The Annals of Statistics
2016, Vol. 44, No. 3, 1038–1068
DOI: 10.1214/15-AOS1397

SLOPE IS ADAPTIVE TO UNKNOWN SPARSITY AND ASYMPTOTICALLY MINIMAX

BY WEIJIE SU[1] AND EMMANUEL CANDÈS[2]

Stanford University

We consider high-dimensional sparse regression problems in which we observe $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}$, where $\mathbf{X}$ is an $n \times p$ design matrix and $\mathbf{z}$ is an n-dimensional vector of independent Gaussian errors, each with variance σ^2. Our focus is on the recently introduced SLOPE estimator [*Ann. Appl. Stat.* **9** (2015) 1103–1140], 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 $\mathbf{X}$ 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

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

as the dimension p increases to ∞, and where $\varepsilon > 0$ is an arbitrary small constant. This holds under a weak assumption on the ℓ_0-sparsity level, namely, $k/p \to 0$ and $(k \log p)/n \to 0$, 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 ℓ_0-sparsity classes. We are not aware of any other estimator with this property.

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 $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 \mathbf{I}_p)$ of independent statistics $\{y_i\}$, and wish to test which means β_i are nonzero. Begin by ordering the observations as

Received April 2015; revised September 2015.

[1] Supported in part by a General Wang Yaowu Stanford Graduate Fellowship.

[2] Supported in part by NSF Grant CCF-0963835 and by the Math + X Award from the Simons Foundation.

MSC2010 subject classifications. Primary 62C20; secondary 62G05, 62G10, 62J15.

Key words and phrases. SLOPE, sparse regression, adaptivity, false discovery rate (FDR), Benjamini–Hochberg procedure, FDR thresholding.

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

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

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

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

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 $\mathbb{P}(|z| \geq \lambda_i^{\mathrm{BH}}) = q \cdot i/p$.] Then BHq rejects all those hypotheses with $|y_i| \geq \widehat{t}_{\mathrm{FDR}}$ and makes no rejection in the case where all the observations fall below the critical curve, that is, when the set $\{i : |y|_{(i)}/\sigma \geq \lambda_i^{\mathrm{BH}}\}$ 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

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

where $p_0 = |\{i : \beta_i = 0\}|$ is the total number of nulls. Unlike the 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 $\mathbf{y}$. 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 $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2\mathbf{I}_p)$ 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:[3] 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

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

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 $\log^5 p \leq k \leq p^{1-\delta}$ for any constant $\delta > 0$. Then as $p \to \infty$, it holds that

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

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

$$\underset{\mathbf{b}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{b}\|_2^2 + \sigma^2 \operatorname{Pen}\bigl(\|\mathbf{b}\|_0\bigr), \tag{1.4}$$

where $\operatorname{Pen}(k) = 2k \log(p/k)$ see [35] and [13, 52] for related ideas. In fact, [2] begins by considering the penalized MLE with

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

[3] See [3] for the use of this word.

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

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

where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of regression coefficients and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ 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 $\mathbf{X}$ has full column rank, this is equivalent to testing a vector of means under arbitrary correlations since the model is equivalent to $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ ($\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ 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 [5], which is a powerful FDR controlling method working when $p \le n$, and other innovative ideas [31, 38, 42, 43] 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 [51]. 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

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

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 $\mathbf{X} = \mathbf{I}_p$, the SLOPE estimate is the solution to

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

in the literature on optimization, this solution is called the prox to the sorted ℓ_1 norm evaluated at $\mathbf{y}$, hence the notation in the left-hand side. [In the case of a general orthogonal design in which $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$, the SLOPE solution is $\operatorname{prox}_{\boldsymbol{\lambda}}(\mathbf{X}'\mathbf{y})$.] Suppose the observations are nonnegative and already ordered, that is, $y_1 \geq y_2 \geq \cdots \geq y_p \geq 0$.[4] Then by [15], Proposition 2.2, SLOPE can be recast as the solution to

