Asymptotics and Optimal Designs of SLOPE for Sparse Linear Regression

Hong Hu and Yue M. Lu
John A. Paulson School of Engineering and Applied Sciences
Harvard University, Cambridge, MA 02138, USA

***Abstract*—In sparse linear regression, the SLOPE estimator generalizes LASSO by assigning magnitude-dependent regularizations to different coordinates of the estimate. In this paper, we present an asymptotically exact characterization of the performance of SLOPE in the high-dimensional regime where the number of unknown parameters grows in proportion to the number of observations. Our asymptotic characterization enables us to derive optimal regularization sequences to either minimize the MSE or to maximize the power in variable selection under any given level of Type-I error. In both cases, we show that the optimal design can be recast as certain infinite-dimensional convex optimization problems, which have efficient and accurate finite-dimensional approximations. Numerical simulations verify our asymptotic predictions. They also demonstrate the superiority of our optimal design over LASSO and a regularization sequence previously proposed in the literature.**

I. Introduction

In sparse linear regression, we seek to estimate a sparse vector $\boldsymbol{\beta} \in \mathbb{R}^p$ from

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{w}, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ is the design matrix and $\boldsymbol{w}$ denotes the observation noise. In this paper, we study the *sorted* ℓ_1 *penalization estimator* (SLOPE) [1] (see also [2], [3]). Given a non-decreasing regularization sequence $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_p]^\top$ with $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_p$, SLOPE estimates $\boldsymbol{\beta}$ by solving the following optimization problem

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{x}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \sum_{i=1}^{p} \lambda_i |x|_{(i)}, \tag{2}$$

where $|x|_{(1)} \le |x|_{(2)} \le \cdots \le |x|_{(p)}$ is a reordering of the absolute values $|x_1|, |x_2|, \ldots, |x_p|$ in increasing order. In [1], the regularization term $J_{\boldsymbol{\lambda}}(\boldsymbol{x}) \overset{\text{def}}{=} \sum_{i=1}^{p} \lambda_i |x|_{(i)}$ is referred to as the "sorted ℓ_1 norm" of $\boldsymbol{x}$. The same regularizer was independently developed in a different line of work [2]–[4], where the motivation is to promote group selection in the presence of correlated covariates.

The classical LASSO estimator is a special case of SLOPE. It corresponds to using a constant regularization sequence, *i.e.*, $\lambda_1 = \lambda_2 = \cdots = \lambda_p = \lambda$. However, with more general λ-sequences, SLOPE has the flexibility to penalize different coordinates of the estimate according to their magnitudes. This adaptivity endows SLOPE with some nice *statistical* properties that are not possessed by LASSO. For example, it is shown in [5], [6] that SLOPE achieves the minimax ℓ_2 estimation rate with high probability. In terms of testing, the authors of [1] show that SLOPE controls false discovery rate (FDR) for orthogonal design matrices, which is not the case in LASSO [7]. In addition, we note that the new regularizer $J_{\boldsymbol{\lambda}}(\boldsymbol{x})$ is still a norm [1], [3]. Thus, the optimization problem associated with SLOPE remains convex, and it can be efficiently solved by using *e.g.*, proximal gradient descent [1], [3].

In the aforementioned studies on analyzing SLOPE, the performance of the estimator is given in terms of non-asymptotic probabilistic bounds. Such bounds provide very limited information about how to optimally design the regularization sequence $\boldsymbol{\lambda}$ in different settings, an important open question in the literature. In this paper, we provide two main contributions:

1) We obtain a characterization of SLOPE in the asymptotic regime: $n, p \to \infty$ and $n/p \to \delta$. Compared with the probabilistic bounds derived in previous work, our results are asymptotically *exact*. Similar asymptotic analysis has been done for LASSO [8] and many other regularized linear regression problems [9]–[11], but the main technical challenge in analyzing SLOPE is the nonseparability of $J_{\boldsymbol{\lambda}}(\boldsymbol{x})$: it cannot be written as a sum of component-wise functions, *i.e.*, $J_{\boldsymbol{\lambda}}(\boldsymbol{x}) \neq \sum_{i=1}^{p} J_i(x_i)$. In our work, we overcome this challenge by showing that the proximal operator of $J_{\boldsymbol{\lambda}}(\boldsymbol{x})$ is *asymptotically separable*.
2) Using our asymptotic characterization, we derive *oracle* optimal $\boldsymbol{\lambda}$ in two settings: (1) the optimal regularization sequence that minimizes the MSE $\mathbb{E}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$; and (2) the optimal sequence that achieves the highest possible power in testing and variable selection under a given level of Type-I error. In both cases, we show that the optimal design can be recast as certain infinite-dimensional convex optimization problems, which have efficient and accurate finite-dimensional approximations.

