsymbol{y}/\alpha - \tau\|\boldsymbol{X}_i\|^2/2$. As shown in Lemma 5.1, $|A| \leq 2/\epsilon$, which gives

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\pi}}&\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \leq (1 - \epsilon/2)\alpha^2\tau^2\right) \\ &\leq \mathbb{P}_{\boldsymbol{\pi}}\left(\boldsymbol{X}_I'\boldsymbol{y}/\alpha - \tau\|\boldsymbol{X}_I\|^2/2 \text{ is at least the } \lceil 2/\epsilon \rceil^{\text{th}} \text{ largest}\right).\end{aligned} \tag{A.37}$$

We complete the proof by showing that the probability in the right-hand side of (A.37) is negligible uniformly over all estimators $\widehat{\boldsymbol{\beta}}$ as $p \to \infty$. By the independence between I and $\boldsymbol{X}, \boldsymbol{z}$, we can assume $I = 1$ while evaluating this probability. With this in mind, we aim to show that there are sufficiently many i's such that

$$\boldsymbol{X}_i'\boldsymbol{y}/\alpha - \frac{\tau}{2}\|\boldsymbol{X}_i\|^2 - \boldsymbol{X}_1'\boldsymbol{y}/\alpha + \frac{\tau}{2}\|\boldsymbol{X}_1\|^2 = \boldsymbol{X}_i'\left(\boldsymbol{z}/\alpha + \tau\boldsymbol{X}_1\right) - \frac{\tau}{2}\|\boldsymbol{X}_i\|^2 - \boldsymbol{X}_1'\boldsymbol{z}/\alpha - \frac{\tau}{2}\|\boldsymbol{X}_1\|^2$$

is positive. Since

$$\boldsymbol{X}_1'\boldsymbol{z}/\alpha + \frac{\tau}{2}\|\boldsymbol{X}_1\|^2 = O_{\mathbb{P}}(1/\alpha) + \frac{\tau}{2}\left(1 + O_{\mathbb{P}}(1/\sqrt{n})\right),$$

it suffices to show that

$$\#\left\{2 \le i \le p : \boldsymbol{X}_i'\left(\boldsymbol{z}/\alpha + \tau\boldsymbol{X}_1\right) - \frac{\tau}{2}\|\boldsymbol{X}_i\|^2 > \frac{C_1}{\alpha} + \frac{\tau}{2} + \frac{C_2\tau}{\sqrt{n}}\right\} \le \lceil 2/\epsilon \rceil - 1 \tag{A.38}$$

holds with vanishing probability for all positive constants C_1, C_2. By the independence between $\boldsymbol{X}_i$ and $\boldsymbol{z}/\alpha + \tau\boldsymbol{X}_1$, we can replace $\boldsymbol{z}/\alpha + \tau\boldsymbol{X}_1$ by $(\|\boldsymbol{z}/\alpha + \tau\boldsymbol{X}_1\|, 0, \ldots, 0)$ in (A.38). That is,

$$\boldsymbol{X}_i'\left(\boldsymbol{z}/\alpha + \tau\boldsymbol{X}_1\right) - \frac{\tau}{2}\|\boldsymbol{X}_i\|^2 \stackrel{d}{=} \|\boldsymbol{z}/\alpha + \tau\boldsymbol{X}_1\|\, X_{i,1} - \frac{\tau}{2}X_{i,1}^2 - \frac{\tau}{2}\|\boldsymbol{X}_{i,-1}\|^2,$$

where $\boldsymbol{X}_{i,-1} \in \mathbb{R}^{n-1}$ is $\boldsymbol{X}_i$ without the first entry. To this end, we point out that the following three events all happen with probability tending to one:

$$\begin{aligned} \#\{2 \le i \le p : \|\boldsymbol{X}_{i,-1}\| \le 1\}/p &\to 1/2, \\ \max_i X_{i,1}^2 &\le \frac{2\log p}{n}, \\ \|\boldsymbol{z}/\alpha + \tau\boldsymbol{X}_1\| &\ge \left(\sqrt{n} - \sqrt{\log p}\right)/\alpha. \end{aligned} \tag{A.39}$$