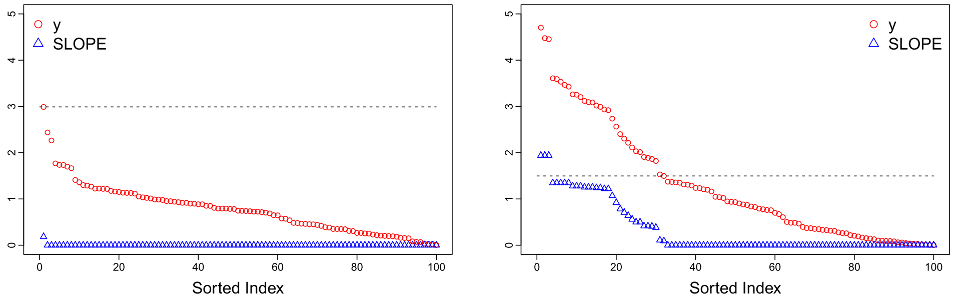
$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|\mathbf{y}-\boldsymbol{\lambda}-\mathbf{b}\|^2 = \frac{1}{2}\sum_i (y_i - \lambda_i - b_i)^2 \\ &\text{subject to} \quad b_1 \geq b_2 \geq \cdots \geq b_p \geq 0 \end{aligned} \tag{1.8}$$

so that it is equivalent to solving an isotonic regression problem with data $\mathbf{y} - \boldsymbol{\lambda}$. Hence, methods like the pool adjacent violators algorithm (PAVA) [6, 41] are directly applicable. Further, 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, that is, $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2\mathbf{I}_p)$, SLOPE controls the FDR [15], namely, FDR(SLOPE) $\leq qp_0/p$, where again p_0 is the number of nulls, that is, 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

[4] For arbitrary data, the solution can be obtained as follows: let $\mathbf{P}$ be a permutation that sorts the magnitudes $|\mathbf{y}|$ in a nonincreasing fashion. Then $\operatorname{prox}_{\boldsymbol{\lambda}}(\mathbf{y}) = \operatorname{sgn}(\mathbf{y}) \odot \mathbf{P}^{-1} \operatorname{prox}_{\boldsymbol{\lambda}}(\mathbf{P}|\mathbf{y}|)$, 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.



(a) Weak signals.

(b) Strong signals.

FIG. 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.*

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.

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 $\mathbf{X}'$ gives the statistically equivalent Gaussian sequence model

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

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_p)$. Estimating a sparse mean vector from Gaussian data is a well-studied problem with a long line of contributions; see [10, 13, 20, 24, 34, 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

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

where the infimum is taken over all measurable estimators; see [25] 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 $\mathbf{X}$ 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 $\delta > 0$, that is, 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. This idea may also be found in the recent work [39] in which the authors prove that some smooth-thresholding procedures uniformly achieve asymptotic minimaxity under the same assumptions as in Theorem 1.1. They also establish some optimality results for these thresholding rules at a fixed $\boldsymbol{\beta}$. 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:

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

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

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

Since $\sum_{1\le j\le k}(\lambda_j^{\text{BH}})^2 = (1 + o(1))2k\log(p/k)$,[5] summing this approximation over the first k coordinates gives

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

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.

[5] 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.

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

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

for all k in the stated range. Since $\lambda_j^2 = \sum_{i=1}^{j} \lambda_i^2 - \sum_{i=1}^{j-1} \lambda_i^2$, we have that λ_j^2 is roughly the derivative of $f(x) = 2x\log(p/x)$ at $x = j$ yielding $\lambda_j^2 \approx f'(j) = 2\log p - 2\log j - 2$, or

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

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 $\sqrt{2\log(p/j)}$.

We speculate that Theorem 1.1—and to some extent Theorem 1.2 below—extend to other loss functions. For instance, from the proofs of Theorem 1.1 we believe that for $r \ge 1$,

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

holds. Furthermore, examining the proof of Theorem 1.1 reveals that for all k not necessarily obeying $k/p \to 0$ (e.g., $k = p/2$),

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

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 $\mathbf{X}$ is a Gaussian random design with X_{ij} i.i.d. $\mathcal{N}(0, 1/n)$ so that the columns of $\mathbf{X}$ 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 [4, 7, 12, 18, 19, 28, 55]. 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* $\boldsymbol{\lambda} = \sigma(1+\varepsilon)\boldsymbol{\lambda}^{\mathrm{BH}}(q)$ *for some arbitrary constant* $0 < \varepsilon < 1$. *Suppose* $k/p \to 0$ *and* $(k\log p)/n \to 0$. *Then*

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

For information, it is known that under some regularity conditions on the design [47, 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* $\varepsilon > 0$, *we have*

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

Similar results dealing with arbitrary designs can be found in the literature, compare Theorem 1 in [57]. However, the notable difference is that our theorem captures the exact constants in addition to the rate.