A caveat of our optimal design is that it requires knowing the limiting empirical measure of $\boldsymbol{\beta}$ (*e.g.*, the sparsity level and the distribution of its nonzero coefficients). For this reason, our results are *oracle optimal*. It provides the first step towards more practical optimal designs that are completely blind to $\boldsymbol{\beta}$.

The rest of the paper is organized as follows. In Sec. II, we first prove the asymptotic separability of the proximal operator associated with $J_{\boldsymbol{\lambda}}(\boldsymbol{x})$. This property allows us to derive our main asymptotic characterization of SLOPE, summarized as Theorem 1. Based on this analysis, we present the optimal design of the regularization sequence in Sec. III. Numerical simulations verify our asymptotic characterizations. They also

demonstrate the superiority of our optimal design over LASSO and a previous sequence design in the literature [5]. Due to space constraints, we only state and illustrate the main results in this paper, and leave the technical proofs to [12].

II. Main Asymptotic Results

A. Technical Assumptions

There are four main objects in the description of our model and algorithm: (1) the unknown sparse vector $\boldsymbol{\beta}$; (2) the design matrix $\boldsymbol{A}$; (3) the noise vector $\boldsymbol{w}$; and (4) the regularization sequence $\boldsymbol{\lambda}$. Since we study the asymptotic limit (with $p \to \infty$), we will consider a sequence of instances $\{\boldsymbol{\beta}^{(p)}, \boldsymbol{A}^{(p)}, \boldsymbol{w}^{(p)}, \boldsymbol{\lambda}^{(p)}\}_{p\in\mathbb{N}}$ with increasing dimensions p, where $\boldsymbol{\beta}^{(p)}, \boldsymbol{\lambda}^{(p)} \in \mathbb{R}^p$, $\boldsymbol{A}^{(p)} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{w}^{(p)} \in \mathbb{R}^n$. A sequence of vectors $\boldsymbol{x}^{(p)} \in \mathbb{R}^p$ indexed by p is called a *converging sequence* [8] if its empirical measure $\mu_p(x) \overset{\text{def}}{=} \frac{1}{p}\sum_{i=1}^{p} \delta(x - x_i^{(p)})$ converges weakly to a probability measure on $\mathbb{R}$.

Our results are proved under the following assumptions:

(A.1) The number of observations grows in proportion to p: $n^{(p)}/p \to \delta \in (0, \infty)$.

(A.2) The number of nonzero elements in $\boldsymbol{\beta}^{(p)}$ grows in proportion to p: $k/p \to \rho \in (0, 1]$.

(A.3) The elements of $\boldsymbol{A}^{(p)}$ are i.i.d. Gaussian distribution: $A_{ij}^{(p)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{1}{n})$.

(A.4) $\{\boldsymbol{\beta}^{(p)}\}_p$, $\{\boldsymbol{w}^{(p)}\}_p$ and $\{\boldsymbol{\lambda}^{(p)}\}_p$ are converging sequences. The distribution functions of the limiting measures are denoted by F_β, F_w and F_λ, respectively. Moreover, we have $\mathbb{P}(|\lambda| \neq 0) > 0$, $\frac{1}{p}\|\boldsymbol{\beta}^{(p)}\|^2 \to \mathbb{E}[\beta^2]$, $\frac{1}{n}\|\boldsymbol{w}^{(p)}\|^2 \to \mathbb{E}[w^2] = \sigma_w^2$ and $\frac{1}{p}\|\boldsymbol{\lambda}^{(p)}\|^2 \to \mathbb{E}[\lambda^2]$, where the probability $\mathbb{P}(\cdot)$ and the expectations $\mathbb{E}[\cdot]$ are all computed with respect to the limiting measures.

B. Asymptotics of the Proximal Operator of $J_{\boldsymbol{\lambda}}(\boldsymbol{x})$

We start by studying the proximal operator associated with the sorted ℓ_1 norm $J_{\boldsymbol{\lambda}}(\boldsymbol{x})$. Given $\boldsymbol{y} \in \mathbb{R}^p$ and a regularization sequence $\boldsymbol{\lambda}$ with $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_p$, the proximal operator is defined as the solution to the following optimization problem:

$$\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y}) \overset{\text{def}}{=} \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2 + \sum_{i=1}^{p} \lambda_i |x|_{(i)}, \qquad (3)$$

where $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_p$.