Making use of this and (A.38), we only need to show that

$$\begin{aligned} N \triangleq \#\left\{2 \le i \le 0.49p : \frac{1}{\alpha}\left(1 - \sqrt{(\log p)/n}\right)\sqrt{n}X_{i,1} > \frac{\tau\log p}{n} + \frac{\tau}{2} + \frac{C_1}{\alpha} + \frac{\tau}{2} + \frac{C_2\tau}{\sqrt{n}}\right\} \\ \#\left\{2 \le i \le 0.49p : \frac{1}{\alpha}\left(1 - \sqrt{(\log p)/n}\right)\sqrt{n}X_{i,1} > \tau + \frac{\tau\log p}{n} + \frac{C_1}{\alpha} + \frac{C_2\tau}{\sqrt{n}}\right\} \end{aligned}$$

obeys

$$N \le \lceil 2/\epsilon \rceil - 1 \tag{A.40}$$

with vanishing probability. The first line of (A.39) shows that there are at least $0.49p$ many i's such that $\|\boldsymbol{X}_{i,-1}\| \le 1$ and we assume they correspond to indices $2 \le i \le 0.49p$ without loss of generality. (Note that N is independent of all $\boldsymbol{X}_{i,-1}$'s.) Observe that

$$\tau' := \frac{\tau + \tau(\log p)/n + C_1/\alpha + C_2\tau/\sqrt{n}}{\left(1 - \sqrt{(\log p)/n}\right)/\alpha} = \alpha\left(1 + 2\sqrt{\frac{\log p}{n}}\right)\tau + O(1)$$

for sufficiently large p (to ensure $(\log p)/n$ is small). Hence, plugging the specific choice of τ and using $\alpha \le 1$, we obtain

$$\tau' \le \left(1 + 2\sqrt{(\log p)/n}\right)\tau + O(1) \le \sqrt{2\log p} - \log\sqrt{2\log p} + O(1),$$

which reveals that $\sqrt{2\log(0.49p)} - \tau' = \sqrt{2\log p} - \tau' + o(1) \to \infty$. Since $\sqrt{n}X_{n,i}$ are i.i.d. $\mathcal{N}(0,1)$, Lemma A.13 validates (A.40). □

Proof of Corollary 1.5. Let $c > 0$ be a sufficiently small constant to be determined later. It is sufficient to prove the claim with p replaced by a possibly smaller value given by $p^\star := \min\{\lfloor cn \rfloor, p\}$ (if we knew that $\beta_i = 0$ for $p^\star + 1 \le i \le p$, the loss of any estimator $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ would not increase after projecting onto the linear space spanned by the first $p^\star$ columns). Hereafter, we assume $\boldsymbol{X} \in \mathbb{R}^{n \times p^\star}$ and $\boldsymbol{\beta} \in \mathbb{R}^{p^\star}$. Observe that $p = O(n)$ implies $p = O(p^\star)$ and, therefore,

$$\log(p^\star/k) \sim \log(p/k). \tag{A.41}$$

In particular, $k/p^\star \to 0$ and $n/\log(p^\star/k) \to \infty$. This suggests that we can apply Theorem 5.4 to our problem, obtaining

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P}\left(\frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2k \log(p^\star/k)} > 1 - \epsilon' \right) \to 1.$$

for every constant $\epsilon' > 0$. Because of (A.41), we also have

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P}\left(\frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{2k \log(p/k)} > 1 - \epsilon' \right) \to 1 \tag{A.42}$$

for any $\epsilon' > 0$.

Since $p^\star/n \le c \le 1$, the smallest singular value of the Gaussian random matrix $\boldsymbol{X}$ is at least $1 - \sqrt{c} + o_{\mathbb{P}}(1)$ (see, for example, [53]). This result, together with (A.42), yields

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{P}\left(\frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2k \log(p/k)} > (1 - \sqrt{c})^2 (1 - \epsilon') \right) \to 1$$

for each $\epsilon' > 0$. Finally, choose c and ϵ' sufficiently small such that $(1 - \sqrt{c})^2(1 - \epsilon') > 1 - \epsilon$. □