Taking Theorems 1.2 and 1.3 together 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 [29, 30]. 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 $\mathbf{X}'\mathbf{y}$. 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 $\operatorname{prox}_{\boldsymbol{\lambda}}(\mathbf{X}'\mathbf{y})$; that is, by applying $\mathbf{X}'$ and then pretending that we are dealing with an orthogonal design. Such estimates turn out to have unbounded risks.

We remark that a preprint [33] 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 $(\mathbf{x}^*, y^*)$, independent from the "training set" $(\mathbf{X}, \mathbf{y})$, obeying the linear model $y^* = \langle \mathbf{x}^*, \boldsymbol{\beta} \rangle + \sigma z^*$ with $\mathbf{x} \sim \mathcal{N}(0, n^{-1}\mathbf{I}_p)$ and $z^* \sim \mathcal{N}(0, \sigma^2)$. Then for any estimate $\widehat{\boldsymbol{\beta}}$, the prediction $\widehat{y} = \langle \mathbf{x}^*, \widehat{\boldsymbol{\beta}} \rangle$ 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

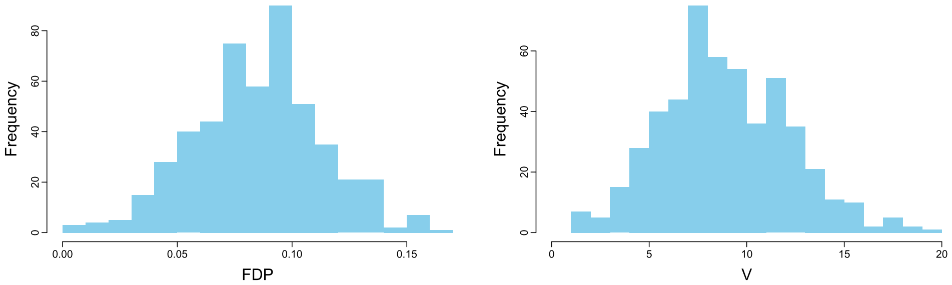
$$V \le \frac{q}{1-q}k. \tag{1.11}$$

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

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

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.

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 $\|\mathbf{X}(\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}} - \boldsymbol{\beta})\|^2$ is very nearly equal to $\|\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}} - \boldsymbol{\beta}\|^2$ since skinny Gaussian matrices are near isometries. Therefore, we can carry our results over to the estimation of the mean vector $\mathbf{X}\boldsymbol{\beta}$.



(a) Histogram of FDP.

(b) Histogram of V.

FIG. 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* $\boldsymbol{\lambda} = 1.1\boldsymbol{\lambda}^{\mathrm{BH}}(0.1)$.

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

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

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

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

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

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 notation.* 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 supplementary materials [50].

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 $\mathbf{a}$, define the support of $\mathbf{a}$ as $\mathrm{supp}(\mathbf{a}) \triangleq \{i : a_i \neq 0\}$. 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 $0 < m < p$,

$\boldsymbol{\lambda}^{[m]} \triangleq (\lambda_1, \dots, \lambda_m)$ and $\boldsymbol{\lambda}^{-[m]} \triangleq (\lambda_{m+1}, \dots, \lambda_p)$. We write $\boldsymbol{\lambda}_\varepsilon$ (the superscript is omitted to save space) for the ε-inflated BHq critical values,

$$\lambda_{\varepsilon,i} = (1+\varepsilon)\lambda_i^{\mathrm{BH}} = (1+\varepsilon)\Phi^{-1}\bigl(1 - iq/(2p)\bigr).$$

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,

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

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 even under the global null and, consequently, the risk when estimating sparse signals would be high. This phenomenon can already be seen in the orthogonal case [34, 40]. This means that if we choose λ in a nonadaptive fashion then we would need to select $\lambda \geq \sigma\sqrt{2\log p}$. 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

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

The proof is in the supplementary materials [50]. Hence, the risk inflation does not decreases as the sparsity level k increases, whereas it does for SLOPE. Note that when $p = n$ and $k = p^{1-\delta}$,