In the case of LASSO, which corresponds to choosing $\lambda_1 = \lambda_2 = \cdots = \lambda_p = \lambda$, the proximal operator is easy to characterize, as it is separable: $[\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i = \text{sign}(y_i)\max(|y_i| - \lambda, 0)$. In other words, the ith element of $\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})$ is solely determined by y_i. However, this separability property does not hold for a general regularization sequence. When p is finite, $[\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i$ depends not only on y_i but also on other elements of $\boldsymbol{y}$. This coupling makes it much harder to analyze the proximal operator. Fortunately, as we show below, when $p \to \infty$, $\text{Prox}_{\boldsymbol{\lambda}}(\cdot)$ becomes *asymptotically separable*.

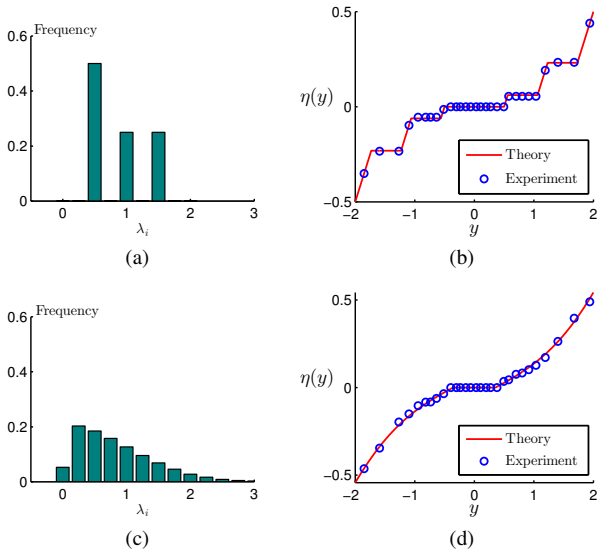


Figure 1: (a) and (c): The histograms of two different $\boldsymbol{\lambda}$-sequences. (b) and (d): Sample points of $(y_i, [\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i)$ (the blue dots) compared against the limiting scalar functions (the red curves). In this experiment, $p = 1024$.

Proposition 1: Let $\{\boldsymbol{y}^{(p)}\}_p$ and $\{\boldsymbol{\lambda}^{(p)}\}_p$ be two converging sequences. Denote by F_y and F_λ the distribution functions of their respective limiting measures. It holds that

$$\lim_{p\to\infty} \frac{1}{p}\|\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y}^{(p)}) - \eta(\boldsymbol{y}^{(p)}; F_y, F_\lambda)\|^2 \to 0, \qquad (4)$$

where $\eta(\cdot; F_y, F_\lambda)$ is a scalar function that is determined by F_y and F_λ, and $\eta(\boldsymbol{y}^{(p)}; F_y, F_\lambda)$ denotes a coordinate-wise application of $\eta(\cdot; F_y, F_\lambda)$ on $\boldsymbol{y}^{(p)}$.

The asymptotic separability of $\text{Prox}_{\boldsymbol{\lambda}}(\cdot)$ greatly facilitates our asymptotic analysis and design of SLOPE, since it allows us to reduce the original high-dimensional problem to an equivalent one-dimensional problem. In what follows, we refer to $\eta(\cdot; F_y, F_\lambda)$ as the *limiting scalar function*. In the appendix, we briefly present a procedure for constructing $\eta(\cdot; F_y, F_\lambda)$ from the limiting distribution functions F_y and F_λ. More details can be found in [12].

Example 1: We compare the proximal operator $\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})$ and the limiting scalar function $\eta(y; F_y, F_\lambda)$, for two different $\boldsymbol{\lambda}$-sequences shown in Fig. 1(a) and Fig. 1(c). The red curves represent the limiting scalar functions obtained in Proposition 1, whereas the blue circles are sample points of $(y_i, [\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i)$, with $y_i \sim \mathcal{N}(0, 1)$. For better visualization, we randomly sample 3% of all $(y_i, [\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i)$. It can be seen that under a moderate dimension $p = 1024$, the proximal operator can already be very accurately approximated by the limiting scalar function.

C. Asymptotics of SLOPE

We are now ready to tackle the original optimization problem (2) associated with SLOPE. Our goal is to characterize the joint empirical measure of $(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$: $\mu_p(\hat{\beta}, \beta) \overset{\text{def}}{=} \frac{1}{p}\sum_{i=1}^{p} \delta(\hat{\beta} - \hat{\beta}_i, \beta - \beta_i)$. Indeed, many quantities of interest, such as the

MSE, type-I error, and FDR, are all functionals of this joint empirical measure. A function $\psi : \mathbb{R}^2 \to \mathbb{R}$ is called *pseudo-Lipschiz* if $|\psi(x) - \psi(y)| \le L(1 + \|x\|_2 + \|y\|_2)\|x - y\|_2$ for all $x, y \in \mathbb{R}^2$, where L is a positive constant. As in [8], we will depict the limit of $\mu_p(\hat{\beta}, \beta)$ through its action on pseudo-Lipschiz functions.

Theorem 1: Assume (A.1) – (A.4) hold. For any pseudo-Lipschiz function ψ, we have

$$\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \psi(\hat{\beta}_i,\, \beta_i) = \mathbb{E}_{B,Z}[\psi(\eta(B + \sigma Z;\, F_y, F_{\tau\lambda}), B)]. \tag{5}$$

Here, B, Z are two independent random variables with $B \sim F_\beta$ and $Z \sim \mathcal{N}(0, 1)$; $\eta(\cdot\,; F_y, F_{\tau\lambda})$ is the limiting scalar function defined in Proposition 1, with F_y denoting the distribution function of $B + \sigma Z$ and $F_{\tau\lambda}$ denoting that of $\tau\lambda$ for some $\tau \ge 0$. Moreover, the scalar pair $(\sigma,\, \tau)$ is the unique solution of the following equations:

$$\sigma^2 = \sigma_w^2 + \frac{1}{\delta}\mathbb{E}_{B,Z}[(\eta(B + \sigma Z; F_y, F_{\tau\lambda}) - B)^2] \tag{6}$$

$$1 = \tau\big(1 - \frac{1}{\delta}\mathbb{E}_{B,Z}[\eta'(B + \sigma Z; F_y, F_{\tau\lambda})]\big). \tag{7}$$

Remark 1: Readers familiar with the asymptotic analysis of LASSO will recognize that the forms of (6) and (7) look identical to the results of LASSO obtained in [8], [11]. Indeed, the first part of our proof directly applies the framework of analyzing LASSO asymptotics using convex Gaussian min-max theorem (CMGT) [10], [11]. Following [11], in the asymptotic regime, the limiting measure of SLOPE is determined by the following fixed point equations:

$$\sigma^2 = \sigma_w^2 + \frac{1}{\delta}\lim_{p\to\infty}\frac{1}{p}\|\mathrm{Prox}_{\tau\boldsymbol{\lambda}}(\boldsymbol{\beta} + \sigma\boldsymbol{Z}) - \boldsymbol{\beta}\|_2^2 \tag{8}$$

$$1 = \tau\left[1 - \frac{1}{\delta}\lim_{p\to\infty}\frac{1}{p}\mathrm{div}(\mathrm{Prox}_{\tau\boldsymbol{\lambda}}(\boldsymbol{\beta} + \sigma\boldsymbol{Z}))\right]. \tag{9}$$

Note that (8) and (9) are similar to (6) and (7), except that they involve an $\mathbb{R}^p \mapsto \mathbb{R}^p$ proximal mapping: $\mathrm{Prox}_{\tau\boldsymbol{\lambda}}(\boldsymbol{\beta}+\sigma\boldsymbol{Z})$. This is where Proposition 1 becomes useful. Using the asymptotic separability stated in that proposition, we can simplify (8) and (9) to the scalar equations given in (6) and (7).

Theorem 1 essentially says that the joint empirical measure of $(\hat{\boldsymbol{\beta}},\, \boldsymbol{\beta})$ converges weakly to the law of $(\eta(B + \sigma Z;\, F_y, F_{\tau\lambda}), B)$. This means that although the original problem (2) is high-dimensional, its asymptotic performance can be succinctly captured by merely two scalars random variables. In (5), if we let $\psi(x, y) = (x - y)^2$, we obtain the asymptotic MSE; by setting $\psi(x, y) = 1_{y=0,\, x\neq 0}$, we can recover the type-I error. (Technically, $1_{y=0, x\neq 0}$ is not pseudo-Lipschiz. However, with additional justifications [1], one can show that the conclusion is still correct.)

III. Oracle Optimality of SLOPE

In this section, we will study the optimal design of the regularization sequence in SLOPE. Using the asymptotic characterization presented in Sec. II, we will derive the optimal limiting distribution F_λ to achieve the best estimation or testing performance, given the oracle knowledge of F_β.