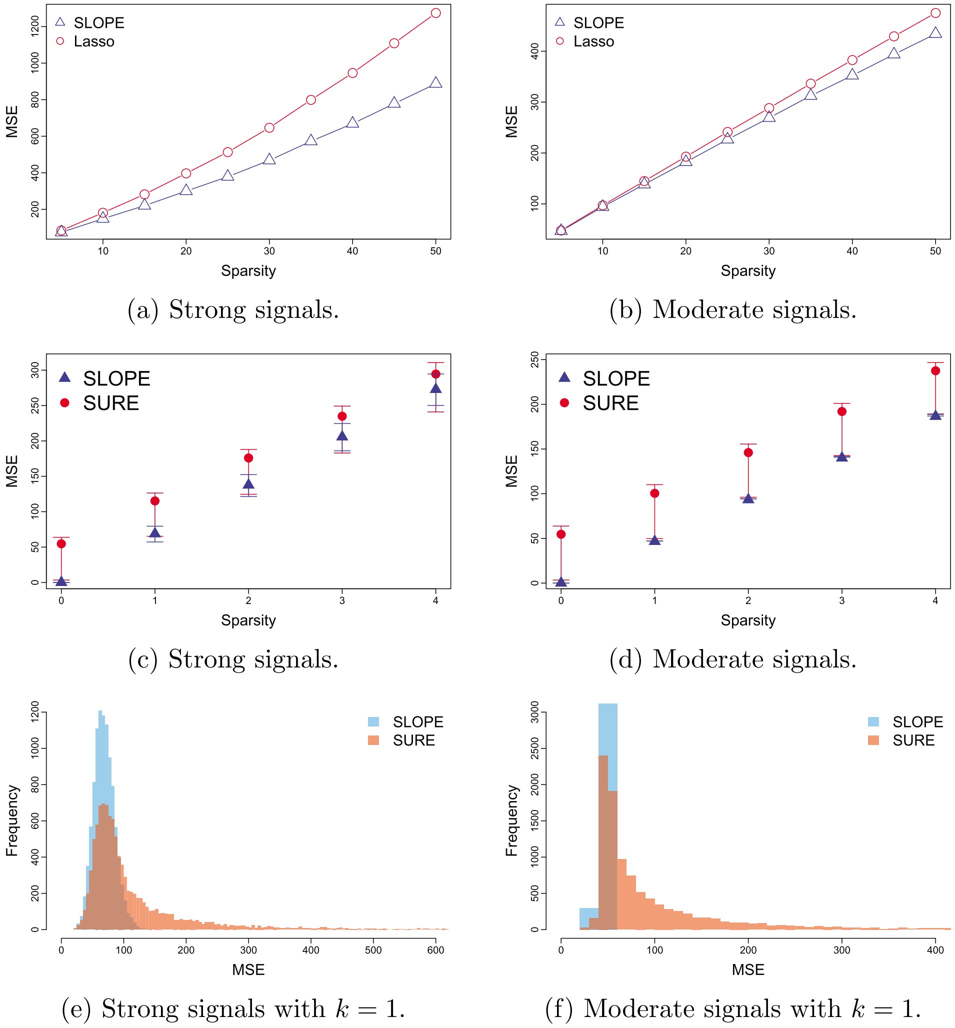
$$\frac{2\sigma^2 k \log p}{2\sigma^2 k \log(p/k)} \to \frac{1}{\delta}.$$

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

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

in the worst case over k.

Figure 3(a) and 3(b) compare SLOPE with Lasso estimates for both strong and moderate signals. SLOPE is more accurate than the Lasso in both cases, and the



(a) Strong signals.

(b) Moderate signals.

(c) Strong signals.

(d) Moderate signals.

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

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

FIG. 3. (a) *and* (b) *compares between SLOPE and Lasso under 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* $0.8\lambda_1^{\mathrm{BH}}$. *Next*, (c), (d), (e), (f) *compare SLOPE with SURE under orthogonal design. Empirical distributions of* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ *is obtained from* 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* (c) *and* (d), *the bars represent* 75% *and* 25% *percentiles.*

comparative advantage increases as k gets larger. This is consistent with the reasoning that SLOPE has a lower bias when k gets larger.

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 [43] 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 [36] are likely to be far too conservative. In a different direction, it would be interesting to compare SLOPE with the Lasso in different settings, where perhaps both k/p and n/p converge to positive constants. While some tools have been developed for the Lasso in this asymptotic regime [7], it is unclear how SLOPE would behave and even what a good sequence of weights $\{\lambda_i\}$ might be in this case.