A. Estimation with Minimum MSE

We first turn to the problem of finding the optimal $\boldsymbol{\lambda}$-sequence which minimizes the MSE of slope estimator. Since we work in the asymptotic regime, it boils down to finding the optimal distribution F_λ^* such that

$$\begin{aligned} F_\lambda^* &= \arg\min_{F_\lambda} \lim_{p\to\infty}\frac{1}{p}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \\ &= \arg\min_{F_\lambda} \mathbb{E}_{B,Z}[(\eta(B + \sigma Z; F_y, F_{\tau\lambda}) - B)^2], \end{aligned}$$

where $B \sim F_\beta$ and the second equality follows from Theorem 1. From (6), this is further equivalent to finding F_λ to minimize σ. However, directly searching over F_λ appears unwieldy, since σ as a functional of F_λ is defined indirectly through a nonlinear fixed point equation.

To simplify this problem, we first note that in (6) and (7), the influence of F_λ on the solution $(\sigma,\, \tau)$ is only through the limiting scalar function η. Therefore, instead of optimizing over F_λ, we can find the optimal η^* and then calculate the corresponding F_λ^*. The next question then becomes finding all possible η that can be realized by some F_λ. In fact, we can compute all possible $\eta(\cdot)$ associated with any converging sequence $\{\boldsymbol{y}^{(p)}\}_{p\in\mathbb{N}}$. Let

$$\mathcal{M} \overset{\text{def}}{=} \big\{\eta(\cdot\,; F_y, F_\lambda) \mid \exists F_\lambda,\, \mathbb{E}\lambda^2 < \infty, \text{ s.t. } (4) \text{ holds}\big\}$$

be the functional space that η belongs to. We have the following result:

Proposition 2: For any converging sequence $\{\boldsymbol{y}^{(p)}\}_{p\in\mathbb{N}}$, we have

$$\begin{aligned} \mathcal{M} =\{&\eta(y) \mid \eta(y) = -\eta(-y) \text{ and} \\ &0 \le \eta(y_1) - \eta(y_2) \le y_1 - y_2,\, \forall y_1 \ge y_2\} \end{aligned}$$

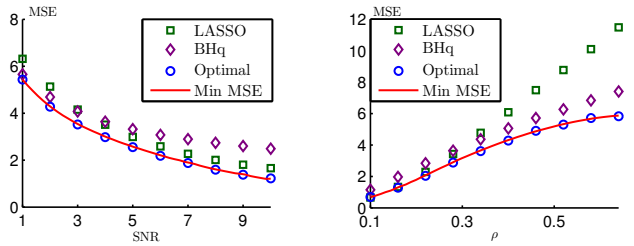
and $\mathcal{M}$ is a convex set. Moreover, for any $\eta \in \mathcal{M}$, the corresponding distribution of the $\boldsymbol{\lambda}$-sequence that yields η can be represented by: $\lambda \sim |Y| - \eta(|Y|)$, where Y follows the limiting distribution of $\{\boldsymbol{y}^{(p)}\}_{p\in\mathbb{N}}$.

Remark 2: Proposition 2 is the key ingredient in our optimal design. It shows that, with different choices of F_λ, we can reach any nondecreasing and odd function that is Lipschitz continuous with constant 1. Clearly, the soft-thresholding functions associated with LASSO belongs to $\mathcal{M}$, but the set $\mathcal{M}$ is much richer. This is the essence of how SLOPE generalizes LASSO: it allows for more degrees of freedom in the regularization.

Due to Proposition 2, the optimization problem can be simplified to that of finding the optimal $\eta \in \mathcal{M}$ such that σ as obtained from (6) and (7) is minimized. Specifically, we need to find

$$\sigma_{\min} \overset{\text{def}}{=} \inf\{\sigma \mid \exists \eta \in \mathcal{M}, \text{ s.t. } \sigma \text{ satisfies (6) and (7)}\}. \tag{10}$$

Note that equations (6) and (7) involve two variables σ and τ. It is not easy to handle them simultaneously. A simplification we can make is to first set τ to 1 and find the minimum σ such that the first equation (6) and the inequality $\mathbb{E}_{B,Z}[\eta'(B + \sigma Z; F_y, F_\lambda)] \le \delta$ hold. Once we



(a) $\rho = 0.256$ (b) SNR $= 5$

Figure 2: Comparison of MSEs obtained by three regularization sequences: LASSO, BHq and the oracle optimal design, under different SNR and sparsity levels. Here, $p = 1024$, $\delta = 0.64$. The red curves show the theoretical minimum MSE that can be achieved by using the oracle optimal sequences.

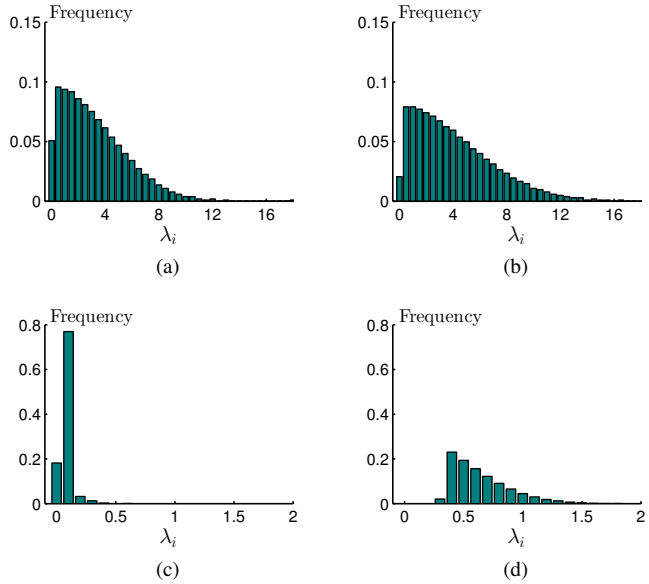


Figure 3: Comparison of distributions of two regularization sequences in Fig. 2a: (a)-(b): SNR $= 1$, (c)-(d): SNR $= 10$.

get $\sigma_{\min}$, the corresponding τ can then be obtained via (7): $\tau = (1 - \frac{1}{\delta}\mathbb{E}_{B,Z}[\eta'(B + \sigma_{\min}Z; F_y, F_{\tau\lambda})])^{-1}$ and λ is in turn updated to be λ/τ. After this replacement, (6) and (7) will be both satisfied. It is not difficult to show that this procedure will lead to the same $\sigma_{\min}$ as defined in (10). Therefore, the remaining task is to solve, for every candidate σ, the following problem:

$$\min_{\eta \in \mathcal{M}} \mathbb{E}_{B,Z}[(\eta(B + \sigma Z) - B)^2] \quad (11)$$
$$\text{s.t. } \mathbb{E}_{B,Z}[\eta'(B + \sigma Z)] \leq \delta.$$

Thanks to the convexity of $\mathcal{M}$, we can show that (11) is an infinite-dimensional convex optimization problem. In practice, we can discretize over $\mathbb{R}$ to solve a finite-dimensional approximation. If the minimum value of (11) is smaller than $\delta(\sigma^2 - \sigma_w^2)$, then it can be shown that $\sigma_{\min} < \sigma$ and vice versa. Clearly, we only need to search for $\sigma_{\min}$ over a compact set: $[\sigma_w, \sqrt{\sigma_w^2 + \frac{1}{\delta}\mathbb{E}[B^2]}$, since from (6), we know $\sigma^2 \geq \sigma_w^2$ and also $\sigma^2 = \sigma_w^2 + \frac{1}{\delta}\mathbb{E}[B^2]$ when $\eta \equiv 0$. As a result, we can do a binary search over $[\sigma_w, \sqrt{\sigma_w^2 + \frac{1}{\delta}\mathbb{E}[B^2]}$ to find the minimum σ. Once we find the optimal $\eta(y)$ and $(\sigma_{\min}, \tau)$, we know from Proposition 2 that the corresponding λ can be represented as: $\lambda \sim \frac{|Y| - \eta(|Y|)}{\tau}$, with $Y \sim B + \sigma_{\min}Z$.

In Fig. 2, we compare the MSEs achieved by our optimal design with those obtained by LASSO and the BHq sequences proposed [5], at different SNR and sparsity levels. For fair comparison, we optimize the parameters of the BHq and LASSO sequences. It can be seen from the figure that the empirical minimum MSEs match well with theoretical ones. We observe from Fig. 2a that, under low SNRs, the BHq sequence can lead to very similar performance as the oracle optimal design. However, at higher SNR levels, the optimal design outperforms the BHq sequence, while it gets close to LASSO. To unravel the underlying reason for this, we plot in Fig. 3 the distributions of the $\boldsymbol{\lambda}$-sequences associated with the optimal design and the BHq design, respectively. It turns out that, in the low SNR case, the optimal design and BHq have similar distributions; at higher SNRs, the distribution of the optimal design is close to a delta-like distribution similar to LASSO.