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 $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 \mathbf{I}_p)$. 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 [26] for estimating a vector of means from $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 \mathbf{I}_p)$ is a cross-validation type procedure in the sense that the thresholding parameter is selected to minimize Stein's unbiased estimate of risk (SURE) [49]. 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 SURE(λ). It has been observed [20, 26] that SURE thresholding loses performance in cases of sparse signals $\boldsymbol{\beta}$, an empirical phenomenon which can perhaps be made theoretically precise. Indeed, our own work in progress aims to show that for any fixed sparsity k, SURE thresholding obeys

$$\frac{\sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{E}\|\widehat{\boldsymbol{\beta}}_{\mathrm{SURE}} - \boldsymbol{\beta}\|^2}{\sup_{\|\boldsymbol{\beta}\|_0 \le k} \mathbb{E}\|\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}} - \boldsymbol{\beta}\|^2} \ge \bigl(1 + o(1)\bigr)\frac{k+1}{k} > 1,$$

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

Figure 3 compares SLOPE with SURE in estimation error. In Figure 3(c) and 3(d), 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. Figure 3(e) and 3(f) 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.* As brought up earlier, the paper [39] suggests a variation on FDR thresholding, where an adaptive smooth-thresholding rule is applied instead of a hard one. Such a procedure is still intrinsically limited to sequence models, and cannot be generalized to linear regression. On this subject, consider the *sequential FDR thresholding* rule,

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

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 $\sigma\lambda^{\mathrm{BH}}_i$ to the ith largest observation (in magnitude). Under the same assumptions as in Theorem 1.1, this estimator also obeys

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

The proof is in [50] 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 $\mathbf{y}$, 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.

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 $\mathbf{a} \in \mathbb{R}^p$ is said to majorize $\mathbf{b} \in \mathbb{R}^p$ if for all $i = 1, \dots, p$,

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

This differs from a more standard definition (e.g., see [44]) where the last inequality with $i = p$ is replaced by an equality (and absolute values are omitted). We see that if $\mathbf{a}$ majorizes $\mathbf{b}$ and $\mathbf{c}$ majorizes $\mathbf{d}$, then the concatenated vector $(\mathbf{a}, \mathbf{c})$ majorizes $(\mathbf{b}, \mathbf{d})$. For convenience, we list below some basic but nontrivial properties of majorization and of the prox to the sorted ℓ_1 norm as defined in (1.7). All the proofs are deferred to the supplementary material [50].

FACT 3.1. If $\mathbf{a}$ majorizes $\mathbf{b}$, then $\|\mathbf{a}\| \geq \|\mathbf{b}\|$.

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

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

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

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

LEMMA 3.1. *For any* $\mathbf{a}$, *it holds that*

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

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

PROOF. The firm nonexpansiveness (e.g., see page 131 of [46]) of the prox reads

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

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

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

where we use the nonnegativity of $\operatorname{sgn}(\mathbf{a}) \odot \operatorname{prox}_{\boldsymbol{\lambda}}(\mathbf{a})$ 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

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

that is, 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* $\boldsymbol{\beta}$,

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

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

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

This gives

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

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

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

We proceed by showing that for $\boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p-k})$, $\mathbb{E}\|\mathrm{prox}_{\boldsymbol{\lambda}^{-[k]}}(\boldsymbol{\zeta})\|^2 = o(1)2k \times \log(p/k)$. To do this, pick $A > 0$ sufficiently large such that $q(1 + 1/A) < 1$ in Lemmas A.3 and A.4 of [50], which then give

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

The conclusion follows from Lemma 3.1 since

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

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

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

Here, the convergence is uniform over ε.

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

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

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

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

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 completes the proof. □

4. Gaussian random designs. When moving from an orthogonal to a nonorthogonal design, the correlations between the columns of **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 [32], the Dantzig selector [21] and MC+ [58]; see, for example, [7, 11, 22, 23, 27, 37, 45, 53, 55, 57, 59] 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 far longer 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, $\boldsymbol{\lambda} = (1+\varepsilon)\boldsymbol{\lambda}^{\mathrm{BH}}(q)$.

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

$$\mathbf{b}^{(m+1)} = \operatorname{prox}_{t_m\boldsymbol{\lambda}}\bigl(\mathbf{b}^{(m)} - t_m\mathbf{X}'(\mathbf{X}\mathbf{b}^{(m)} - \mathbf{y})\bigr),$$

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

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

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*

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

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 $\boldsymbol{\beta}$; that is, $T \supset \operatorname{supp}(\widehat{\boldsymbol{\beta}}) \cup \operatorname{supp}(\widetilde{\boldsymbol{\beta}}) \cup \operatorname{supp}(\boldsymbol{\beta})$. Suppose all the eigenvalues of $\mathbf{X}_T'\mathbf{X}_T$ 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 $|\mathbf{X}_i'\mathbf{z}|$ 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

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

One can view the resolvent solution as a sophisticated type of a dual certificate method, better known as primal-dual witness method [22, 48, 55] 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, that is, 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 in [50]) provides a well-known deviation bound about chi-square random variables, namely,

$$\mathbb{P}\bigl(\|\mathbf{z}\| \le (1+\varepsilon)\sqrt{n}\bigr) \ge 1 - \mathrm{e}^{-\varepsilon^2 n/2} \to 1$$

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

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

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

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

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

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

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

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

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

4.3. *$\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ are close when* $\mathbf{X}$ *is nearly orthogonal.* We prove Lemma 4.2 in the case where $T = \{1, \ldots, p\}$, first. Set $J_{\boldsymbol{\lambda}}(\mathbf{b}) = \sum_{1 \le i \le p} \lambda_i |b|_{(i)}$, by definition $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ respectively, minimize

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

Next, the assumptions about the eigenvalues of $\mathbf{X}'\mathbf{X}$ 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

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

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

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

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

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

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

and

$$\underset{\mathbf{b} \in \mathbb{R}^{|T|}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{\beta}_T + \mathbf{X}_T'\mathbf{z} - \mathbf{b}\right\|^2 + J_{\boldsymbol{\lambda}^{[m]}}(\mathbf{b}).$$

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

$$L_1(\mathbf{b}) := \tfrac{1}{2}\left\|\mathbf{X}_T(\boldsymbol{\beta}_T - \mathbf{b})\right\|^2 + \mathbf{z}'\mathbf{X}_T(\boldsymbol{\beta}_T - \mathbf{b}) + J_{\boldsymbol{\lambda}^{[m]}}(\mathbf{b}),$$
$$L_2(\mathbf{b}) := \tfrac{1}{2}\left\|\boldsymbol{\beta}_T - \mathbf{b}\right\|^2 + \mathbf{z}'\mathbf{X}_T(\boldsymbol{\beta}_T - \mathbf{b}) + J_{\boldsymbol{\lambda}^{[m]}}(\mathbf{b}).$$

From now on, the proof is just as before.

4.4. *Support localization.* Below we write $\mathbf{a} \preceq \mathbf{b}$ as a short-hand for $\mathbf{b}$ majorizes $\mathbf{a}$ 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 $\widehat{\mathbf{b}}_T$ be the solution to the reduced SLOPE problem* (4.6), *which only fits regression coefficients with indices in T. If*

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

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

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 $\|\mathbf{X}_{\overline{T}}'(\mathbf{y} - \mathbf{X}_T\widehat{\mathbf{b}}_T)\|_\infty \le \lambda$. In this case, it is well known that this implies that a solution to the Lasso is supported on T; see, for example, [22, 48, 55].

The main result of this section is this.

LEMMA 4.4. *Suppose*

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

for an arbitrary small constant $c > 0$, *where* d *is a deterministic sequence diverging to infinity*[6] *in such a way that* $k^\star/p \to 0$ *and* $(k^\star \log p)/n \to 0$. *Then*

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

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

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

which would follow from

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

and

$$\mathbf{X}'_{\overline{S^\star}}\mathbf{z} \preceq (1+\varepsilon/2)(\lambda^{\mathrm{BH}}_{k^\star+1}, \dots, \lambda^{\mathrm{BH}}_p). \tag{4.10}$$

Lemma A.12 in the supplementary material [50] concludes that (4.9) holds with probability tending to one, since, by assumption, $\varepsilon > 0$ 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 in [50]. Conditional on $\mathbf{z}$, $\mathbf{X}'_{\overline{S}}\mathbf{z} \sim \mathcal{N}(0, \|\mathbf{z}\|^2/n \cdot \mathbf{I}_{p-k})$. By definition, $\mathbf{X}'_{\overline{S^\star}}\mathbf{z}$ is formed from $\mathbf{X}'_{\overline{S}}\mathbf{z}$ by removing its $k^\star - k$ largest entries in absolute value. Denoting by $\zeta_1, \dots, \zeta_{p-k}$ i.i.d. standard Gaussian random variables, (4.10) thus boils down to

$$\bigl(|\zeta|_{(k^\star-k+1)}, |\zeta|_{(k^\star-k+2)}, \dots, |\zeta|_{(p-k)}\bigr) \preceq \frac{(1+\varepsilon/2)\sqrt{n}}{\|\mathbf{z}\|}(\lambda^{\mathrm{BH}}_{k^\star+1}, \dots, \lambda^{\mathrm{BH}}_p). \tag{4.11}$$