Note that for small sparsity-level ρ, LASSO can outperform BHq and achieve performance close to that of the optimal sequence, but it is prone to higher bias when ρ grows. From Fig. 2b, we can find that LASSO's performance degrades much faster than the other two as ρ increases, This is because LASSO's penalization is not adaptive to the underlying sparsity levels [5].

B. Multiple Testing with Maximum Power

Next we consider using SLOPE for variable selection. For a given level of type-I error, we want to find the optimal regularization sequence to achieve the highest possible power.

As we have shown in the last section, in the asymptotic region, this is equivalent to optimizing over $\eta \in \mathcal{M}$. Let $y_{\text{thresh}} = \sup_{y \geq 0}\{y \mid \eta(y) = 0\}$. It follows from Theorem 1 that, in order to ensure $\mathbb{P}_{\text{type-I}} = \alpha$, we need to have $\frac{y_{\text{thresh}}}{\sigma} = \Phi^{-1}(1 - \frac{\alpha}{2})$, where $\Phi(\cdot)$ is the CDF of the standard normal distribution. Similarly, we can compute the power of the test as $\mathbb{P}(|\frac{\beta}{\sigma} + Z| \geq \Phi^{-1}(1 - \frac{\alpha}{2}))$. It can be shown that for any fixed β, $\mathbb{P}(|\frac{\beta}{\sigma} + Z| \geq \Phi^{-1}(1 - \frac{\alpha}{2}))$ is a nonincreasing function of σ. Thus, under a given type-I error rate α, maximizing the power is equivalent to minimizing σ.

Similar to the procedure described in Sec. III-A, we can traverse through a bounded set of σ to find $\sigma_{\min}$. The difference here is that we need to enforce additional constraints that guarantee $P_{\text{type-I}} = \alpha$. We omit further details, which can be found in [12]. In Fig.4, we compare the FDR curve of the optimal design with that of the BHq sequence. We verify that the optimal design indeed dominates the BHq sequence and that the empirical FDR curve matches well with theoretical prediction.

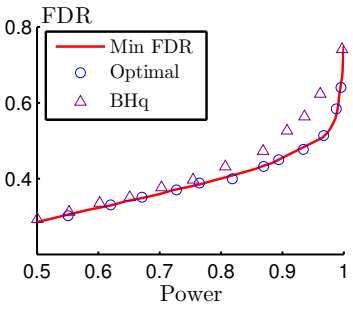


Figure 4: Hypothesis testing using oracle optimal and BHq sequences. Here, $p = 1024$, $\beta_i \overset{i.i.d.}{\sim} (1-\rho)\delta(0) + \rho\mathcal{N}(\mu_0, \sigma_0^2)$ with $\rho = 0.25$, $\mu_0 = 2.125$ and $\sigma_0 = 0$, $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.25$. The results are average over 100 realizations.

APPENDIX

In this appendix, we briefly describe the procedure for constructing the limiting scalar function $\eta(\cdot; F_y, F_\lambda)$ in Proposition 1. The procedure, as summarized in Algorithm 1, can be viewed as the asymptotic limit of a fast algorithm proposed in [1], [2] for solving (3). Here, $\lambda(y)$ can be intuitively understood as a function that assigns each y a regularization strength λ and $G(y)$ just thresholds y using the assigned λ. This is exactly in the same spirit of magnitude-dependent regularization applied in SLOPE. The `WHILE LOOP` in Step 2-14 is essentially an adjustment of $G(y)$ obtained in Step 1: within certain intervals $[y_L, y_R]$, the original $G(y)$ is replaced by $\mathbb{E}\left(G(y) \mid y \in [y_L, y_R]\right)$. The `WHILE LOOP` ends until $G(y)$ is nondecreasing.

For illustrations, we consider the simplest scenario when F_Y is continuous and when $G(y)$ as obtained in Step 1 is nondecreasing. In this case, `WHILE LOOP` in Step 2-14 will not be executed. It follows that

$$\eta(y; F_y, F_\lambda) = \text{sign}(y)\max\{0, |y| - F_\lambda^{-1}[F_{|y|}(|y|)]\}.$$

Clearly, this reduces to the soft-thresholding function associate with LASSO, with the regularization parameter given by $F_\lambda^{-1}[F_{|y|}(|y|)] \equiv \lambda$. Illustrations of several more complicated examples can be found in [12].

Algorithm 1 Limiting proximal mapping

Input: Distribution: $y \sim F_y(y)$, $\lambda \sim F_\lambda(\lambda)$

Output: Limiting proximal mapping $\eta(y; F_y, F_\lambda)$

1: Compute

$$G(y) := y - \lambda(y), y \geq 0$$

where

$$\lambda(y) := \begin{cases} F_\lambda^{-1}(F_{|y|}(y)) & \text{if } \mathbb{P}(|Y| = y) = 0, \\ \frac{\int_{F_{|y|}(y^-)}^{F_{|y|}(y)} F_\lambda^{-1}(u)du}{F_{|y|}(y) - F_{|y|}(y^-)} & \text{if } \mathbb{P}(|Y| = y) > 0. \end{cases}$$

2: **while** $\exists$ MDI [a] of $G(y)$ **do**
3: Find the first MDI of $G(y)$: $D_1 = [y_1, y_2]$.
4: **if** $\mathbb{P}(|Y| > y_2) = 0$ **then**
5: $y_L \leftarrow \arg\max_{v \in [0, y_1]} \mathbb{E}\left(G(y) | y \in [v, y_2]\right)$
6: $y_R \leftarrow y_2$
7: **else**
8: Find $I_1 = [y_2, y_3]$, neighbouring MNDI [b] of D_1.
9: Solve

$$\min_{w \in [0, y_1]} \max_{v \in [y_2, y_3]} \mathbb{E}\left(G(y) | y \in [v, w]\right)$$

10: Get optimal solution: $[w^*, v^*]$.
11: $y_L \leftarrow v^*$, $y_R \leftarrow w^*$
12: **end if**
13: For $y \in [y_L, y_R]$, replace original $G(y)$ by $\mathbb{E}\left(G(y) \mid y \in [y_L, y_R]\right)$ to get a new $G(y)$.
14: **end while**
15: Obtain: $\eta(y) = \text{sign}(y)\max\{0, G(|y|)\}$

[a] We call $[L, R]$ a maximal decreasing interval (MDI) of $G(y)$, if $G(y)$ is strictly decreasing on $[L, R]$ while not strictly decreasing on $[L-\varepsilon, R+\varepsilon]$ for any $\varepsilon > 0$.

[b] Maximal nondecreasing interval (MNDI) is defined in the same way as MDI.

REFERENCES

[1] M. Bogdan, E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès, "SLOPE-adaptive variable selection via convex optimization," *The annals of applied statistics*, vol. 9, no. 3, p. 1103, 2015.

[2] L. W. Zhong and J. T. Kwok, "Efficient sparse modeling with automatic feature grouping," *IEEE transactions on neural networks and learning systems*, vol. 23, no. 9, pp. 1436–1447, 2012.

[3] X. Zeng and M. A. Figueiredo, "Decreasing weighted sorted ℓ_1 regularization," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1240–1244, 2014.

[4] M. Figueiredo and R. Nowak, "Ordered weighted ℓ_1 regularized regression with strongly correlated covariates: Theoretical aspects," in *Artificial Intelligence and Statistics*, 2016, pp. 930–938.

[5] W. Su, E. Candes *et al.*, "SLOPE is adaptive to unknown sparsity and asymptotically minimax," *The Annals of Statistics*, vol. 44, no. 3, pp. 1038–1068, 2016.

[6] P. C. Bellec, G. Lecué, A. B. Tsybakov *et al.*, "SLOPE meets LASSO: improved oracle bounds and optimality," *The Annals of Statistics*, vol. 46, no. 6B, pp. 3603–3642, 2018.

[7] W. Su, M. Bogdan, E. Candes *et al.*, "False discoveries occur early on the LASSO path," *The Annals of Statistics*, vol. 45, no. 5, pp. 2133–2150, 2017.

[8] M. Bayati and A. Montanari, "The LASSO risk for gaussian matrices," *IEEE Transactions on Information Theory*, vol. 58, no. 4, pp. 1997–2017, 2012.

[9] N. E. Karoui, "Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results," *arXiv preprint arXiv:1311.2445*, 2013.

[10] C. Thrampoulidis, S. Oymak, and B. Hassibi, "Regularized linear regression: A precise analysis of the estimation error," in *Conference on Learning Theory*, 2015, pp. 1683–1709.

[11] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Precise error analysis of regularized M-estimators in high-dimensions," *IEEE Transactions on Information Theory*, 2018.

[12] H. Hu and Y. M. Lu, "Asymptotic characterizations and oracle optimal designs of SLOPE," 2019.