Borell's inequality (Lemma A.5 [50]) gives

$$\mathbb{P}\bigl((1+\varepsilon/2)\sqrt{n}/\|\mathbf{z}\| < 1\bigr) = \mathbb{P}\bigl(\|\mathbf{z}\| - \sqrt{n} > \varepsilon\sqrt{n}/2\bigr) \le \mathrm{e}^{-n\varepsilon^2/8} \to 0.$$

The conclusion follows from Lemma A.9 in [50].

We turn to (ii) and note that

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

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

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

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

In summary, $\mathbf{X}'_{\overline{S^\star}}\mathbf{z} \preceq \boldsymbol{\lambda}_\varepsilon^{-[k^\star]} \Rightarrow \operatorname{supp}(\widetilde{\boldsymbol{\beta}}) \subset S^\star$. This completes the proof. □

4.5. *Proof of Theorem* 1.2 *and Corollary* 1.4. Put

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

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

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

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

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

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

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

which diverges to ∞, and

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

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 [50] 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 $\varepsilon^2 n \to \infty$.

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,

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

On $\mathscr{A}_2 \cap \mathscr{A}_3$, all the eigenvalues values of $\mathbf{X}'_{S^\diamond}\mathbf{X}_{S^\diamond}$ 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

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

and similarly,

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

This completes 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 the supplementary materials [50]. 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 $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\beta}, \mathbf{I}_p)$. 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 [50].

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 $\{1, \ldots, p\}$, and sets $\beta_i = \tau_p$ and $\beta_j = 0$ for $j \ne i$. For each $\varepsilon > 0$,*

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

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 $\varepsilon > 0$, we have*

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

PROOF. The fundamental duality between "min max" and "max min" gives

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

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

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

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

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

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,

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

On the other,

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

All in all, this gives

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

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

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

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

5.2. *Random designs.* We return to the regression setup $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}_p)$, where $\mathbf{X}$ 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 $\varepsilon > 0$,

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

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

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

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

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 $\mathbf{X}$. Thus, write the linear model as

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

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

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

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

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

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 $\mathbf{\Sigma}$, 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. Su would like to thank Małgorzata Bogdan and Iain Johnstone for helpful discussions. We thank Yuxin Chen, Rina Barber and Chiara Sabatti for their helpful comments about an early version of the manuscript. We would also like to thank the anonymous Associate Editor and two reviewers for many constructive comments.

SUPPLEMENTARY MATERIAL

Supplement to "SLOPE is adaptive to unknown sparsity and asymptotically minimax" (DOI: 10.1214/15-AOS1397SUPP; .pdf). The supplementary materials contain proofs of some technical results in this paper.

REFERENCES

[1] Abramovich, F. and Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Comput. Statist. Data Anal.* **22** 351–361. MR1411575

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

[3] Adke, S. R., Waikar, V. B. and Schuurmann, F. J. (1987). A two-stage shrinkage testimator for the mean of an exponential distribution. *Comm. Statist. Theory Methods* **16** 1821–1834. MR0903358

[4] Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Statist.* **6** 127–146 (electronic). MR1918295

[5] Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876

[6] Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions. the Theory and Application of Isotonic Regression*. Wiley, New York. MR0326887

[7] BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inform. Theory* **58** 1997–2017. MR2951312

[8] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

[9] BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

[10] BICKEL, P. J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.* **9** 1301–1309. MR0630112

[11] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[12] BIRGÉ, L. (2004). Model selection for Gaussian regression with random design. *Bernoulli* **10** 1039–1051. MR2108042

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

[14] BLAND, J. M. and ALTMAN, D. G. (1995). Multiple significance tests: The Bonferroni method. *The British Medical Journal* **310** 170.

[15] BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J. (2015). SLOPE—Adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140. MR3418717

[16] BOGDAN, M., VAN DEN BERG, E., SU, W. and CANDÈS, E. J. (2013). Statistical estimation and testing via the sorted ℓ_1 norm. Preprint. Available at arXiv:1310.1969.

[17] BONDELL, H. D. and REICH, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64** 115–123, 322–323. MR2422825

[18] BROWN, L. D., CAI, T. T., LOW, M. G. and ZHANG, C.-H. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.* **30** 688–707. MR1922538

[19] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101

[20] CAI, T. T. and ZHOU, H. H. (2009). A data-driven block thresholding approach to wavelet estimation. *Ann. Statist.* **37** 569–595. MR2502643

[21] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[22] CANDÈS, E. J. and PLAN, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.* **37** 2145–2177. MR2543688

[23] CANDÈS, E. J., ROMBERG, J. K. and TAO, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* **59** 1207–1223. MR2230846

[24] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

[25] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields* **99** 277–303. MR1278886

[26] DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464

[27] DONOHO, D. L., JOHNSTONE, I. M., MALEKI, A. and MONTANARI, A. (2011). Compressed sensing over ℓ_p-balls: Minimax mean square error. In *Proceedings of the IEEE International Symposium on Information Theory* 129–133. IEEE, New York.

[28] DONOHO, D. L. and MONTANARI, A. (2013). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. Preprint. Available at arXiv:1310.7320.

[29] DONOHO, D. L. and TANNER, J. (2009). Counting faces of randomly projected polytopes when the projection radically lowers dimension. *J. Amer. Math. Soc.* **22** 1–53. MR2449053

[30] DONOHO, D. L. and TANNER, J. (2010). Exponential bounds implying construction of compressed sensing matrices, error-correcting codes, and neighborly polytopes by random sampling. *IEEE Trans. Inform. Theory* **56** 2002–2016. MR2654490

[31] FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1035. MR3010887

[32] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[33] FIGUEIREDO, M. and NOWAK, R. (2014). Sparse estimation with strongly correlated variables using ordered weighted ℓ_1 regularization. Preprint. Available at arXiv:1409.4005.

[34] FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975. MR1329177

[35] FOSTER, D. P. and STINE, R. A. (1999). Local asymptotic coding and the minimum description length. *IEEE Trans. Inform. Theory* **45** 1289–1293. MR1686271

[36] G'SELL, M., WAGER, S., CHOULDECHOVA, A. and TIBSHIRANI, R. (2013). Sequential selection procedures and false discovery rate control. Preprint. Available at arXiv:1309.5352.

[37] GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039

[38] JI, P. and ZHAO, Z. (2014). Rate optimal multiple testing procedure in high-dimensional regression. Preprint. Available at arXiv:1404.2961.

[39] JIANG, W. and ZHANG, C.-H. (2013). A nonparametric empirical Bayes approach to adaptive minimax estimation. *J. Multivariate Anal.* **122** 82–95. MR3189309

[40] JOHNSTONE, I. M. (2013). Gaussian estimation: Sequence and wavelet models. Available at http://statweb.stanford.edu/~imj/GE06-11-13.pdf.

[41] KRUSKAL, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **29** 115–129. MR0169713

[42] LIU, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.* **41** 2948–2978. MR3161453

[43] LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. MR3210970

[44] MARSHALL, A. W., OLKIN, I. and ARNOLD, B. C. (2011). *Inequalities*: *Theory of Majorization and Its Applications*, 2nd ed. Springer, New York. MR2759813

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

[46] PARIKH, N. and BOYD, S. (2013). Proximal algorithms. *Foundations and Trends in Optimization* **1** 123–231.

[47] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274

[48] RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343

[49] STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

[50] SU, W. and CANDÈS, E. (2015). Supplement to "SLOPE is adaptive to unknown sparsity and asymptotically minimax." DOI:10.1214/15-AOS1397SUPP.

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

[52] TIBSHIRANI, R. and KNIGHT, K. (1999). The covariance inflation criterion for adaptive model selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 529–546. MR1707859

[53] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316

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

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

[56] WU, Z. and ZHOU, H. H. (2013). Model selection and sharp asymptotic minimaxity. *Probab. Theory Related Fields* **156** 165–191. MR3055256

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

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

[59] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

DEPARTMENTS OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: wjsu@stanford.edu

DEPARTMENTS OF STATISTICS
AND MATHEMATICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: candes@stanford.edu